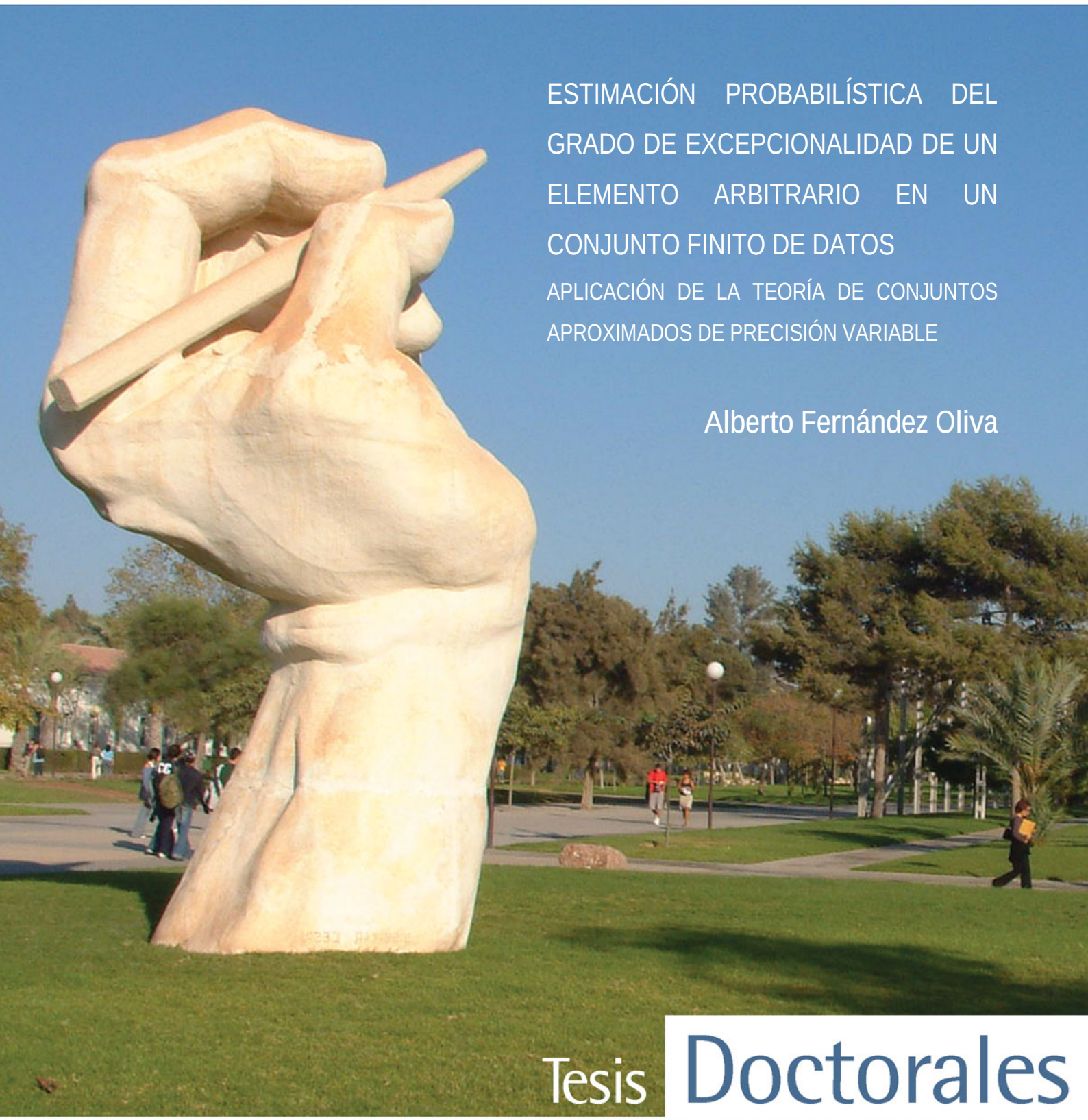
Universitat d'Alacant
Universidad de Alicante



ESTIMACIÓN PROBABILÍSTICA DEL GRADO DE EXCEPCIONALIDAD DE UN ELEMENTO ARBITRARIO EN UN CONJUNTO FINITO DE DATOS

APLICACIÓN DE LA TEORÍA DE CONJUNTOS APROXIMADOS DE PRECISIÓN VARIABLE

Alberto Fernández Oliva

Tesis Doctorales

www.eltallerdigital.com

UNIVERSIDAD de ALICANTE

Tesis Doctoral

Estimación Probabilística del Grado de Excepcionalidad de un Elemento Arbitrario en un Conjunto Finito de Datos

Aplicación de la Teoría de Conjuntos Aproximados de Precisión Variable

Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

UNIVERSIDAD DE ALICANTE

TESIS DOCTORAL

ESTIMACIÓN PROBABILÍSTICA DEL GRADO DE EXCEPCIONALIDAD DE UN ELEMENTO ARBITRARIO EN UN CONJUNTO FINITO DE DATOS

APLICACIÓN DE LA TEORÍA DE CONJUNTOS APROXIMADOS DE PRECISIÓN VARIABLE

Universitat d'Alacant
Universidad de Alicante

Presentada por
ALBERTO FERNÁNDEZ OLIVA

Dirigida por
DRA. Mª COVADONGA FERNÁNDEZ BAIZÁN
DR. FRANCISCO MACIÁ PÉREZ

DEPARTAMENTO DE TECNOLOGÍA INFORMÁTICA Y COMPUTACIÓN

SEPTIEMBRE DE 2010

Universitat d'Alacant
Universidad de Alicante

A la memoria de Raúl Delgado de la Cruz

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

«Si realmente deseas hacer algo y pones todo tu empeño, aprovechas las oportunidades que se te presentan y nunca te rindes, lo lograrás»

Jane Goodall

Naturalista y primatóloga

Doctora *Honoris Causa* por la Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Agradecimientos

Al llegar a este momento trascendente de mi carrera profesional, hay muchas personas a las cuales debo, directa o indirectamente, el haber podido llegar al mismo. Por tanto, sería inadmisible no dejar constancia explícita de mi más sincera gratitud a:

Mi madre, por haberme dado la vida y por darme cada día su bendición desde el lugar del firmamento en que se encuentre

Marcel –mi hijo–, por ser la consumación de mi realización personal. Su existencia me supone la mayor felicidad que pueda concebir y me inspira a alcanzar nuevas metas cada día

Alina –mi esposa–, por su inmenso cariño, incondicionalidad e infinita paciencia. Especialmente, por haberme hecho el mejor regalo del mundo: *MI HIJO*

Mi padre, por haber sido el mejor ejemplo de ser humano, humilde y bondadoso, que he conocido. Por haberme enseñado que en la vida hay principios y normas que son inviolables

Mis abuelos, por haber dedicado, con sumo esmero, los últimos años de su vida a mi crianza y a mi educación

Universitat d'Alacant
Universidad de Alicante

"El Brother", por su infinita nobleza, por su ayuda de siempre y por haber sido mi verdadero hermano

Cova –Ma. Covadonga Fernández Baizán–, por su inmensa generosidad e infinito cariño. Por haber sido siempre, una amiga excepcional. Sin lugar a dudas, a su empeño y dedicación debo la culminación de este trabajo

Paco –Francisco Maciá Pérez–, por poner su increíble talento y su inmensa capacidad de trabajo en función de mi Tesis. Por haber sido, más que un tutor, un excelente amigo

Miguel Alfonso Abreu Ortega, por haber puesto toda su inteligencia y sus ideas geniales en función de este trabajo. Sus aportes han sido decisivos en la realización del mismo

Nelson Melo, *Carlos Alberto Iglesias Álvarez –Mel–, Armando Rodríguez Fonte –Mandy–* y *Arturo Pié Joa* –estudiantes talentosos y excelentes seres humanos–, por su entrega a este proyecto

Prof. Dr. Carlos Narciso Bouza Herrera, por su amistad, por su ayuda incondicional y por su asesoría en los temas estadísticos

Todos mis compañeros y amigos de toda la vida –los que están y los que ya no están– de la Facultad de Matemática y Computación de la Universidad de La Habana, por soportarme durante todos estos años

Todos los buenos alumnos que, a pesar del paso del tiempo, me han seguido regalando su cariño sincero y me han seguido alentando para que cada día sea un mejor profesional

Universitat d'Alacant
Universidad de Alicante

Mis buenos amigos españoles que, a lo largo de todos estos años, me han regalado su generosidad, hospitalidad y ayuda sin límites: María y Rafael Portencasa, Tency Rojo, Raúl Martín y familia, Lucy Bruno y familia, Octavio Santana, Ernestina Menasalvas, Coro Pérez, Mar González y familia, Juanma Barbará y familia, María Jesús Fiteni, Jesús Peral, Antonio Vaquero, Carmen Fernández Chamizo y Marian Sánchez, entre otros

El colectivo de profesores y trabajadores del Departamento de Tecnología Informática y Computación de la Universidad de Alicante, especialmente a su Grupo M, así como a varios becarios de dicho departamento, por permitirme compartir junto a ellos momentos inolvidables de mi vida y por poner a mi disposición toda su inteligencia y los medios técnicos de que disponen. En especial, quiero destacar, por el cariño y el afecto brindado, a: Juanma, Diego, Virgilio, Rafa, Luis Felipe, Felipe, Juan Carlos, Jose, Mora, Héctor, Pablo, Jorge, Jorge Gea, Iván, Antonio, Miguel, María José, Marisa, Dany, Andrés y Ángel Grediaga

Todas las buenas personas que, a lo largo de toda mi vida, me han regalado un poco de afecto y cariño

Todas las malas personas que, para contrarrestar el daño que me han querido hacer, me han obligado a crecer y a madurar como ser humano

Mis más sinceras disculpas a quien se sienta olvidado. En mi corazón, seguro no lo está.

La Habana, a 10 de septiembre de 2010

Alberto Fernández Oliva

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Prefacio

Cuando tuve frente a mí el reto de escribir este *prefacio*, mi tutor me comentó que esto era algo personal que en su momento reflejó en la Memoria de su Tesis pero que nada me obligaba a hacerlo en la mía. No obstante, me motivó la idea y lo primero que me vino a la mente fue que esta misma situación coyuntural venía como *anillo al dedo* para ilustrar el tema central de mi tesis, en una primera aproximación: encontrar un *prefacio*, en una Memoria de Tesis Doctoral, es un *caso excepcional* en este tipo de documentos pero no por ello deja de ser útil al lector.

Reflexionando en torno a las interpretaciones que podría darle al tema para poder expresar, desde mi apreciación personal, el significado más intuitivo de la *detección de casos excepcionales — outliers—,* se me ocurrió asociarlo con fenómenos que están incorporados a la cotidianeidad pero que, muchas veces, uno ni siquiera medita en torno a ellos.

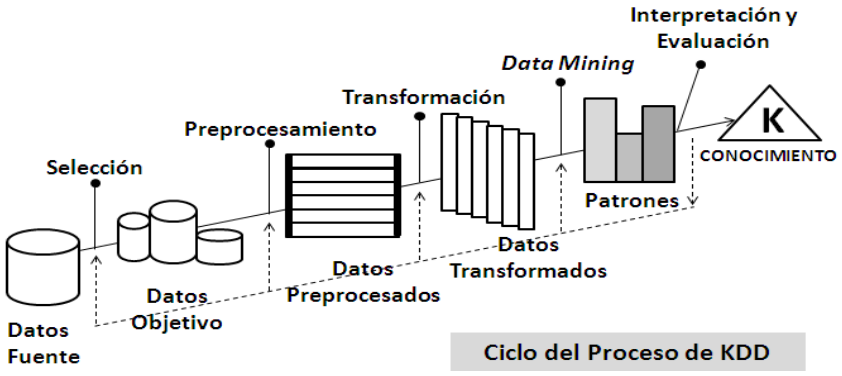
Teniendo en cuenta que la enorme cantidad de datos generados en la dinámica propia de las organizaciones ha pasado, en los últimos años, de ser un engorro a tener la consideración de recurso importantísimo para fundamentar el diseño de estrategias óptimas de actuación, me surgió la idea de establecer un paralelismo entre la utilidad de los *outliers* y el *Reciclado de residuos.*

Hasta hace poco tiempo, los *residuos* generados en el normal acontecer de la actividad humana eran considerados,

simplemente, un problema —y no precisamente pequeño—. La aparición de técnicas de *reciclado*, consecuencia de la cada vez más asentada conciencia ecológica en un mundo globalizado, ha hecho que esta perspectiva haya cambiado drásticamente. En el mundo actual, el nivel de desarrollo de un país se mide por la cantidad de *residuos* que genera; pero también, el nivel de *reciclado* de los mismos es un indicador fiable de su grado de cultura y desarrollo cívico.

Habiendo transcurrido dos décadas desde que se acuñó el término *KDD (Knowledge Discovery from Data)* para designar el *proceso de búsqueda de información en grandes volúmenes de datos*, me pregunté: ¿qué pueden tener en común el *Reciclado* y el *KDD*? Lo primero que valoré fue que ambos son procesos que parten de considerar —y la experiencia demuestra que con bastante éxito— los *residuos* como *un nuevo tipo de materia prima para la obtención de un producto nuevo.* Este *producto* puede ser material —en el caso del *Reciclado*— o intangible —en el caso del *KDD*—.

Recordé entonces el *ciclo del proceso de KDD* que observé en uno de los primeros artículos que, cuando comenzaba a meterme en esta investigación, me propició mi buen amigo Mayito —Mario Eugenio Díaz Laguardia—. Acudí, nuevamente, al referido documento y comencé a reflexionar en torno al mismo. Tras ello, me resultó fácil darme cuenta que, en esencia, no es nada diferente a un ciclo de *reciclado* convencional.



Ciclo del Proceso de KDD

En el esquema, donde aparecen los *Datos Fuente* —inicio del proceso—, situé mentalmente un *Contenedor de Residuos* y donde aparece el *Conocimiento* —final del proceso—, situé los *objetos manufacturados* que llevarán en alguna parte impreso el tan manoseado *símbolo internacional del reciclaje*: 

Resultaba trivial darse cuenta que, en ambos casos y entre ambos extremos, existe una cadena de procesos que, en etapas sucesivas, realiza *la transformación*. Y es aquí donde enmarcaría la detección de *casos excepcionales o raros* —que es como, en tono informal, se llama a los *outliers*—.

Centrándome en la etapa del *preprocesado,* en el caso *del reciclaje,* me hice una nueva pregunta: ¿Qué sucedería si entre los *residuos* se hubiese colado un diamante? y rápidamente acoté, si la *trituradora* no lo capta, pues,… ¡se acabó!

Por fortuna, la Informática es más sutil. Al principio, en el *preprocesado* de los datos, simplemente se eliminaban los datos *raros*, discordantes con el resto. Los *pobres outliers* fueron *basura* hasta un buen día en el que alguien cayó en la cuenta de que esta singularidad podía ocultar información valiosa: una aplicación al universo del *KDD* de lo que se conoce con el término de *serendipicidad (serendipity)*, que viene a ser el hallazgo, por casualidad, de algo realmente importante, pero que no era, ni mucho menos, lo que andábamos buscando.

Desde entonces, el ciclo *KDD* se desdobla —o debería hacerlo— en dos: El *preprocesado de datos convencionales*, que producirá reglas o patrones de adecuado soporte y fiabilidad y la *detección de outliers* —que luego serán objeto de estudio e interpretación— para no dejar a la *serendipicidad* el posible hallazgo de *diamantes* ocultos en nuestros *residuos.*

Finalmente, me convencí de que trabajar con *outliers* es como *reciclar residuos*: sacar algo de provecho de lo que antes se *tiraba a la basura*.

Alberto Fernández Oliva

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Resumen

En un proceso de *minería de datos (Data Mining-DM)* se entiende por *casos excepcionales —outliers—* los objetos que muestran un comportamiento *anormal* en relación al contexto donde se encuentran o que tienen valores inesperados en algunos de los parámetros considerados. Por la importancia que ello reviste en los *procesos de búsqueda de información en grandes volúmenes de datos* (*KDD*), los investigadores, en los últimos años, han prestado especial atención al desarrollo de técnicas de detección eficientes. Si, en general, tales procesos se enfocan en el sentido de descubrir patrones de comportamiento representativos —alta fiabilidad y soporte—, la detección de *outliers* aprovecha justamente la elevada marginalidad de estos objetos para detectarlos midiendo su grado de desviación respecto a dichos patrones y a partir de ello deducir conocimiento relevante.

Por otra parte, la *Teoría de Conjuntos Aproximados —Rough Sets—* juega un papel importante cuando se desea hacer razonamiento a partir de datos imprecisos o, siendo más exactos, cuando se desea establecer relaciones entre los datos. En especial, el enfoque del proceso de *minería de datos* bajo la concepción de esta teoría está basado en el conocimiento que un *agente* (o grupo de *agentes*) tiene

acerca de cierta realidad y su capacidad para discernir algunos fenómenos, procesos, objetos, etc. Por tanto, su enfoque se basa en la capacidad de clasificar datos que han sido obtenidos por diversas vías. De igual manera, su aplicación ha resultado efectiva en la solución de múltiples problemas de la vida real.

Si bien la aplicación de la Teoría de *Rough Sets* al campo de los procesos *KDD* viene realizándose desde su formulación por Z. Pawlak en la década de los 80 del pasado siglo, en los últimos años se ha comenzado a considerar la detección de *outliers* como un proceso de *KDD* con interés en sí mismo. La combinación de ambos enfoques (*Rough Sets* como fundamento para la caracterización y detección de *outliers*), es un punto de vista absolutamente nuevo, con un gran potencial de interés teórico y aplicabilidad práctica, todo lo cual nos animó a orientar nuestra investigación en esta dirección.

Para abordar este planteamiento se realizó un detallado estudio del estado actual de la técnica así como de propuestas existentes. Todo ello posibilitó el establecimiento de un marco teórico general que permitió proporcionar un método de detección de *outliers*, computacionalmente viable y no determinista, basado en el *Modelo de Conjuntos Aproximados de Precisión Variable (VPRSM).*

Por último, se realizó una ampliación del marco teórico alcanzado para establecer una aproximación estocástica a la solución del problema de determinar si un elemento dado es *outlier* dentro de un determinado *universo* de datos. Esto significa que, a partir de ella, se facilita el establecimiento de un criterio probabilístico sobre dicha condición para dichos elementos.

Universidad de Alicante

Resum

Els casos excepcionals —*outliers*— són objectes que mostren un comportament anormal dintre del context on es troben o que tenen valors inesperats en alguns dels seus paràmetres. Per la importància que això revesteix en els processos de recerca d'informació en grans volums d'informació, els investigadors, en els últims anys, han prestat especial atenció al desenvolupament de tècniques de detecció eficients. Si, en general, tals processos s'enfoquen en el sentit de descobrir patrons de comportament representatius —alta fiabilitat i suport—, la detecció de *outliers* aprofita justament l'elevada marginalitat d'aquests objectes per a detectar-los amidant el seu grau de desviació respecte a aquests patrons i a partir d'això deduir informació —*coneixement*— rellevant.

Per altra banda, la *Teoria de Conjunts Aproximats —Rough Sets*— juga un paper important quan es desitja fer raonaments a partir de dades imprecises o, sent més exactes, quan es desitgen establir relacions entre les dades. Especialment, l'enfocament de la mateixa està basat en el coneixement que un agent (o grup d'agents) tingui sobre certa realitat i la seva capacitat per a destriar alguns fenòmens, processos, objectes, etc. Per tant, el seu enfocament es basa en l'habilitat per a classificar dades

Universitat d'Alacant
Universidad de Alicante

que han estat obtinguts per diverses vies. D'igual manera, la seva aplicació ha resultat efectiva en la solució de múltiples problemes de la vida real.

Tenint en compte que ambdós aspectes —problema de la detecció de *outliers* i teoria de *Rough Sets*— incideixen directament en els processos abans esmentats, en aquesta Tesi vam pretendre vincular els mateixos aplicant aquesta teoria al referit problema.

Si bé l'aplicació de la Teoria de *Rough Sets* al camp dels processos *KDD* ve realitzant-se des de la seva formulació per Z. Pawlak en la dècada dels 80 del passat segle, en els últims anys s'ha començat a considerar la detecció de *outliers* com un procés de *KDD* amb interès en si mateix. La combinació d'ambdós enfocaments (*Rough Sets* com fonament per a la caracterització i detecció de *outliers*), és un punt de vista absolutament nou, amb un gran potencial d'interès teòric i aplicabilitat pràctica, tot la qual cosa ens va animar a orientar la nostra investigació en aquesta adreça.

Per a abordar aquest plantejament es va realitzar un detallat estudi de l'estat actual de la tècnica així com de propostes existents. Tot la qual cosa va possibilitar l'establiment d'un marc teòric general que ens va permetre proporcionar un mètode de detecció de *outliers*, computacionalment viable i no determinista, basat en el *Model de Conjunts Aproximats de Precisió Variable* (*VPRSM*).

Pretenent optimitzar el cost, quant a temps, del procés general d'anàlisi de les dades, es va realitzar una ampliació del marc teòric arribat a de manera tal que el mateix va possibilitar la proposta d'un mètode que permet, de forma desatesa —pel que fa a l'elecció dels paràmetres que intervenen en l'anàlisi—, establir una aproximació estocàstica a la solució del problema de determinar si un

Universitat d'Alacant
Universidad de Alicante

element donat és *outlier* dintre d'un determinat univers de dades. Això significa que, a partir d'aquesta solució es facilita l'establiment d'un criteri probabilístic sobre aquesta condició per a aquests elements.

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Abstract

In *Data Mining* processes, *outliers* are objects that display an abnormal behavior, that is to say: They have unexpected values in some of their parameters taken into account in the analysis. Given their importance in the process of information retrieval from large amounts of data, researchers have dedicated, in recent years, much attention to the development of efficient detection techniques. Although, in general, these processes focus on the recognition of representative behavior patterns having high reliability and support, *outliers detection* rests on the high marginality of these objects, measuring their deviation from recognized patterns to extract information and knowledge.

Rough Set Theory plays an important role when reasoning is performed on imprecise data. This approach is based on the knowledge that an agent has about a certain reality and on their ability to discern phenomena, processes, objects, etc. In other words, it's based on the capability of classifying data obtained from different sources. This theory has been applied effectively in many real life problems.

If, on the one hand, the implementation of the *Rough Sets Theory* to the field of the *KDD* process has been occurring ever since its formulation by Z. Pawlak in the decade of

80's, last century, in the past years the *outliers detection* has started to be seen as a *KDD* process with interests in itself. The combination of both approaches (*Rough Sets* as a foundation for the characterization and *outliers detection*), constitutes an absolutely new point of view, with a great potential for it's the theoretical interest and its practical applicability. All of the above encouraged us to orient our investigation in this regard.

A detailed study of the current state of the art is included. Then, we have drawn a theoretical frame that allowed us to design a method for *outliers detection* which is computationally feasible and non-deterministic, based on the *Variable Precision Rough Sets Model*.

With the goal of optimizing the time complexity of the general process of data analysis, the theoretical frame was enlarged. This made possible the development of a method which allows, in an unattended way (regarding the selection of the parameters involved in the analysis), to obtain a stochastic approximation to the solution of the problem of determining whether a given element is an *outlier* in a given data set.

Universitat d'Alacant
Universidad de Alicante

Contenido

AGRADECIMIENTOS ..... **I**
PREFACIO ..... **V**
RESUMEN ..... **IX**
RESUM ..... **XI**
ABSTRACT ..... **XV**
CONTENIDO ..... **XVII**
FIGURAS ..... **XXIII**
TABLAS ..... **XXVII**

CAPÍTULO 1
INTRODUCCIÓN **1**

MOTIVACIÓN Y JUSTIFICACIÓN 5
FORMULACIÓN DEL PROBLEMA 11
HIPÓTESIS DE PARTIDA 12
OBJETIVOS 13
PROPUESTA DE SOLUCIÓN 14
METODOLOGÍA 16
PLAN DE TRABAJO 19

Universitat d'Alacant
Universidad de Alicante

CAPÍTULO 2
ESTADO DEL ARTE 23

PRINCIPALES MÉTODOS ESTADÍSTICOS PARA LA DETECCIÓN DE *OUTLIERS* 24

Detección Univariante 24
Detección Bivariante 29
Detección Multivariante 30

ASPECTOS BÁSICOS PARA LA DETECCIÓN DE *OUTLIERS* CON MÉTODOS ESTADÍSTICOS 39

La Calificación de un Caso de Outlier 40
La Descripción de los Outliers y su Especificación 40
El Mantenimiento o la Eliminación de los Outliers 40

PRINCIPALES CRÍTICAS A LOS MÉTODOS ESTADÍSTICOS DE DETECCIÓN DE *OUTLIERS* 42
APLICACIÓN DE LA DETECCIÓN DE *OUTLIERS* 44
CAUSAS DE LA APARICIÓN DE *OUTLIERS* 48
CÓMO PROCEDER ANTE LA APARICIÓN DE *OUTLIERS* 49
TÉCNICAS DE DETECCIÓN DE *OUTLIERS* 51

Técnicas basadas en Distribuciones 53
Técnicas basadas en Profundidades 53
Técnicas basadas en Distancias 54
Técnicas basadas en Densidades 55
Métodos basados en Técnicas de Clusters 57
Técnicas basadas en el Uso de Redes Neuronales 58
Métodos basados en Redes Supervisadas 58
Métodos basados en Redes No Supervisadas 59
Técnicas basadas en Subespacios 60
Técnicas basadas en Soporte Vectorial (Support Vector) 60
Otras Técnicas de Detección de Outliers 61

CONSIDERACIONES PARA LA CONCEPCIÓN DE MÉTODOS DE DETECCIÓN DE *OUTLIERS* 62
CONSIDERACIONES SOBRE COMPARACIONES DE MÉTODOS 64
FINANCIACIÓN DA LAS INVESTIGACIONES 66
TEORÍA DE *ROUGH SETS* APLICADA A LA DETECCIÓN DE *OUTLIERS* 68
CONCLUSIONES 71

Universitat d'Alacant
Universidad de Alicante

CAPÍTULO 3
ANTECEDENTES DE LA INVESTIGACIÓN 75

CONCEPTOS FUNDAMENTALES DE LA TEORÍA DE *RS* 77
ANÁLISIS CRÍTICO DE UN MÉTODO DE DETECCIÓN DE *OUTLIERS* BASADO EN *ROUGH SETS* 77
CONCLUSIONES DEL ANÁLISIS 84

CAPÍTULO 4
DETECCIÓN EFICIENTE DE *OUTLIERS* BASADA EN *RSBM* 87

PROPUESTA DE AMPLIACIÓN DEL MARCO TEÓRICO 89
MÉTODO DE DETECCIÓN DE *OUTLIERS* BASADO EN *RSBM* 95
Pseudo-código del Algoritmo RSBM 95
IMPLEMENTACIÓN COMPUTACIONAL. ALGORITMO *RSBM* 102
Complejidad Espacial 103
Complejidad Temporal 103
LOCALIZACIÓN DE *OUTLIERS* DENTRO DE UN CONJUNTO DE DATOS 105
VALIDACIÓN DE LOS RESULTADOS 108
Prueba 4-1 108
Prueba 4-2 111
CONCLUSIONES 115

CAPÍTULO 5
DETECCIÓN NO DETERMINISTA DE *OUTLIERS* BASADA EN *VPRSM* 119

RSBM FRENTE A *VPRSM* 121
VPRSM. NOTACIONES BÁSICAS Y PROPIEDADES 123
MÉTODO DE DETECCIÓN NO DETERMINISTA DE *OUTLIERS* BASADO EN *VPRSM* 131
Pseudo-código del Algoritmo VPRSM 132
Fase 1. Construcción de β-Fronteras Internas 132
Fase 2. Construcción del Conjunto E y Detección de Outliers 134
IMPLEMENTACIÓN COMPUTACIONAL. ALGORITMO *VPRSM* 135

Universitat d'Alacant
Universidad de Alicante

LOCALIZACIÓN DE *OUTLIERS* DENTRO DE UN CONJUNTO DE DATOS....136
VALIDACIÓN DE LOS RESULTADOS....139

Prueba 5-1....139
Prueba 5-2....142
Prueba 5-3....146
Prueba 5-4....152

CONCLUSIONES....165

CAPÍTULO 6
ALGORITMO *BETA-MIU* PROBABILÍSTICO....169

FASE 1. DETERMINACIÓN PARA CADA ELEMENTO DEL *UNIVERSO* DE SU REGIÓN DE EXCEPCIONALIDAD....172

Marco Teórico. Algoritmo BM....172
Implementación computacional. Algoritmo BM....202
Estudio de la complejidad espacial y temporal. Algoritmo BM....213
Conclusiones....220

FASE 2. ESTIMACIÓN DE LA PROBABILIDAD DE CADA ELEMENTO DEL *UNIVERSO* DE SER *OUTLIER*....221

Marco Teórico. Algoritmo BM/Probabilístico....221
Optimización....224
Implementación Computacional. Algoritmo BM/Probabilístico....231
Validación de los resultados....233

CAPÍTULO 7
COMPARACIÓN DE MÉTODOS....245

CONSIDERACIONES GENERALES....246
ALGORITMOS DE DETECCIÓN DE *OUTLIERS* BASADOS EN LA TEORÍA DE *ROUGH SETS*....247

Algoritmo RSBM....248
Algoritmo VPRSM....250
Consideraciones comunes para RSBM y VPRSM....251

Universitat d'Alacant
Universidad de Alicante

Algoritmo BM *252*
Algoritmo BM/Probabilístico *253*
Resumen *255*

COMPARACIÓN CON OTROS MÉTODOS DE DETECCIÓN DE *OUTLIERS* 257

Métodos Estadísticos *257*
Métodos basados en Distancias *259*
Enfoques basados en Densidades *261*
Enfoques basados en Profundidades (depth-based) *264*
Enfoques basados en Particiones (Partition-based) *265*
Enfoques basados en Redes Neuronales *268*

CAPÍTULO 8
CONCLUSIONES 273

ÁMBITO DE APLICACIÓN DE LA PROPUESTA 276
PRINCIPALES APORTACIONES 277
PROBLEMAS ABIERTOS Y LÍNEAS FUTURAS DE INVESTIGACIÓN 281

Problema No. 1 ..... *281*
Problema No. 2 ..... *282*

CONCLUSIONES SOBRE PROCESO DE INVESTIGACIÓN 283

REFERENCIAS BIBLIOGRÁFICAS 285

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Figuras

Figura 1-1 Marco teórico general 15
Figura 1-2 Esquema general de solución propuesto 16

Figura 2-1 Ejemplo de distribución Normal 25
Figura 2-2 Criterios de detección de *outliers* a partir de la distribución Normal 26
Figura 2-3 Ejemplo de un gráfico de *Box & Whiskers* 28
Figura 2-4 Gráfico de *B&W* mostrando la detección de *outliers* 28
Figura 2-5 Diagrama de dispersión 30
Figura 2-6 Interpretación bivariante de los gráficos de *B&W* 30
Figura 2-7 Ejemplo de gráficos de *B&W* para el caso bivariante 31
Figura 2-8 Ejemplo de aplicación de la regla del *NN* 37
Figura 2-9 Aplicación de la regla de los *k-vecinos más cercanos* (*k*=3) 38
Figura 2-10 Aplicación de la técnica basada en distancias para una distancia *d* y *k*=3 54

Figura 3-1 Aproximación Inferior, Superior y Frontera 76

Figura 4-1 Estructuras de datos utilizadas 101
Figura 4-2 Partición que estable r_1 sobre U y Frontera de X respecto a r_1 104
Figura 4-3 Partición que estable r_2 sobre *U* y Frontera de *X* respecto a r_2 105
Figura 4-4 Variando número de columnas del conjunto de datos 107
Figura 4-5 Variando la cardinalidad del conjunto de datos 108
Figura 4-6 Variando el número de relaciones de equivalencia 109
Figura 4-7 Resultados de las pruebas de detección 113

Figura 5-1 Inclusión de conjuntos estándar 122
Figura 5-2 Ejemplo de inclusión mayoritaria 124
Figura 5-3 La relación de inclusión mayoritaria no es transitiva 124
Figura 5-4 Regiones representativas para $\beta=0$. Correspondencia con *RSBM* 127
Figura 5-5 Variación de las regiones significativas para un *error de clasificación* $\beta=0,1$ 127

Universidad de Alicante

Figura 5-6	Variación de las regiones significativas a partir de la variación del *β-error*	129
Figura 5-7	Partición que establece r_1 sobre *U* y *frontera* de *X* respecto a r_1. $\beta=0$; $\beta=0,25$	135
Figura 5-8	Partición que establece r_2 sobre *U* y *frontera* de *X* respecto a r_2. $\beta=0$; $\beta=0,25$	137
Figura 5-9	*RS vs. VPRSM* en cuanto a tiempo de ejecución. Variando número de columnas	139
Figura 5-10	*RS vs. VPRSM* en cuanto a tiempo de ejecución. Variando número de relaciones	139
Figura 5-11	*RS vs. VPRSM* en cuanto a tiempo de ejecución. Variando número de filas	140
Figura 5-12	*Tiempo de ejecución* del algoritmo *VPRSM* sobre un *conjunto de datos* generado de forma sintética	144
Figura 5-13	*RS* modelo básico *(RSBM) vs. VPRSM* en cuanto a detección de *outliers*	147
Figura 5-14	Detección de *outliers* Prueba 5-4.1: Arrhythmia DS	154
Figura 5-15	Detección de *outliers* Prueba 5-4.2*: Arrhythmia DS*	162
Figura 6-1	Interpretación gráfica de la función $\varphi(A, B)$	173
Figura 6-2	Interpretación de la función φ(C, X) desde el punto de vista teórico. X⊆U es el conjunto de todos los elementos del universo que cumplen el concepto y C⊂U, C≠ϕ	174
Figura 6-3	Posibles opciones para el Caso 1: φ(P1, X) = ϕ	178
Figura 6-4	Intervalo de valores de β para el cual se cumple $C_1 \subseteq B_j$ cuando estamos en el Caso 1: $\varphi(P_1, X) = \phi$	179
Figura 6-5	Restricción inicial de valores de β para el Caso 2: $\varphi(P_1, X) = X$	180
Figura 6-6	Rango de valores de β para los cuales se cumple el Caso 2: $\varphi(P_1, X) = X$	183
Figura 6-7	Conjunto de valores de β para los cuales $B_i=\phi$	186
Figura 6-8	Conjunto de valores de β determinado por $M_i(a)$	189
Figura 6-9	Relación entre los intervalos $M_i(a)$ y $M_j(a)$	191
Figura 6-10	Ordenación de los $\lambda_i(a)$ a partir de la permutación $Z_i(a)$	191
Figura 6-11	Interpretación gráfica de la función Total(a, β)	193
Figura 6-12	Valores de μ para los cuales $\mu \leq GrExcep(a,\beta_0)$	194
Figura 6-13	Región de valores β-μ para los cuales un elemento *a* cualquiera del *universo* es *outlier* en *U*	196
Figura 6-14	Región de excepcionalidad para un elemento muy contradictorio	197
Figura 6-15	Región de excepcionalidad para un elemento muy poco contradictorio	198
Figura 6-16	*Región de excepcionalidad* para un elemento que, para determinados valores de β, a pesar de su *grado de excepcionalidad*, no pertenece a ningún *conjunto excepcional no redundante*	199
Figura 6-17	Región de valores beta-miu asociada al país Brasil	224
Figura 6-18	Aplicación de un corte, a partir del valor t=0,1, al área total de la región beta-miu	225
Figura 6-19	Efectos del *corte* al ser aplicado a la región que caracteriza a un elemento poco contradictorio para muchos valores de β	226
Figura 6-20	Efectos del *corte* al ser aplicado a la región que caracteriza a un elemento muy contradictorio para pocos valores de β	227

Figura 6-21 Comparación de tiempos de ejecución entre los algoritmos *VPRSM* y BM/Probabilístico 232
Figura 6-22 Resultados de la prueba Prueba 6-2 realizada al algoritmo *BM*/Probabilístico 235
Figura 6-23 Resultados de la prueba Prueba 6-3.1 realizada al algoritmo *BM*/Probabilístico 238
Figura 6-24 Resultados de la prueba Prueba 6-3.2 realizada al algoritmo *BM*/Probabilístico 240

Figura 7-1 *Clustering Outliers Removal. Clusters* y *outliers* 264
Figura 7-2 *Clustering Outliers Removal. Clusters* sin *outliers* 265

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Tablas

Tabla 3-1 Ejemplo de un *universo* de datos sobre *personas* 80

Tabla 4-1 Datos sobre *pacientes* que representan un *universo* dado U 103
Tabla 4-2 *outliers* introducidos en el conjunto de datos 111

Tabla 5-1 Datos del ejemplo que representa al *universo* U 136

Tabla 5-2 *Outliers* introducidos en el conjunto de datos: *Census Bureau DB* 147

Tabla 5-3 *Outliers* introducidos Prueba 5-4.1*: Arrhythmia DS* 153

Tabla 5-4 Detección de *outliers* Prueba 5-4.1: *Arrhythmia DS* 155

Tabla 5-5 *Outliers* introducidos Prueba 5-4.2*: Arrhythmia DS* 160

Tabla 5-6 Detección de *outliers* Prueba 5-4.2: *Arrhythmia DS* 161

Tabla 6-1 Algoritmo BM. *complejidad espacial* de las estructuras de datos 217
Tabla 6-2 Algoritmo BM. *complejidad temporal* de los métodos y del algoritmo 217
Tabla 6-3 Conjunto de datos sobre *países* 223
Tabla 6-4 Probabilidad para los *outliers* introducidos. Prueba 6-3.1: *Arrhythmia DS* 239
Tabla 6-5 Probabilidad para los *outliers* introducidos. Prueba 6-3.2: *Arrhythmia DS* 241

Tabla 7-1 Cuadro comparativo, atendiendo a VENTAJAS y DESVENTAJAS, de los algoritmos *RSBM, VPRSM, BM* y *BM*/Probabilístico 253
Tabla 7-2 Principales limitaciones de otros métodos de detección que resuelven los algoritmos de detección basados en *RS* 269

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Capítulo 1

Introducción

Las investigaciones más recientes en aspectos relacionados con el *proceso de búsqueda de información en grandes volúmenes de datos* (*KDD*) prestan, en la actualidad, especial atención a la detección de los *casos excepcionales* (*outliers)* que puedan observarse dentro de los grandes volúmenes de información. Los *outliers* se detectan con el objetivo de ser interpretados en el contexto del análisis, con propósito diverso, especialmente con el objetivo de ser evaluados por el tipo de información que pueden proporcionar. Dicha información puede posteriormente tener incidencia directa en la toma de decisiones dentro de cualquier contexto de aplicación.

Donde primero se abordó este problema fue en el campo de la *Estadística* pero, como ya hemos señalado, en los últimos años ha tomado gran trascendencia en otras áreas de investigación. Dentro del proceso general de *KDD,* la *minería de datos* (*Data Mining —DM*) adquiere especial trascendencia. En su concepción más tradicional, a los datos se les aplica un conjunto de técnicas de *Inteligencia Artificial* encaminadas, fundamentalmente, a la búsqueda de patrones de comportamiento representativos dentro de

los mismos. En este importante paso dentro del *KDD,* la detección de *outliers* adquiere especial relevancia. La oportuna detección de dichos objetos es, en algunos casos, la motivación principal del análisis, ya que pueden aportar conocimiento oculto o información relevante en relación a diferentes aspectos del dominio específico de aplicación en que se esté trabajando. En otros casos, la oportuna detección de *outliers* puede garantizar la confiabilidad de los resultados. Su presencia puede hacer poco confiable los patrones que se obtengan como resultado del propio proceso de *minería de datos*.

Aunque no existe una definición formal de *outlier* universalmente aceptada por todos, se han propuesto algunas que son las más referenciadas por la comunidad científica internacional.

La de Grubbs (Grubbs, 1969) es una de las primeras que describe acertadamente el concepto de *outlier* a partir de su comportamiento:

> *«Una observación sobresaliente, o excepcional, que se desvía marcadamente de los otros miembros de la muestra en la que ocurre»*

Universitat d'Alacant
Universidad de Alicante

Igualmente sucede con la de Hawkins (Hawkins, 1980) en la que también se describe a un *outlier* por su comportamiento:

> *«Un outlier es una observación que se desvía tanto de las otras, que permite suponer que se ha generado por un mecanismo diferente»*

Barnett y Lewis (Barnett & Lewis, 1994), por su parte, toman como base la definición de Grubbs, y en base a ella, enuncian una nueva:

> *«Un outlier en un conjunto de datos, no es más que una observación, o un conjunto de observaciones, que parece ser inconsistente con el resto»*

Como se puede observar fácilmente, el núcleo de todas ellas, es el mismo:

> *… se desvía, marcadamente de los otros…*
>
> *… se desvía tanto de las otras…*
>
> *… parece ser inconsistente con el resto de los datos…*

En todas se destaca la *diferencia*, la *distinción*, la *marginalidad*, el *grado de desviación* o la *excepcionalidad* del *objeto* que se considera *outlier,* en cuanto a su comportamiento, con respecto al resto de los *objetos* que integran el *universo* de datos sujeto a estudio.

El tema de la detección de *casos excepcionales* se ha venido asociando a un amplio conjunto de técnicas. Por lo general, la mayoría de estas técnicas son esencialmente idénticas y sólo se diferencian por el nombre con el que los autores las identifican. Entre ellas, podemos citar:

- *Novelty detection*
- *Anomaly detection*
- *Noise detection*
- *Deviation detection*
- *Exception mining*

No obstante, el término generalmente más usado para identificar la técnica es: *detección de casos excepcionales* (*outlier detection*).

La detección de *outliers* ha resultado ser de gran importancia en diferentes campos. Para poder comprender

el alcance de este problema es importante tener una idea de los diferentes contextos de aplicación del mismo.

En muchos casos, la detección de *casos excepcionales* puede estar vinculada con aspectos relacionados con la seguridad. Su oportuna detección puede evitar que se caiga en una situación de *peligro* o de *emergencia*. Por ejemplo, la detección de *outliers* puede proporcionar indicios del mal funcionamiento del motor de una nave aérea o permitir detectar la presencia de algún *intruso* que intenta burlar la protección de algún sistema. En sentido general, en la sociedad actual, invadida por las nuevas tecnologías de la información, la detección de *outliers* es de trascendental importancia. La detección de un cambio repentino en el normal funcionamiento de los sistemas puede implicar la ocurrencia de un *fraude* o de un robo.

La industria representa otro contexto donde la detección de *outliers* puede ser trascendente. Por ejemplo, puede incidir en el control del flujo de producción de una determinada organización donde se monitoriza en tiempo real el control de la calidad de los productos que en ella se fabrican.

En el comercio, los métodos de detección de *outliers* pueden ser aplicados en *estudios de mercado* y en la identificación de nuevas tendencias a partir de las cuales se abren nuevas oportunidades de negocio para las empresas.

Son, quizás, las grandes *bases de datos* el escenario donde más interpretaciones pueden tener hoy en día la presencia de *outliers*. Podrían indicar un uso indebido de los datos, un error en la entrada de los mismos o una mala interpretación de algún valor ausente. En todos los casos, su oportuna detección es de suma importancia para mantener la integridad y la consistencia de la misma.

En la actualidad, como ya se ha comentado, este tema toma especial relevancia en los *procesos de búsqueda de*

información en grandes volúmenes de datos (*Knowledge Discovery on Data - Data Mining —KDD-DM*). Si dichos procesos se centran fundamentalmente en el descubrimiento de patrones de comportamiento representativos, la detección de *outliers* aprovecha justamente la elevada marginalidad de dichos objetos para descubrirlos midiendo su *grado de desviación* con respecto a dichos patrones.

Motivación y Justificación

Sin que sea necesario profundizar mucho desde el punto de vista teórico, una primera aproximación a la dimensión del problema de la detección de *outliers*, a partir de los ejemplos que se acaban de mencionar, permite darnos cuenta de la importancia y la trascendencia del mismo.

Una eficiente detección de *outliers* puede evitar que se tomen decisiones basadas en datos erróneos, además de ayudar en la detección, prevención y reparación de los efectos negativos que puede traer consigo el uso indebido de los mismos. La presencia de *outliers* en un *conjunto de datos* puede entorpecer la detección de patrones confiables dentro del proceso de *minería de datos*.

En otros casos, a través de la detección de *outliers* se puede descubrir gran cantidad de conocimiento oculto, inesperado, interesante y útil dentro de los grandes volúmenes de información.

Esto último nos hace reflexionar en relación al siguiente aspecto: aunque el término *outlier*, por la *excepcionalidad* que lleva implícita su concepción, podría hacer suponer que siempre trae aparejado consigo una interpretación negativa del fenómeno, en la práctica, esto no siempre es así. En diversas ocasiones, la detección de *casos*

excepcionales, como casos distinguidos por su *excepcionalidad*, puede ser el objetivo fundamental de algún proceso, análisis o estudio. Por tanto, no siempre se debe asumir el término *excepcional* a partir de una interpretación negativa del mismo.

Por norma, los *casos excepcionales* no pueden ser caracterizados categóricamente como *beneficiosos* o como *problemáticos*, sino que deben ser interpretados en el contexto del análisis. Deben ser evaluados por el tipo de información que proporcionan. Esto puede llegar a marcar la diferencia entre el objetivo de una u otra investigación.

En la actualidad, los investigadores crean modelos, algoritmos y funciones, definidos en espacios cada vez más abstractos. Todo esto hace pensar que, en el futuro, el desarrollo de las investigaciones se verá cada vez más condicionado por la naturaleza de los datos. A partir de esta realidad, se hace necesario concebir técnicas de análisis de datos cada vez más novedosas y eficientes. La *minería de datos* se ha consolidado como un área de la *Inteligencia Artificial* que brinda técnicas, teorías y herramientas que posibilitan, de forma eficiente, el análisis de los complejos *conjuntos de datos* del mundo actual.

Universidad de Alicante

Teniendo en cuenta que el presente trabajo de investigación se enmarca dentro de un proyecto más general que tiene que ver directamente con los procesos de *KDD-DM,* consideramos relevante abordar, a grandes rasgos, aspectos que justifican la trascendencia de la detección de *outliers* en dicho contexto.

Las primeras aplicaciones en el campo del *DM* centraban su atención solo en los patrones que se observaban con más frecuencia en los datos. Despreciaban, o descartaban, otros que se observaban con menor frecuencia y que contenían objetos de comportamiento *extraño* o *excepcional* que,

quizás, podrían también aportar información de mucha importancia. Tomando conciencia de este hecho, las investigaciones más recientes relacionadas con el *KDD* y especialmente, las relacionadas con el importante paso dentro del mismo (Fayyad *et al.*, 1996) donde se realiza el *DM*, consideran el problema de la detección de *outliers* como un problema medular. Algunos (Chawla & Sun, 2006), justifican la trascendencia de la detección de *outliers* dentro del *DM,* considerándolo uno de los cuatro aspectos fundamentales que garantizan el equilibrio del mismo:

- Clasificación.
- Agrupamiento (*Clustering*).
- Reglas de Asociación.
- Detección de *outliers*.

Si los procesos de *KDD-DM, en general,* se orientan en el sentido de descubrir patrones de comportamiento representativos —alta fiabilidad y soporte—, la detección de *outliers* sirve para medir el *grado de desviación* de estos objetos respecto a dichos patrones y, en muchos casos, el fenómeno de la detección se ve como un proceso de descubrimiento de información (*conocimiento*) de gran utilidad en el análisis e interpretación de los datos.

En general, desde la perspectiva del *KDD-DM*, el tema de la detección de *outliers* se enfoca desde dos puntos de vista diferentes:

- Los *outliers* como objetos indeseables que deben ser atendidos o eliminados en la fase de *preparación de los datos.* Su presencia en el *conjunto de datos* puede entorpecer notablemente la detección de patrones confiables.
- Los *outliers* como objetos susceptibles a ser identificados, a partir del interés implícito que tienen para el propio proceso. En tal caso, no deben ser

eliminados del *conjunto de datos*. Para algunas aplicaciones los eventos *excepcionales* son más representativos e interesantes que los eventos más comunes desde el punto de vista del descubrimiento de información. Un ejemplo de aplicaciones en este sentido son las referidas a la detección de *fraudes* en el uso de tarjetas de créditos y en el comercio electrónico. En un caso, los *outliers* podrían brindar información para tipificar patrones de conducta indebida y, en el otro, para suministrar información útil para el *marketing*.

Todo lo anterior pone de manifiesto que los procesos de *KDD-DM* requieren métodos cada vez más efectivos para la detección de *outliers.* En los *conjuntos de datos* actuales aparecen datos, estructuras de representación y formas de almacenamiento cada vez más sofisticadas. Por tanto, hay que trabajar en función de la obtención de modelos de detección efectivos, en correspondencia con los retos que imponen tales particularidades, así como el uso de las nuevas tecnologías, en general.

El estudio del *estado del arte* realizado en el marco de esta investigación ha permitido identificar el alcance del problema de la detección de *outliers* a partir de su aplicación en múltiples contextos. Su ámbito de aplicación es amplio y diverso. Algunos ejemplos concretos se mencionan a continuación:

- Con la universalización del uso de las redes de computadoras, los temas de *seguridad* toman especial importancia. En ese aspecto, los métodos de detección de *outliers* constituyen un elemento esencial en lo referido a la detección de *fraudes* e *intrusos*. (Lazarevic *et al.*, 2003)

- Otros ejemplos de aplicación son: aplicaciones bancarias o financieras (Last & Kandel, 2001); aplicaciones médicas (Hawkins *et al.*, 2002); diagnóstico de fallos en diversos contextos. (Qinglin *et al.* 2007); tratamiento de imágenes y videos (Knorr *et al.*, 2000); investigaciones químico–farmacéuticas (Cramer *et al.*, 2004); *e-business/e-commerce* (He *et al.*2004); investigaciones socio–ambientales (Li *et al.*, 2006); investigaciones socio–culturales (Stomoimenova *et al.*, 2005) y estudios poblacionales (Otey *et al.*, 2005c).

La diversidad de contextos donde se enmarcan las aplicaciones mencionadas, evidencia que los problemas relacionados con la detección de *outliers* inciden en aspectos de interés social, económico, científico y cultural del mundo actual. Por tanto, las investigaciones encaminadas a aportar soluciones en esta dirección, tienen impacto y trascendencia en lo relativo al desarrollo de los mismos.

Universitat d'Alacant
Universidad de Alicante

Esta diversidad de ámbitos de aplicación, en los cuales, la naturaleza de los datos y los *espacios* donde ellos están definidos adquieren particularidades diferentes, quizás sea uno de los motivos que más justifique la gran diversidad de métodos de detección existentes. Cada uno se ajusta a los datos y a los contextos donde serán aplicados. Esto supone el reto de concebir métodos de detección cada vez más flexibles que puedan ser aplicados en diferentes entornos.

Con el afán de hacer cada vez más eficiente el proceso de detección de *outliers*, se aprecia el interés de los investigadores por aplicar nuevas técnicas a dicho problema. La Teoría de Conjuntos Aproximados (*Rough Sets Theory*), (Pawlak, 1982), (Pawlak, 1991) se ha venido

aplicando, en los últimos años, de manera eficiente en los procesos de *KDD-DM*. Con ello, se ha puesto de manifiesto su capacidad para modelar un amplio conjunto de situaciones reales, así como su efectividad en la solución de problemas de índole muy diversa, en diversos ámbitos de aplicación, todo lo cual avala su *madurez*. En el estudio del *estado del arte* realizado resultó especialmente atractiva una propuesta en la que se presenta un marco teórico para un método de detección de *outliers* basado en la Teoría de *Rough Sets (RS)* (Jiang *et al.*, 2005). Dicho trabajo constituye el primer antecedente de la utilización de este modelo al problema de *outlier detection*. Esta investigación justifica, en cierta medida, que este marco teórico puede ser un punto de partida idóneo para novedosas propuestas que mejoren los algoritmos de detección de *outliers*.

A modo de resumen, puede señalarse lo siguiente:

- Algunos autores señalan (Kollios *et al.*, 2003) que, con la evolución y el desarrollo actual de la ciencia, las líneas de investigación que trabajan en la propuesta de métodos de detección y de teorías que expliquen la aparición de *outliers* en diferentes campos de aplicación, se encuentran a la *cabeza* de dicho desarrollo.

- El principal problema, en lo referido a la detección de *outliers* es que aún no se cuenta con una aproximación universalmente aplicable al mismo. Esto establece la necesidad de seleccionar métodos de detección eficientes para el análisis particular de cada *conjunto de datos*.

- Los métodos de detección de *outliers* deben estar en correspondencia con el desarrollo actual de las nuevas tecnologías de la información.

Universidad de Alicante

Todo lo que acabamos de expresar permite afirmar que aún existen varios *problemas abiertos* para la comunidad científica internacional en relación al problema de la detección de *outliers*. Como consecuencia, siguen apareciendo referencias a nuevos modelos y a nuevos métodos basados en diversos enfoques.

Formulación del Problema

A partir del estudio del *estado del arte* realizado en relación al problema de la detección de *outliers*, cuyo resultado se recoge más extensamente en el **capítulo 2**, se ha identificado como uno de esos *problemas abiertos* dentro de dicho campo, el que a continuación se describe.

El **ámbito del problema** identificado en esta investigación lo podemos ubicar dentro del dominio general de los problemas de *Inteligencia Artificial*, especialmente en los *procesos de búsqueda de información en grandes volúmenes de datos* (*Bases de Datos*) —*procesos de KDD*— y, dentro de éstos, más concretamente, en el importante paso donde se realiza la *minería de datos o DM*, todo ello aplicado al problema general de la detección de *outliers* y en especial, en lo referido a la búsqueda de métodos y algoritmos que mejoren la eficiencia del proceso de detección.

Teniendo en cuenta el contexto de la investigación y los antecedentes analizados, se puede definir el **problema** identificado en la investigación de la siguiente forma: aunque en la actualidad existen varios métodos para la detección de *casos excepcionales* (*outliers*) basados en una gran cantidad de técnicas y con mejores o peores órdenes de *complejidad temporal*, todos ellos requieren que se establezcan previamente unas condiciones particulares que dependen directamente del contexto en el que serán

aplicados y bajo las cuales los algoritmos son capaces de proporcionar un conjunto de *outliers* dentro de un *universo* de datos dado.

Lo anterior hace necesario, en primer lugar, un análisis previo del contexto de dicho *universo* y, con posterioridad a la ejecución de los algoritmos, un análisis del conjunto de *outliers* señalados por los mismos. Este proceso puede tener que repetirse varias veces para lograr que los resultados se adecuen a los intereses iniciales del análisis.

Por lo tanto, aunque esté determinada la *complejidad temporal* de los algoritmos para el *caso peor*, no se puede decir lo mismo del proceso global del análisis de *outliers* que incorpora también el análisis previo y posterior del contexto, así como la necesidad, en la inmensa mayoría de los casos, de ejecutar repetidas veces un mismo algoritmo.

A partir de la **formulación del problema** planteado, se establece una hipótesis que guiará la investigación.

Hipótesis de Partida

Tomando como referencia los antecedentes estudiados se establece como **hipótesis** de partida para solucionar este problema que: es posible desarrollar una nueva teoría basada en la extensión de los conceptos básicos y las herramientas formales que nos proporciona la Teoría de Conjuntos Aproximados (Pawlak, 1982), (Pawlak, 1991) y el *Modelo de Conjuntos Aproximados de Precisión Variable* (*VPRSM*) (Ziarko, 1993), aplicados al problema de la detección de *outliers,* que permita obtener, de forma no supervisada, para cada elemento de un *universo* de datos, la región de valores de los umbrales en la cual dicho elemento es *outlier*. A partir de dicho resultado, es posible

determinar la probabilidad de que cada elemento del *universo* sea un *outlier* en el mismo.

Objetivos

El **objetivo general** planteado en esta investigación, partiendo de la hipótesis establecida, se puede sintetizar en: establecer un método computacionalmente viable que proporcione la probabilidad que tiene cada elemento de un *universo* de datos dado de ser *excepcional*, sin necesidad de haber establecido las condiciones previas —referidas a la determinación de los umbrales que intervienen en el análisis— en función de un contexto específico de aplicación.

Para garantizar el cumplimiento del **objetivo general** planteado, se propone alcanzar, paulatinamente, un conjunto de **objetivos parciales**. Son los siguientes:

- Ampliar el marco teórico de la teoría *Rough Sets,* de forma tal que pueda ser aplicable al *problema de la detección de outliers* —tomando como punto de partida las propuestas previamente realizadas en este ámbito— a partir de soluciones computacionalmente viables a dicho problema.
- Ampliar el marco teórico obtenido a partir del cumplimiento del objetivo parcial anterior, de manera tal que posibilite establecer un método de detección no determinista basado en el *Modelo de Rough Sets de Precisión Variable* (*VPRSM*), que supere las limitaciones de carácter determinista del modelo básico de *RS* en lo relativo a la *clasificación* y que, al mismo tiempo, mantenga la viabilidad computacional.

Universidad de Alicante

Propuesta de Solución

A partir de los resultados anteriores, establecer un marco teórico que permita establecer el diseño de un método que, dada la región de valores de los umbrales que intervienen en el análisis y que constituyen las condiciones previas para las cuales un elemento del *universo* es *outlier,* sea capaz de proporcionar, de forma no supervisada, la probabilidad que tiene cada elemento de dicho *universo* de ser *excepcional*.

El diagrama que se muestra en la **Figura 1-1** ilustra el proceso a partir del cual, de forma gradual, se alcanza el marco teórico general al que se hace referencia en el párrafo anterior.

Inicialmente, tomando como punto de partida el marco teórico del modelo básico de *Rough Sets* (*RSBM*) y de propuestas previamente realizadas (*ODRSBM*), se ampliará el mismo de forma tal que pueda ser aplicable al *problema de la detección de outliers,* a partir de soluciones computacionalmente viables a dicho problema (*ODRSBM-Determinista*). Este resultado se completará con nuevas propuestas teóricas que permitan establecer un método de detección no determinista, basado en el *Modelo de Rough Sets de Precisión Variable (VPRSM),* que supere las limitaciones de carácter determinista del modelo básico de *RS* en lo relativo a la *clasificación* y que, al mismo tiempo, mantenga la viabilidad computacional (*ODRSBM-No Determinista*).

La propuesta anterior deberá contener los elementos teóricos necesarios que servirán de base para la conceptualización de un método (*Método BM*) que permita obtener, de forma no supervisada, para cada elemento del *universo*, la región de valores de los umbrales en la cual dicho elemento es *outlier*. A partir de ello, se establecerá el

marco teórico general que garantizará el cumplimiento del objetivo general planteado en la investigación (*Método Probabilístico*).

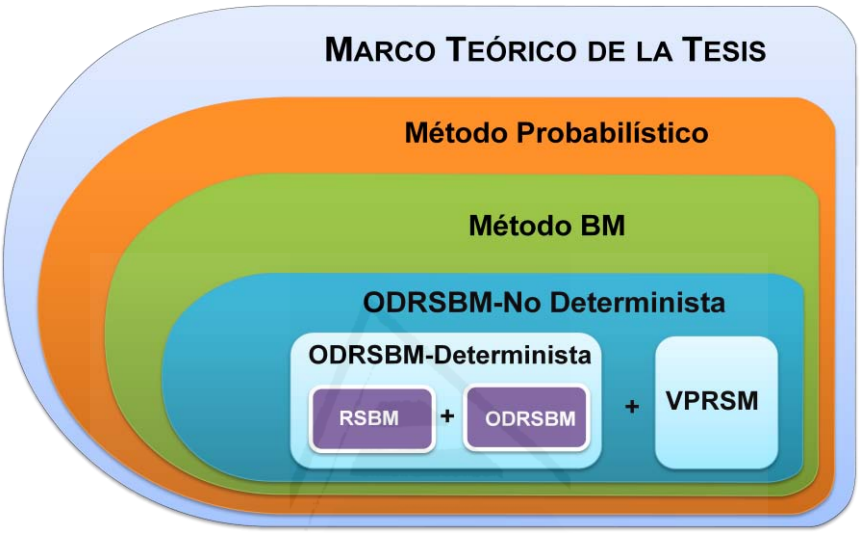


Figura 1-1 Marco teórico para la formalización de un algoritmo computacionalmente viable para la estimación probabilística, de forma no supervisada, de la condición de *outlier* de cada elemento de un universo dado.

Universitat d'Alacant
Universidad de Alicante

En todos los casos, la viabilidad computacional de los métodos propuestos deberá ser validada a partir de algoritmos concretos. En particular, sobre el marco teórico general alcanzado, se formalizará un algoritmo que permita estimar, de forma no supervisada, la probabilidad de cada elemento de un *universo* dado, de ser *outlier* en dicho *universo*.

La **Figura 1-2** muestra el esquema general de solución propuesto.

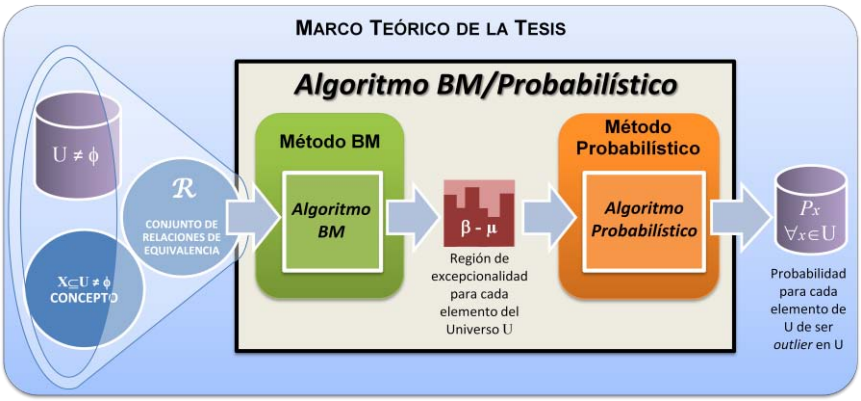


Figura 1-2 Esquema general de solución propuesto, basado en la implementación de un algoritmo computacionalmente viable para la estimación probabilística, de forma no supervisada, de la condición de *outlier* de cada elemento de un universo dado, todo ello a partir del desarrollo del marco teórico creado en la investigación

Metodología

Para alcanzar los objetivos planteados en la investigación, se aplicaron durante todo el desarrollo de la misma algunas de las prácticas consideradas por los científicos como las más adecuadas para exponer y validar los planteamientos teóricos propuestos, así como los resultados de la investigación en general. Esto es, en esencia, la aplicación del *método científico*.

El *método científico* o *camino hacia el conocimiento*, como también suele llamársele; ha sido definido de diversas formas pues no resulta evidente expresar con exactitud la profundidad conceptual del mismo.

Francis Bacon, en su obra "El Avance del Conocimiento" (Bacon, 1605), define el método científico a partir de las siguientes etapas: observación, introducción, hipótesis,

experimentación, demostración o refutación (antítesis) de la hipótesis y tesis o teoría científica (conclusiones).

Esta definición es aplicable a áreas del conocimiento donde sea posible la experimentación. Por tanto, su aplicación resulta factible en las investigaciones relacionadas con la *Ciencia de la Computación* y la *Ingeniería Informática*.

Intentando interpretar la esencia del *método científico*, puede decirse que parte de la observación de los fenómenos naturales. A partir de lo observado, se formulan hipótesis. Las hipótesis se intentan verificar mediante la experimentación. Se analizan los resultados alcanzados y, a partir de ellos, se llega a conclusiones. Si los resultados respaldan las hipótesis, entonces las mismas adquieren validez y grado de tesis o de teoría científica.

En sentido general, lo expresado en el párrafo anterior resume la estrategia seguida en el desarrollo de la presente investigación. En el marco de la misma, se han aplicado los siguientes métodos contemplados en el *método científico*:

Se aplicó el *método hipotético-deductivo* para la elaboración de las hipótesis centrales que guiaron el desarrollo de la investigación. Consecuentemente, el cumplimiento del objetivo general propuesto favoreció el establecimiento de nuevos objetivos (objetivos parciales) de trabajo, a partir de los cuales se derivaron nuevas hipótesis.

La aplicación del método *histórico-lógico* y el *dialéctico*, facilitó el estudio de trabajos previos que resultaron atractivos. De ellos se extrajeron sus aspectos positivos y negativos. Lo que se derivó de este estudio sirvió de antecedente para el establecimiento del objetivo general y los objetivos parciales planteados en la investigación.

Entre los *métodos lógicos* aplicados, el *analítico–sintético* permitió estructurar el problema central de la investigación

en varios subproblemas. A partir del planteamiento teórico inherente a cada uno de ellos, se llegó al nivel de concreción y realización práctica deseado, de forma tal que se hizo posible la integración de todos para dar solución al problema general formulado al comienzo de la investigación

En la revisión bibliográfica realizada, durante todo el desarrollo de la investigación, se siguió el método *inductivo-deductivo.* Esta revisión permitió identificar las tendencias actuales en relación al problema de la detección de *outliers,* así como los problemas que aún siguen estando abiertos dentro de este campo. Todo esto permitió plantearnos retos en la investigación a partir de los objetivos propuestos.

La aplicación del método de la *comparación-clasificación* posibilitó establecer una comparación de métodos a partir de la cual se resumen las ventajas de nuestra propuesta en relación a otras ya existentes, todo lo cual permitió identificar sus principales aportaciones.

El método de *idealización-modelación* se aplicó en cada una de las ampliaciones que, de forma incremental, fue necesario realizar del marco teórico formal inicialmente existente. Los modelos teóricos establecidos en cada caso, constituyen la base teórico-matemática de las distintas versiones del método de detección propuesto, cuya implementación se concretó con algoritmos que permitieron comprobar la viabilidad computacional de los mismos. En el establecimiento de dichos modelos, se aplicaron varios *métodos matemáticos*, que permitieron demostrar los resultados teóricos propuestos.

En todo el desarrollo del trabajo ha estado presente el estudio o la aplicación de diversos *métodos estadísticos*. Su aplicación fue decisiva en la *generación sintética* de *conjuntos de datos* sobre los que se validaron los resultados alcanzados. Para garantizar el cumplimiento del objetivo

general de la Tesis, resultó indispensable la aplicación de técnicas estadísticas.

El *método experimental* se aplicó en la validación de los resultados alcanzados y en algunos casos, permitió también fundamentar los estudios comparativos realizados entre los métodos de detección de *outliers* existentes y nuestras propuestas.

El *método coloquial* se siguió para la presentación y discusión de los resultados de la investigación en diferentes contextos. La aplicación del *método de la entrevista* permitió la interacción con otros especialistas, lo que favoreció nuestra integración en grupos de investigación multidisciplinarios.

Plan de Trabajo

Para el cumplimiento de los objetivos propuestos y, por tanto, para desarrollar la solución al problema general planteado, se establece el siguiente **plan de trabajo**:

- Estudiar el *estado del arte* en relación al problema de la detección de *casos excepcionales* centrándonos especialmente en la aplicación de la Teoría de Conjuntos Aproximados (*RS*) a dicho problema.
- Establecer un método, computacionalmente viable, para la detección de *outliers* basado en el Modelo Básico de Conjuntos Aproximados (*RSBM*)
  - Análisis de propuestas previas que suponen los antecedentes del presente trabajo de investigación.
  - Establecimiento de un marco teórico que permita superar las limitaciones de los trabajos previos.

Universitat d'Alacant
Universidad de Alicante

- Diseñar e implementar un algoritmo basado en el método propuesto a partir del cual se valide la viabilidad computacional del mismo.
- Validación de los resultados alcanzados.
- Conclusiones.

- Establecer un método que mantenga la viabilidad computacional del anterior y además sea no determinista en lo referido a la *clasificación*.
  - Estudio de los conceptos básicos de *VPRSM*.
  - *RSBM* frente a *VPRSM.*
  - Ampliación del marco teórico introduciendo en él aspectos del *Modelo de Conjuntos Aproximados de Precisión Variable* (*VPRSM*).
  - Diseñar e implementar un algoritmo basado en el método propuesto a partir del cual se valide la viabilidad computacional del mismo.
  - Validación de los resultados alcanzados.
  - Conclusiones.

- Crear un método que permita, de forma no supervisada, establecer la región de valores de los umbrales que intervienen en el análisis en la cual cada elemento de un *universo* dado será considerado *excepcional* dentro del mismo. A partir de dicha región, establecer un método general que determine, de forma no supervisada también, la probabilidad que tiene cada elemento del *universo* de ser *excepcional* dentro del mismo. Ambos métodos deben ser computacionalmente viables.
  - A partir del marco formal alcanzado hasta el momento, introducir en el mismo nuevos aspectos

Universitat d'Alacant
Universidad de Alicante

teóricos a partir de los cuales se garanticen los resultados previstos.

- Diseñar e implementar algoritmos que validen la viabilidad computacional de los métodos propuestos.
- Validación de los resultados alcanzados.
- Conclusiones.

Para facilitar el seguimiento, se ha estructurado el resto del documento en función del *Plan de Trabajo* propuesto de forma tal que en el Capítulo 2, se exponen los aspectos más significativos obtenidos como resultado del estudio del *estado del arte* realizado en relación al problema de la detección de *outliers*; en el Capítulo 3, se hace un análisis crítico de una propuesta concreta de un método de detección de *outliers* basado en *RS,* cuyo marco formal resultó ser lo más cercano a nuestro interés y constituía el primer antecedente de la aplicación de la Teoría de *RS* al problema de la detección de *outliers*. Teniendo en cuenta las limitaciones que se señalan a la propuesta antes mencionada; en el Capítulo 4, se establece un marco formal a partir del cual se propone un método de detección de *outliers*, computacionalmente viable, basado en el modelo básico de *RS*. Atendiendo al carácter determinista de este modelo; en el Capítulo 5 se introducen nuevos aspectos teóricos que amplían el marco formal existente y, a partir de ellos, se propone un método de detección, basado en *VPRSM,* que mantiene la viabilidad computacional del anterior y, además, es no determinista; los resultados expuestos en los capítulos precedentes sirven de antecedentes para la concepción de un método que determina, mediante la aplicación de técnicas estocásticas, la probabilidad de cada elemento de un *universo* de datos dado de ser *excepcional* en dicho *universo,* desarrollado,

fundamentalmente, en el Capítulo 6; en el Capítulo 7 se establece una comparación entre los diferentes métodos propuestos en nuestra investigación y se comparan los mismos con otros métodos de detección existentes; finalmente, incorporamos un último capítulo de conclusiones (Capítulo 8) en el que se resaltan las principales aportaciones de la investigación, así como las líneas futuras de trabajo.

Universitat d'Alacant
Universidad de Alicante

Capítulo 2

Estado del Arte

Los modelos estadísticos fueron los primeros que trataron el problema de la detección de *outlier*s. Una buena parte de los métodos de detección, incluso los más recientes, incorporan, en mayor o menor medida, algún aspecto heredado de algún método estadístico. Las técnicas estadísticas, en general, están estrechamente ligadas a las técnicas de *minería de datos*, por tanto, dedicaremos especial atención a describir los aspectos esenciales que caracterizan a dichos métodos (Hair *et al.*, 1999) dentro de esta importante rama de la Matemática.

Como ya se ha señalado, los *casos excepcionales* pueden ser considerados tanto *útiles* como *problemático*s. Desde un punto de vista estadístico, cuando los *casos excepcionales* son *útiles* pueden ser un indicativo de las características de un segmento de la población que se llegaría a descubrir en el curso normal del análisis.

Por el contrario, los *casos excepcionales problemáticos*, no son representativos de la población y están en contra de los

objetivos del análisis. Pueden llegar a distorsionar, seriamente, los *test* estadísticos.

Cuando se detectan *casos excepcionales*, es necesario que el investigador examine los datos con el fin de averiguar el tipo de influencia que ejercen dichos elementos sobre los mismos.

Los *casos excepcionales* se deben situar en un contexto adecuado para evaluar la influencia de las observaciones individuales y determinar si esta influencia es *beneficiosa* o *perjudicial*.

Principales Métodos Estadísticos para la Detección de *Outliers*

En Estadística, los *casos excepcionales* pueden identificarse desde tres tipos de perspectivas: *univariante*, *bivariante* y *multivariante*. El investigador debería utilizar cuantas perspectivas sean necesarias, buscando una consistencia entre los métodos de identificación de tales casos. A continuación, se detallan los procesos resultantes de cada una de ellas.

Universidad de Alicante

Detección Univariante

La perspectiva *univariante* está asociada a la *distribución Normal*, es decir, examina la distribución de las observaciones y selecciona como *casos excepcionales* aquéllos que caigan fuera de los rangos de la *distribución* establecida.

Es un método de detección típicamente *paramétrico*. En él, se asume una *distribución* (el enfoque típico asume la

Normal) que depende de ciertos parámetros. En el caso de la *Normal* serían: la *media* (μ) y la *varianza* (σ^2).

Un aspecto fundamental del método es el establecimiento de un umbral para la clasificación de un dato como *caso excepcional*. El enfoque típico basado en la *Normal* considera la conversión de los datos originales a valores estándar, es decir, que tengan una *media* 0 y una *desviación estándar* de 1. Esto se logra centrando los datos con respecto a su *media* y reduciéndolos, dividiendo por la *desviación típica* (σ). La probabilidad de superar ciertos valores críticos (definidos por el investigador convenientemente) es computable con facilidad.

Dado que los valores están expresados en un formato normalizado, se pueden realizar fácilmente comparaciones entre las variables. Las pautas sugeridas identifican como excepcionales aquellos casos con valores absolutos estándar de 2,5σ –multiplicar por 2,5 la desviación típica que es una medida de cuanto se separan los datos de la media– o superiores. En otras ocasiones, se pueden usar valores entre 3σ y 4σ, lo cual establece una mayor restricción a la hora de clasificar un *outlier*. Esto supone más seguridad en la *clasificación*.

Universidad de Alicante

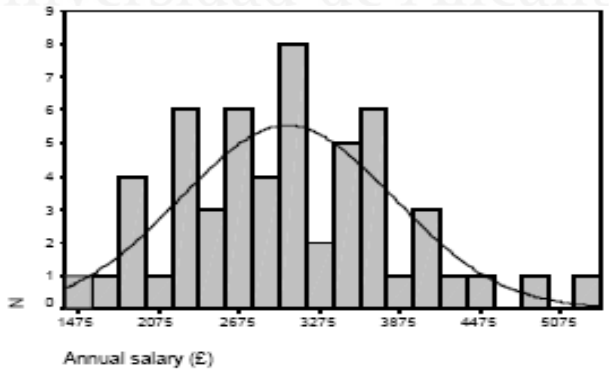


Figura 2-1 Ejemplo de distribución normal

En la **Figura 2-1** se observa un histograma para datos que representan el salario anual de los trabajadores de un cierto país en un año. Dichos datos distribuyen *normalmente*.

La **Figura 2-2** ilustra un *criterio de clasificación* de *outliers* a partir del uso de la distribución *Normal* de los datos. En este caso se supone que los valores están expresados en un formato estandarizado.

En cualquier caso, el investigador debe darse cuenta de que es posible que un cierto número de observaciones puedan caer fuera de esos rangos de la *distribución*. En tal caso, debería esforzarse en identificar sólo aquellas observaciones verdaderamente distintivas y designarlas como *casos excepcionales*.

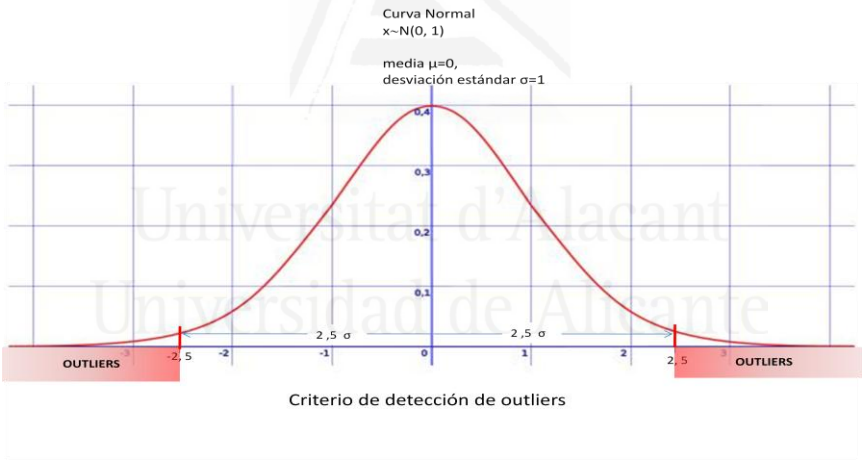


Figura 2-2 Criterios de detección de *outliers* a partir de la distribución Normal

En (Tukey, 1977) se propone el uso de unos gráficos llamados de *Box & Whiskers* (*B&W*) para ser usados en la detección de *outliers*. Su uso en el estudio de los rangos o características de la distribución de valores de una variable seleccionada se basa en el trazado del *B&W* para las observaciones tenidas en cuenta. Las mediciones de la

variable (por ejemplo, *salario*) se agrupan en categorías definidas por el experto. Por ejemplo, *salario según los diferentes perfiles laborales* (profesionales, obreros, ejecutivos, etc.) y para cada uno se traza el gráfico correspondiente. En estos gráficos bidimensionales se fija la ubicación de, al menos, una medida de la tendencia central de los datos. Por lo general, se fijan la *mediana* o la *media* y se da una valoración de su variación utilizando:

- El *recorrido* (diferencia entre el máximo y el mínimo de las observaciones que se tienen en consideración).
- Los *cuartiles* del 25%, 50% y el 75% (generalmente).
- Los errores *normales* o las desviaciones *normales*.

Estos se computan para cada grupo y se presentan en una *Box* seleccionada. Los posibles *outliers* son aquellos que aparecen fuera del rango considerado como aceptable.

Al detectar que un punto está fuera de los límites, este es considerado como un posible *outlier* para un *grupo* concreto de datos. Para ilustrar lo que podrían ser tipos de *grupos* se presenta el siguiente ejemplo: en *grupo* 1: *hombres* y *grupo 2*: *mujeres*. Este proceso se sigue para cada *grupo*.

En la **Figura 2-3** se utiliza la *media* como medida de localización. Los grupos quedan establecidos por el *sexo.* Para cada grupo se analiza la variable *Wishes Controled Consciously* (*WCC*) medida en una investigación psicológica. Dicha variable se mide en una escala 100-150. Los límites dados en la figura fijan el intervalo donde deben encontrarse los datos que no se consideran como posibles *outliers* (dentro de la *Box* o en alguno de los *Whiskers*).

Algunos sistemas computacionales comerciales (por ejemplo, *StatGraph*) diseñados para la obtención de gráficos de este tipo permiten visualizar los *outliers* de acuerdo al análisis estadístico de los datos tenidos en cuenta. El

gráfico que se observa en la **Figura 2-4** ilustra lo antes dicho. En él, tras el análisis, se muestran los *outliers*. Estos quedan identificados por los puntos de mayor intensidad en el color.

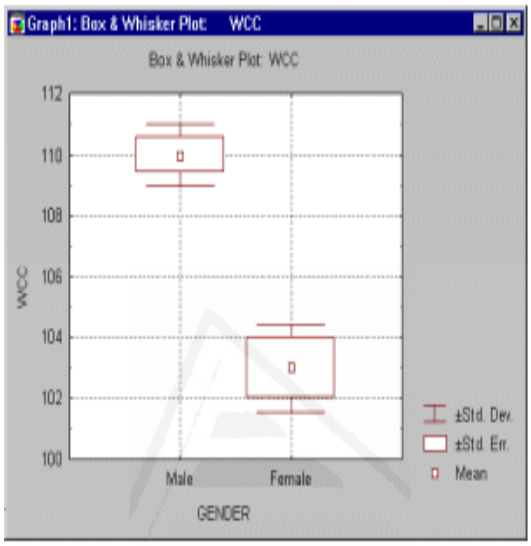


Figura 2-3 Ejemplo de un gráfico de *Box & Whiskers*

Universitat d'Alacant
Universidad de Alicante

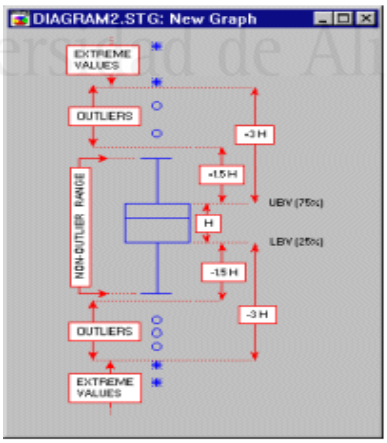


Figura 2-4 Gráfico de *B&W* mostrando la detección de *outliers*

Muchas variantes de estos gráficos aparecen en la literatura especializada de análisis de datos.

Detección *Bivariante*

La detección *bivariante* se caracteriza porque en ella se evalúan conjuntamente pares de variables. Los resultados de la misma, por lo general, se representan en un *diagrama de dispersión*. En dichos gráficos, los casos que caigan manifiestamente fuera del rango del resto de las observaciones se identifican como puntos aislados y en tal caso, se les considera como posibles *outliers*.

Para ayudar a determinar el rango esperado de las observaciones, se suele superponer sobre el diagrama una *elipse*. Dicha *elipse* representa un *intervalo de confianza* específico para una *distribución normal bivariante*. Con ello se proporciona una representación gráfica de los *límites de confianza*, lo cual facilita la identificación de los *casos excepcionales* (**Figura 2-5**).

Existen otros tipos de *gráficos de dispersión* como por ejemplo, el *gráfico de influencia*. En éste, el tamaño de cada *punto* varía según su *influencia* en las relaciones que intervienen en el análisis. Los métodos, a partir de los cuales se genera este tipo de gráficos, proporcionan una cierta evaluación de la *influencia* de cada observación, lo cual sirve de complemento al designar los *casos excepcionales*.

En el caso de la *detección bivariante* también se pueden usar los gráficos de *Box* & *Whiskers*, vistos en otra dimensión. La **Figura 2-6** muestra una interpretación *bivariante* de los mismos. Por su parte, la **Figura 2-7** ilustra un ejemplo particular de este tipo de gráficos bajo dos

niveles de temperatura y en dos instantes diferentes (dos grupos) del *polysulfido* 6 y 7.

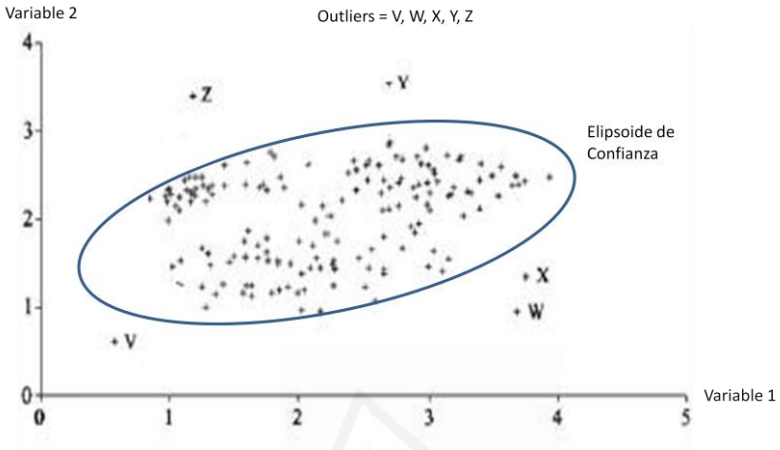


Figura 2-5 Diagrama de dispersión

Un aspecto importante a resaltar es el hecho de que al aumentar la dimensión se hace más difícil la interpretación de los datos y por tanto la determinación de los *outliers*.

Detección *Multivariante*

Desde la perspectiva *multivariante* de identificación de *casos excepcionales*, la evaluación de cada observación se hace a través de un conjunto de variables $x_1,\ldots,x_k$ ($k \geq 2$).

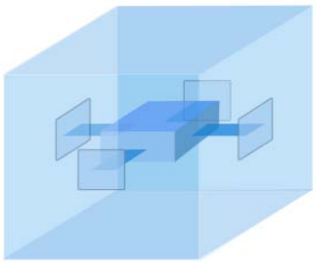


Figura 2-6 Interpretación *bivariante* de los gráficos de *B&W*

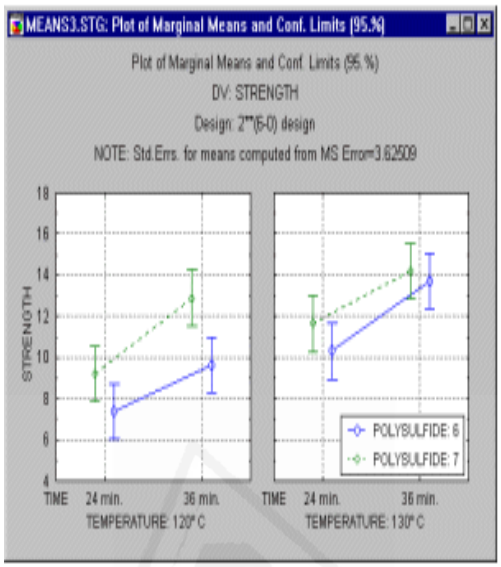


Figura 2-7 Ejemplo de gráficos de *B&W* para el caso *bivariante*

Por lo general, en la mayoría de los análisis *multivariantes* se tienen en consideración más de dos variables, por tanto, el investigador necesita una forma objetiva de medir la posición multidimensional de cada observación relativa a un punto común. En gran medida, esto justifica el hecho de que la mayoría de los métodos estadísticos de detección de *outliers* que trabajan desde esta perspectiva estén basados en algún criterio de distancia.

La *medida D^2 de Mahalanobis* constituye el criterio de *distancia* más popular para la detección de *outliers* sobre *distribuciones multivariadas*. No es más que una medida de la *distancia* de cada observación en un *espacio multidimensional* con respecto al *centro medio* de las observaciones, y proporciona una medida común de *centralidad multidimensional*.

Desde el punto de vista probabilístico, a partir de D^2 se pueden deducir propiedades estadísticas que permiten hacer pruebas de hipótesis para el estudio de *outliers*. Valores significativamente altos de dicha medida indican la presencia de un posible *outlier*.

Dada la naturaleza de los *test* estadísticos, se debe usar un nivel muy conservador para determinar el valor de los umbrales a partir de los cuales se designe a un dato como *caso excepcional*. Por ejemplo, si el D^2 asociado a un dato es mayor que el valor de un umbral dado, entonces dicho dato es considerado un posible *outlier*.

La *métrica de Mahalanobis* para datos procedentes de una *distribución multivariada* p-dimensional parte de la observación de n vectores p-dimensionales de la forma:

$$x_i = \begin{bmatrix} x_{1i} \\ \vdots \\ x_{pi} \end{bmatrix} i = 1, \dots, n \qquad (1)$$

Como ya hemos expresado, la métrica o *distancia de Mahalanobis* se denota por D^2 y la fórmula a partir de la cual se expresa la misma para una cierta observación i es la siguiente:

$$MD_i = \left((x_i - t)' C^{-1} (x_i - t)\right)^{1/2} \quad i=1,\dots,n \qquad (2)$$

Dicho valor expresa la *distancia* entre la observación i y el promedio de los datos expresado por el valor de t. Tal valor t, es una medida de posición escogida. Generalmente, la *media aritmética multivariada*, o sea, un vector cuyos valores son la *media* de las componentes de los n vectores que están en observación.

Universidad de Alicante

Tenidos en consideración todos los vectores x_i de la forma $\begin{pmatrix} x_{i1} \\ x_{i2} \\ \cdot \\ \cdot \\ \cdot \\ x_{ip} \end{pmatrix} \longrightarrow t = \begin{pmatrix} \sum \frac{x_{i1}}{n} \\ \sum \frac{x_{i2}}{n} \\ \cdot \\ \cdot \\ \cdot \\ \sum \frac{x_{ip}}{n} \end{pmatrix}$ (3)

$C = [\ S_{ij}\ ]_{pxp}$, es la *matriz de covarianzas muestral*, donde:

$$S_{ij} = \frac{\sum_{h=1}^{n} (x_{ih} - \bar{x}_i)(x_{jh} - \bar{x}_j)}{n-1} \quad (4)$$

$\bar{X}$:es la *media*

Los *outliers multivariados* pueden ser caracterizados como observaciones con un valor muy grande de *MDi*.

A este procedimiento se le señala como principal limitación el hecho de que la métrica D^2 debe ser estimada por un procedimiento que pueda producir medidas fiables para la detección de *outliers*.

Este método basado en la *distancia de Mahalanobis* y en general todos los métodos estadísticos, dada su naturaleza precisamente estadística, pueden calificar de *outliers* ciertos datos, sólo por el hecho de estar por encima del valor crítico dado por el percentil y sin embargo, pueden no serlo.

Recientemente, y manteniendo como base el principio teórico de los métodos basados en *distancias*, se han propuesto implementaciones originales de dichos métodos

para el caso *multivariado*. Estos métodos se comportan de forma eficiente sobre *conjuntos de datos k*-dimensionales ($k \geq 2$) de gran tamaño. Casi todas estas implementaciones se basan en la siguiente idea para clasificar un objeto como *outlier*:

«Un objeto *O* en un *conjunto de datos U* es un *outlier* en *U* si existe un porcentaje preestablecido de objetos en *U* que se encuentran a una *distancia* mayor o igual que *D* de *O*, para un valor *D* dado» (Knorr & Ng, 1997), (Knorr & Ng, 1998), (Knorr & Ng, 1999), (Knorr *et al.*, 2000).

Otros enfoques en esta línea se basan en la *distancia* de un objeto a sus *k-ésimos vecinos más cercanos*. A partir de esto se establece el concepto de: *clustering* basado en criterios de *distancia***.**

El método *k-NN* (*K-Nearest Neighbor*) fue desarrollado por Fix y Hodges en la década de los 50 (Fix & Hodges, 1951). Su núcleo es la estimación de una *función de densidad* que sirve para discriminar los *outliers* de las observaciones aceptables. Para ello se estima dicha función para cada categoría sujeta al análisis.

Suponiendo que existen *s* categorías, la *función de densidad conjunta* asociada a una categoría particular, C_j, $1 \leq j \leq s$, se denota de la siguiente forma: $f(x/Cj)$, donde x es un vector, según se definió en (1), en el cual están representadas las variables sujetas a estudio.

Se establece un conjunto de atributos que caracterizan a los individuos de la clase. A cada atributo particular se asocian determinadas categorías.

El siguiente ejemplo ilustra los conceptos antes expuestos.

Ejemplo:

Atributo_1: *hombre - provincia de procedencia*

Atributo_2: *mujer - provincia de procedencia*

Cada valor particular de algún atributo define una clase o categoría, por ejemplo, *hombres_de_La_Habana* sería una clase asociada al *Atributo_1*.

Cada *hombr*e tiene un *salario*, una *edad*, un *nivel de escolaridad*, etc. Estas son, por ejemplo, algunas de las variables que pueden aparecer representadas en el vector *x*. Cada instancia de una clase tendrá su representación en un vector particular.

Hay dos problemas esenciales asociadas al método *K-NN*:

- La elección de la *distancia* o *métrica*.
- La elección del número de *vecinos* (*k*) a considerar.

La métrica más elemental usada como *criterio de distancia* es la *euclidiana*.

$$d(x, y)=(x-y)^T(x-y) \quad (5)$$

Esta métrica puede causar problemas si las variables han sido medidas en unidades muy distintas (por ejemplo, la *edad* en años y los *salarios* en miles de euros) entre sí.

Algunos prefieren reescalar los datos antes de aplicar el método y usar la variable tipificada. Lo cual no es más que un valor *z* calculado a través de la siguiente fórmula:

$$z = (x- \text{ media de } x) \ / \ \text{desviación típica de } x \quad (6)$$

donde, *x* es un vector según (1) y por tanto, *z* será también un vector que se obtiene según la fórmula expresada.

Otra *distancia* bien usada para los fines que acabamos de mencionar es la *Manhatan o City Block* definida por

$$d(x, y)=|x-y| \quad (7)$$

En algunos casos, la propia *distancia de Mahalanobis* es usada para tales fines.

Las aproximaciones algorítmicas al método *K-NN* parten del conocimiento que se tiene acerca de un juego de datos, consistente en n puntos (cada uno de los cuales representa uno de los vectores x_i). Algorítmicamente se sigue el principio básico de asumir que las observaciones que están próximas, a partir de alguna métrica dada, son clasificadas en un mismo grupo. Bajo este enfoque, al intentar clasificar una muestra x desconocida, juega en ello un papel decisivo la manera en que está(n) clasificado(s) su(s) vecino(s) más cercano(s) x_j. En tal caso, se supondrá que x tiene la misma *clasificación* que él (ellos).

Este es un procedimiento de *clasificación no paramétrico* que se establece, fundamentalmente, a partir de dos reglas:

- Regla del *NN* (del *vecino más cercano*).
- Regla de los *K- NN* (de los *k vecinos más cercanos*).

La Regla del *NN,* o del *vecino más cercano,* se basa en el siguiente criterio:

La observación j es el *vecino más cercano* a x si:

$$d(x, x_j) = min\ d(x, x_i) \quad i, j = 1, \ldots, n \tag{8}$$

Esta regla clasifica a x en la categoría θ_n dónde x_n es el *vecino más cercano* a x. θ_n es la clase a la cual pertenece x_n, mientras que θ es la clase verdadera a la que pertenece x. Si θ_n no es igual a θ , entonces se ha cometido un error en la *clasificación*.

La **Figura 2-8** muestra un ejemplo de la regla del *NN* en el que hay representadas dos clases:

clase_1: θ_1 (los triángulos)

clase_2: θ_2 (los cuadrados)

Universitat d'Alacant
Universidad de Alicante

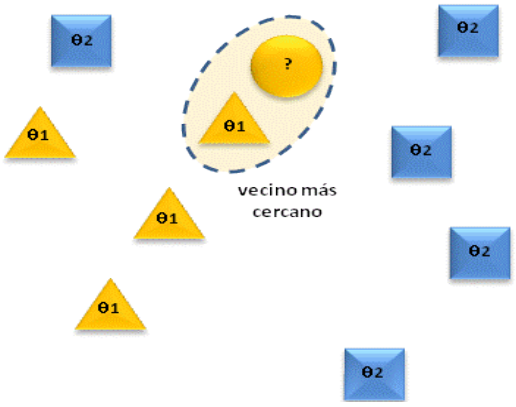


Figura 2-8 Ejemplo de aplicación de la regla del *NN*

El *círculo* representa la muestra desconocida x y como su *vecino más cercano* pertenece a la clase θ_1, entonces a él se le clasifica también como miembro de dicha clase. Es decir, se supone que el *círculo* pertenece a la clase de los *triángulos* porque está más cerca de un *triángulo*.

Si el número de puntos pre-clasificados es grande, tiene sentido usar no sólo al *vecino más cercano,* sino analizar los grupos y clasificar la observación a partir de la clase que tenga los *k vecinos más cercanos* a ella.

La Regla de los *K- NN* o de los *k vecinos más cercanos* en esencia expresa lo siguiente:

Si tenemos k observaciones que pertenecen a una clase θ_n y constituyen los *vecinos más cercanos* a x en dicha clase, entonces se dice que $x \in \theta_n$ siempre que la suma de las *distancias* de x a esos k elementos de θ_n sea la mínima entre todas las sumas de igual tipo para las restantes clases, es decir, se suman las *distancias* de x a los k elementos *más cercanos* de cada clase y se clasifica dentro

de la clase para la cual dicha suma es la mínima. El siguiente ejemplo, ilustra este tipo de regla.

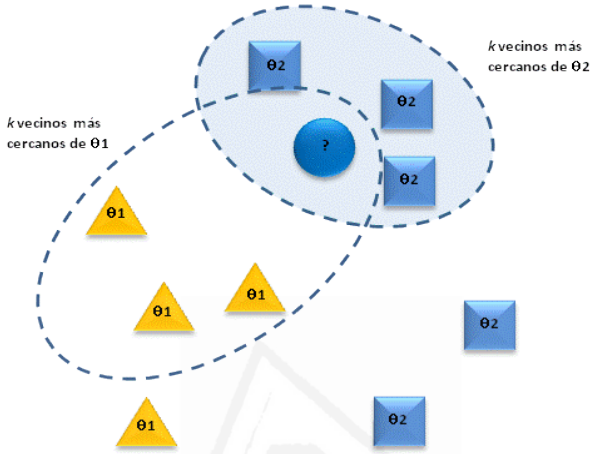


Figura 2-9 Aplicación de la regla de los *k-vecinos más cercanos* (*k*=3)

La **Figura 2-9** muestra un ejemplo de aplicación de la regla de los *K-NN*, para *k*=3. Igual que en el ejemplo anterior, en éste se han considerado dos clases: θ_1 (los *triángulos*) y θ_2 (los *cuadrados*). El *círculo* representa la muestra desconocida *x*. Como el grupo de los tres *vecinos más cercanos* a él proviene de la clase θ_2, entonces se clasifica al *círculo* como miembro de dicha clase.

El problema teórico asociado a los métodos que acabamos de describir está referido a la probabilidad de cometer un error en la *clasificación* y se aborda estadísticamente a partir de enfoques probabilísticos.

En la práctica, en el problema del *vecino más cercano* se tiene un juego de datos en el espacio *p*-dimensional. Los *puntos* que componen dicho juego de datos son procesados para que estén listos estadísticamente para aplicar sobre

ellos alguna de las variantes de las reglas del *vecino más cercano*. Dado cualquier punto *q*, se busca el *punto más cercano* o los *k puntos más cercanos* a él. El *punto q* se considera un *outlier* si no tiene *k vecinos suficientemente cercanos* a él. Para ello se fija un valor máximo de la *distancia* a partir de la cual se considera que existe un *vecino suficientemente cercano*. Si no hay *k* individuos con *distancia* menor que ese valor fijado, entonces el *punto* se clasifica como *outlier*.

Existen diferentes enfoques del algoritmo de los *k vecinos más cercanos* (*K-NN*). En todos se usa una métrica adecuada para el cálculo de la *distancia* entre vecinos. Por ejemplo, la *distancia euclidiana* o la *distancia de Mahalanobis*. Existen varias mejoras del algoritmo básico de *K-NN*. Algunas de las publicaciones más recientes que citan el uso de esta técnica, son (Ramaswamy *et al.*, 2000), (Angiulli & Pizzuti, 2002), (Bay & Schwabacher, 2003), (Angiulli & Pizzuti, 2005), (Angiulli *et al.*, 2006), (Angiulli *et al.*, 2007).

Universitat d'Alacant
Universidad de Alicante

Aspectos Básicos para la Detección de *Outliers* con Métodos Estadísticos

Hasta aquí hemos dado una panorámica general de los principales métodos estadísticos de detección de *outliers,* por tanto, estamos en condiciones de sacar algunas conclusiones que, desde el punto de vista estadístico, son relevantes a la hora de abordar el problema de la detección de *outliers.* Centrémonos en los siguientes aspectos que resultan esenciales bajo dicho enfoque.

La Calificación de un Caso de *Outlier*

Cuando las observaciones candidatas a ser consideradas como *casos excepcionales* han sido identificadas por métodos *univariantes*, *bivariantes* o *multivariantes*, el investigador debe seleccionar aquellas observaciones que demuestran una unicidad real en comparación con el resto de la población. Deberá abstenerse de calificar muchas observaciones de *excepcionales* y no deberá caer en la tentación de eliminar aquellos casos que no son consistentes con los casos restantes por el hecho de ser simplemente diferentes.

La Descripción de los *Outliers* y su Especificación

Una vez que se han identificado los casos potencialmente valorados como *excepcionales*, el investigador deberá identificar cada uno de ellos y examinar cuidadosamente si los datos asociados a las variables responsables son *casos excepcionales* o no. Además de estos exámenes visuales, el investigador puede emplear también técnicas *multivariantes* como el *análisis discriminante* o la *regresión múltiple* para identificar las diferencias entre los *casos excepcionales* y las otras observaciones. Este análisis se debe continuar hasta determinar el(los) aspecto(s) de los datos que distingue(n) al *caso excepcional* del resto de las observaciones.

El Mantenimiento o la Eliminación de los *Outliers*

Una vez que se han identificado, especificado y catalogado los *casos excepcionales*, el investigador debe decidir entre mantenerlos o destruirlos. Hay muchas opiniones al respecto. El criterio más generalizado es que deben mantenerse a menos que existan pruebas convincentes de que son verdaderas aberraciones y por tanto, no son datos

representativos de la población bajo observación. Pero en caso de que, a pesar de su *excepcionalidad*, representen un segmento de la población, entonces deben mantenerse dentro del *conjunto de datos*. Si los *casos excepcionales* son eliminados, el investigador corre el riesgo de mejorar el análisis pero a la vez, limitar su generalidad. Si los *casos excepcionales* son *problemáticos* al ser aplicada una técnica particular, en muchas ocasiones pueden ser mejorados de forma tal que se ajusten al análisis sin distorsionarse significativamente.

Estas consideraciones se realizan teniendo en cuenta que en muchas ocasiones los datos serán utilizados para evaluar determinadas *reglas de decisión* que pueden ser afectadas por los *outliers*. Por sólo citar un ejemplo, hagamos referencia al caso de la *media*, que puede llegar a ser poco representativa ante la presencia de una observación aberrante. Consideremos el caso de 4 personas que son entrevistadas en relación a su salario mensual y supongamos que los salarios de esos 4 individuos son los siguientes: 1.000 euros, 1.500 euros, 1.200 euros y 1.300 euros, respectivamente. En este caso, el promedio salarial de esas 4 personas sería de 1.250 euros mensuales. Si se agrega una quinta persona con un salario de 100.000 euros, la *media*, calculada ahora en base a 5 salarios, sería poco representativa para 4 de los 5 entrevistados pues el valor calculado daría 21.000 euros. En tales casos, el estadístico deberá tomar en cuenta si el supuesto *outlier* es un caso a considerar: ¿Realmente esa persona tiene ese salario? o ha ocurrido un error en la medición. Muchas veces, se elimina esa observación sin hacer un análisis como el recomendado, lo cual es erróneo.

Principales Críticas a los Métodos Estadísticos de Detección de *Outliers*

A pesar de la primicia que tienen los estadísticos en cuanto a la aplicación de métodos de detección de *outliers* al análisis de los datos, las nuevas técnicas de Inteligencia Artificial, que en los últimos años se vienen aplicando sobre los voluminosos y complejos *conjuntos de datos* del mundo actual a partir de la óptica del *KDD-DM,* hacen que muchos de los métodos estadísticos de detección de *outliers* resulten obsoletos a tales efectos. Algunos de los aspectos que ponen de manifiesto esta afirmación son los siguientes:

- Los modelos estadísticos son, generalmente, apropiados para el procesamiento de *conjuntos de datos* con valores *reales continuos*, *cuantitativos* o, al menos, datos *cualitativos* con valores *ordinales*. En la actualidad, cada vez más, se necesita procesar datos expresados de manera *categórica* (no *ordinales*). Esto limita considerablemente la aplicación de los métodos estadísticos. Por lo general, para vencer tales limitaciones se hace necesario realizar complejas transformaciones de los datos, con lo cual aumenta considerablemente el tiempo de procesamiento de los mismos.
- En (Otey, *et al.*, 2005c) se señala que los *métodos paramétricos* suponen que los datos deben seguir una *distribución paramétrica* (como caso típico, *univariada*). En tales casos, los métodos no funcionan correctamente en contextos *multivariados*. Como solución a estos aspectos, se usan los *métodos no paramétricos* basados en *distancias*, *clustering* o *densidades*. De los cuales, los basados en distancias ya han sido comentados.
- Dentro de la referida publicación, también se dice que la mayoría de los métodos basados en *distancias* tienen

como limitación fundamental el hecho de que la *complejidad temporal* de los mismos es cuadrática para el *caso peor*. Este aspecto adquiere especial importancia si se trabaja con *conjuntos de datos muy grandes* o *dinámicos*.

- Otro problema a considerar en el uso de métodos estadísticos para la detección de *outliers,* es el incremento de la dimensionalidad de los datos. Esto puede llevar también a un aumento del tiempo de procesamiento y a una distorsión de la *distribución* de los mismos. El problema de la alta dimensionalidad de las actuales bases de datos es otro de los factores que implica que determinados métodos de detección no funcionen correctamente. Al aumentar la dimensión del *conjunto de datos* la efectividad de ciertos algoritmos puede verse afectada. Por ejemplo, el concepto de *distancia* en un espacio de dimensión *k*, no es el mismo que en un espacio de dimensión *k+1*.
- Siguiendo un razonamiento similar, el problema de la alta dimensionalidad puede también incidir sobre la efectividad de los métodos que se basan en criterios de *proximidad* entre los datos. En contextos de alta dimensionalidad, por lo general, los datos están *esparcidos* (cada *tupla* puede tener atributos en el orden de los cientos) y en tal caso, la noción de *proximidad* no resulta tan significativa. Dentro de este contexto, diferentes autores (Aggarwal, 2007), (Aggarwal & Yu, 2008) han trabajado en el desarrollo de técnicas y enfoques encaminados a la solución de este problema. Algunas de ellas se centran en los atributos más sobresalientes, para con ello lograr que los algoritmos manejen un número menor de *dimensiones*. Otras técnicas, se basan en el uso de algoritmos que permiten proyectar los datos en un espacio de menor dimensión (métodos de detección basados en *subespacios*).

Universitat d'Alacant
Universidad de Alicante

- Otro problema asociado a la aplicación de métodos estadísticos, cuando se trabaja con *conjuntos de datos* de alta dimensionalidad, es el siguiente: en las técnicas estadísticas de detección de *outliers*, por lo general, los datos son modelados usando una *distribución estocástica (*aleatoria), y calificar de *outlier* a un dato, depende en gran medida de su relación con dicho modelo. Al aumentar considerablemente la *dimensión* del *conjunto de datos*, resulta muy difícil estimar *distribuciones multidimensionales* de los mismos, es decir, generalmente es muy difícil encontrar el modelo adecuado (Hodge & Austin, 2004). En tales casos, los métodos de estimación pueden tener un margen de error mucho mayor.
- De igual forma, cuando se trabaja en el contexto de la *minería de datos*, la *distribución* de los valores de los atributos es casi siempre desconocida. Para ajustar las observaciones dentro de una *distribución estándar* y seleccionar el *test* adecuado, se requieren esfuerzos computacionales no triviales, especialmente cuando se trabaja con *conjuntos de datos* grandes. Esto constituye también una limitación para la aplicación, en dicho contexto, de métodos de detección basados en *distribuciones*.

Universitat d'Alacant
Universidad de Alicante

Aplicación de la Detección de *Outliers*

A continuación haremos un análisis de la incidencia del problema de la *detección de outliers* en la actualidad. Puede afirmarse que, en las últimas décadas, ha tomado especial relevancia en múltiples y diversos contextos científicos. En ello ha jugado un papel trascendente el vertiginoso desarrollo que han alcanzado las tecnologías de la información en los últimos años. En la **Introducción** de

este trabajo se mencionaron a grandes rasgos varios ejemplos que ilustran lo que se acaba de expresar. Una lista más exhaustiva, que refleja algunos de los campos donde la detección de *outliers* tiene un significado especial, es la siguiente:

- Detección de fraudes:
  - en el uso de tarjetas de crédito,
  - en la telefonía celular,
  - en el procesamiento de aplicaciones de préstamos bancarios. En este caso, además de *fraudes*, la detección de *outliers* puede permitir identificar usuarios potencialmente conflictivos.
- Detección de *intrusos* en las redes de computadoras (detección de accesos no autorizados a las mismas).
- Monitorización de actividades de diverso tipo (Por ejemplo, actividad de un teléfono móvil, comercio electrónico, etc.). A partir de lo cual se detectan acciones sospechosas e indebidas en las mismas.
- Monitorización del funcionamiento o rendimiento de una red de computadoras (Por ejemplo, para detectar *cuellos de botella* que se produzcan en la misma).
- Diagnósticos de fallos o desperfectos en el funcionamiento de motores, generadores, tuberías, instrumentos de medición, instrumentos espaciales, etc.
- Detección de defectos estructurales (Por ejemplo, en el control automatizado de líneas de fabricación para detectar producciones defectuosas).
- Análisis de imágenes obtenidas vía satélite (Por ejemplo, para ayudar a identificar y/o detectar una mala *clasificación* de elementos de interés).

Universidad de Alicante

- Detección de aspectos nuevos en imágenes (Por ejemplo, en robótica, en sistemas de vigilancia o de supervisión, etc.).
- Monitorización automatizada de parámetros médicos (Por ejemplo, el ritmo cardiaco).
- Investigaciones farmacéuticas (Por ejemplo, para identificar nuevas estructuras moleculares).
- Detección de entradas inesperadas en una Base de Datos (Por ejemplo, en *minería de datos*, para determinar errores, *fraudes*, valores válidos pero inesperados, etc.)

Como complemento a esta información general, se incluye una relación de trabajos de investigación que ilustran lo expresado anteriormente. En especial, nos centraremos en los que están vinculados a los procesos de *data mining/web mining*:

- Detección de *intrusos* en las redes y aplicaciones médicas vinculadas con las investigaciones del cáncer (*Wisconsin Breast Cancer Dataset*) (Hawkins *et al.*, 2002).
- Aplicación de la *minería de datos* al diagnóstico médico (Angiulli *et al.*, 2006).
- Minería de *bases de datos* de manuscritos en el marco de un proyecto de digitalización de la herencia cultural y científica de Bulgaria (Stomoimenova *et al.*, 2005).
- Aplicación de técnicas de *minería de datos* a *bases de datos* de alto grado de dimensionalidad. Se incluyen diversas aplicaciones en las que juega un papel decisivo el diagnóstico (Por ejemplo, aplicaciones referidas al diagnóstico médico) (Aggarwal, 2007).

Universitat d'Alacant
Universidad de Alicante

- Aplicación de la detección de *outliers* a estudios climáticos (Intervalos de temperaturas en diferentes ciudades del mundo) (Li *et al.*, 2006).
- Procesamiento de datos poblacionales del *US Census Bureau's Income* (*CENSUS*, 2009), (Otey *et al.*, 2005a), (Otey *et al.*, 2005b).
- Aplicaciones referidas al tránsito vial (Shekhar *et al.*, 2003).
- Minería de datos en el contexto del *Customer Relationship Management* (*CRM*) (He *et al.*, 2004).
- Aplicación de técnicas de detección de *outliers* en *conjuntos de datos* con información procedente de aplicaciones referidas al uso de tarjetas de crédito (Last & Kandel, 2001).
- Aplicaciones para la *clasificación* de tipos de cristales a partir de *test* químicos y físicos (Last & Kandel, 2001).
- Aplicación en el ámbito deportivo (Análisis de *conjuntos de datos* con información de la *NBA*) (Papadimitriou *et al.*, 2003).
- Aplicación en el proceso del *OLAP* (*Online Analytical Processing*) en el contexto del *Data Warehouse* (Lin & Brown, 2001).
- Aplicación en el contexto de un sistema de *video-vigilancia* (*video surveillance*) para garantizar seguridad en áreas públicas (*video/image data mining*) (Knorr *et al.*, 2000).
- *MINDS*: *Minnesota Intrusion Detection System* (Lazarevic *et al.*, 2003).
- Aplicación al análisis de datos químicos (Cramer *et al.*, 2004).

Universitat d'Alacant
Universidad de Alicante

Todo este conjunto de aplicaciones, de actualidad, en las que el *problema de la detección de outliers* tiene una incidencia directa hace evidente la existencia del mismo así como la necesidad y el interés de la comunidad científica de resolverlo con técnicas y métodos eficientes.

Causas de la Aparición de *Outliers*

Una vez presentado el *problema general de la detección de outliers* y los diferentes campos de aplicación del mismo en la vida real, valdría la pena preguntarse:

¿Cuáles son las diversas causas que pueden provocar la aparición de los *outliers*?

De acuerdo con la naturaleza de los mismos, es evidente que las causas de aparición de *outliers* pueden ser muchas, pero consideramos que sería suficiente, a los efectos del presente trabajo, señalar sólo algunas de las más significativas, entre las que se encuentran las siguientes:

- Como consecuencia de algún error de procedimiento. Por ejemplo, errores humanos en la introducción de los datos o algún error en la codificación de los mismos. Estos se cometen con frecuencia y de forma involuntaria. Este tipo de *casos excepcionales* debe identificarse durante el proceso de filtrado de los datos.
- A partir de una observación que ocurre como consecuencia de un acontecimiento extraordinario. En este caso, existe una explicación para la unicidad de la observación.
- Por errores que se comenten en las *lecturas* que se hacen de determinados instrumentos de medición.

Universidad de Alicante

- Como resultado de la variación natural que sufren las poblaciones o determinados procesos que no pueden ser controlados.
- A consecuencia de intentos de *fraudes* en las redes de computadoras y otros usos indebidos de los datos.
- Producto de cambios o fallos que suceden en el comportamiento de determinados sistemas.

Por supuesto que, ante cada causa diferente, existe una manera diferente de reaccionar. Es decir, los sistemas de detección de *outliers* proceden de manera diferente, dependiendo del área de aplicación concreta donde tienen incidencia.

Cómo Proceder ante la Aparición de *Outliers*

Veamos cómo proceder en algunos casos representativos:

- Si la presencia de un *outlier* obedece a un error humano que se ha cometido durante el proceso de introducción de los datos, entonces al ser detectado, se teclea correctamente el dato erróneo y con esto se soluciona el conflicto. Si se decide pasar por alto el error, entonces se debe eliminar el dato, considerándolo como un *dato ausente*.
- Si el *outlier* se origina por un error en la lectura de algún instrumento de medición, simplemente, se borra el mismo y queda resuelta la situación.
- Si ocurre como consecuencia de un acontecimiento extraordinario y existe una explicación para la unicidad de la observación, el investigador debe decidir si el *outlier* debe quedar representado en la muestra o no. Si es así, el *caso excepcional* debe ser retenido en el análisis. En caso contrario, habrá que suprimirlo. Por ejemplo, en una encuesta sobre las características de

los individuos de una población entre cuyos datos significativos se encuentra la estatura de los mismos, un *outlier* puede ser un dato que represente a una persona (o quizás varias) con una altura muy desproporcionada con respecto al resto, pero que puede ser real y por tanto, debe quedar representada en la muestra.

- Si el *outlier* ha sido detectado en algún contexto de extrema seguridad, la manera de proceder es bien diferente a las descritas anteriormente. Por ejemplo, si se detecta por algún sistema de detección de *fraudes* o por algún sistema de detección de *intrusos*, o en algún contexto cuya aparición puede traer aparejada una situación de peligro, la acción debe ser inmediata. Quizás ello implique activar algún sistema de alarma.
- En muchos casos, estos datos anómalos, una vez tratados, se almacenan separadamente, como muestras representativas, para poder establecer, posteriormente, comparaciones ante nuevos indicios que hagan suponer una reiteratividad en su aparición.

Universitat d'Alacant
Universidad de Alicante

Como conclusión, a partir de los aspectos que se acaban de exponer, podemos plantear que la manera de proceder cuando se detecta un *outlier* dentro de un determinado contexto, requiere de un tratamiento personalizado dependiendo de las particularidades de dicho contexto. De igual manera, las particularidades del propio contexto de aplicación condicionan, en gran medida, el tipo de técnica a aplicar para la detección de *outliers*. Por tal motivo, desde que hubo conciencia de la necesidad de resolver este problema han surgido diversos métodos de detección. En la actualidad, a partir del auge que ha tomado el problema en múltiples ámbitos investigativos, ha aumentado considerablemente el número de técnicas de detección utilizadas a tales fines. Abordemos entonces, a

continuación, un análisis en relación a este importante aspecto.

Técnicas de Detección de *Outliers*

En (Hodge & Austin, 2004) se señala que dos de los aspectos fundamentales que deben ser tenidos en consideración para establecer un adecuado proceso de detección de *outliers* son los siguientes:

- Seleccionar un algoritmo que modele con precisión los datos y que destaque con exactitud la *excepcionalidad* de los objetos a partir de alguna técnica concreta. El algoritmo debe ser eficiente, desde el punto de vista computacional, y debe ser *escalable* para los posibles *conjuntos de datos* sobre los cuales será aplicado.
- Seleccionar un *entorno* o *vecindad* de interés para los *outliers*. Esta selección, en ningún caso resulta trivial. Varios algoritmos establecen *fronteras* en relación a la normalidad durante el procesamiento de los datos y de manera particular, fijan un umbral adecuado para el establecimiento de la *excepcionalidad*. Estas aproximaciones con frecuencia son *paramétricas* e imponen un modelo de *distribución* específico. En otros casos, requieren determinados parámetros que deben ser especificados por el usuario. En algunos enfoques, el usuario debe definir el tamaño o la *densidad* de la *vecindad* que se tendrá en cuenta.

Según (Tang *et al.*, 2002), los métodos de detección de *outliers* provienen, por lo general, de dos campos fundamentales de la Matemática y la Ciencia de la Computación: la Estadística y la Inteligencia Artificial (*IA*).

En cuanto a los métodos basados en *IA*, las técnicas fundamentales en las que se basan los mismos son las referidas a *Machine Learning* (Last & Kandel, 2001): Árboles de Decisión y Redes Neuronales.

Todo lo anteriormente expuesto avala el hecho de que en la actualidad se describa una gran variedad de técnicas para la detección de *outliers*. Varias de ellas, puede decirse que son lo suficientemente *maduras* a partir de los resultados positivos que se han alcanzado como consecuencia de su aplicación en diversos contextos.

En (Hodge & Austin, 2004) y (Tang *et al.*, 2002) se describe un amplio conjunto de métodos de detección de *outliers* que se instrumentan a partir de algoritmos de diverso tipo. En general, los algoritmos o métodos de detección de *casos excepcionales* se han clasificado según la técnica en que se basan. En la inmensa mayoría de las clasificaciones hay un pequeño grupo de técnicas que siempre se mencionan. Entre ellas se encuentran las seis primeras que aparecen en la relación que se mostrará en párrafos sucesivos. Haciendo una revisión más profunda, nos hemos percatado que hay muchas más y existe una tendencia a que las mismas proliferen en correspondencia con el acelerado y constante desarrollo de las nuevas tecnologías de la información.

Cada técnica resalta los aspectos nuevos que la distinguen del resto. Varias de ellas tienen como base alguna de las técnicas estadísticas que ya han sido mencionadas y se complementan con el uso de estructuras de datos específicas y de otras técnicas recientes; como, por ejemplo, diversas técnicas de Inteligencia Artificial.

En la documentación revisada se ha podido observar que existe un gran número de técnicas de detección de *outliers*.

Una lista bastante amplia de estas se detalla en los siguientes apartados.

Técnicas basadas en *Distribuciones*

Se basan en algún modelo de *distribución estándar* (Por ejemplo, la *distribución Normal*, la *regresión lineal*, etc.) y consideran como *outliers* a aquellos puntos que se desvían o se alejan considerablemente de dicho modelo. Otros métodos más recientes, usan un modelo *Gaussian* mixto para representar el comportamiento de la *Normal* y a cada elemento del *conjunto de datos* se le asigna un valor basado en los cambios que ocurren en el modelo. Mientras más alto es este valor en un objeto, más posibilidades hay que el mismo sea *outlier*. La mayoría de los métodos basados en *distribuciones* son aplicables solamente a *conjuntos de datos* de una sola *dimensión* (Hawkins, 1980), (Rousseeuw & Leroy, 1987), (Barnett & Lewis, 1994), (Yamanishi *et al.*, 2000), (Yamanishi & Takeuchi, 2001).

Técnicas basadas en *Profundidades*

Los enfoques basados en *profundidades* han sido propuestos con el objetivo de eliminar las limitaciones de las técnicas basadas en la *distribución* de los datos. Cada objeto o dato se representa como un *punto* en un espacio *k*-dimensional y se le asigna una *profundidad*. Los objetos se organizan, por ejemplo, en *capas* (*layers*) de *envolturas convexas* (*convex hulls*) en el espacio de los datos, suponiendo que los *outliers* se encuentran en las capas menos profundas. Estos métodos pueden superar el problema de la adecuación de la *distribución* y conceptualmente pueden procesar datos en un espacio multidimensional. Sin embargo, en la práctica, hay un problema computacional en el enfoque: para calcular las

capas *k*-dimensionales, la técnica depende del cómputo de la *envoltura convexa* (*convex hull*) *k*-dimensional, que tiene una *complejidad temporal* $O(n^{\lfloor k/2 \rfloor})$, *n*: *cardinalidad del conjunto de datos*. Por lo tanto, los métodos basados en la *profundidad* de los datos no son prácticos para *conjuntos de datos* de más de 4 *dimensiones*. De hecho, los algoritmos existentes solo son aceptables cuando la dimensionalidad de los datos es menor o igual que 2 (Ruts & Rousseeuw, 1996), (Jhonson *et al.*, 1998), (Knorr & Ng, 1998).

Técnicas basadas en *Distancias*

Existen varios enfoques del problema de la detección de *outliers* a partir de criterios de *distancia*. Quizás sean los criterios más usados a tales fines. Algunos de ellos ya han sido mencionados al explicar los *métodos estadísticos*. No obstante, podemos señalar otros enfoques además de los ya expuestos (Knorr & Ng, 1998), (Knorr & Ng, 1999), (Angiulli & Pizzuti, 2002), (Bay & Schwabacher, 2003), (Angiulli & Pizzuti, 2005).

Formalmente, algunos autores (Knorr y Ng, 1997), (Knorr y Ng, 1998) consideran que un objeto *x* en un *conjunto de datos* es un *outlier* con respecto a los parámetros *k* y *d,* si la cantidad de objetos que se encuentran a una distancia menor o igual que *d* de *x*, no supera el valor de *k*.

La **Figura 2-10** ilustra un ejemplo de la aplicación de esta definición para *k*=3 y una distancia dada *d*. Claramente, x_i y x_j son *outliers* pues ninguno de los dos tiene más de 3 objetos a una distancia menor o igual que *d*. En cambio, x' no lo es pues la cantidad de objetos que se encuentran a una distancia menor o igual que *d* excede el valor fijado para *k*.

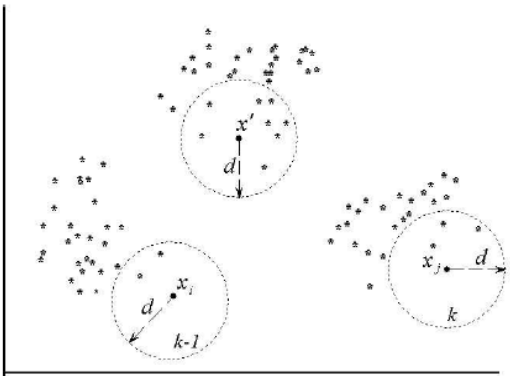


Figura 2-10 Aplicación de la técnica basada en *distancias* para una distancia *d* y *k*=3

Esta técnica es capaz de manipular datos *p*-dimensionales para cualquier valor de *p* y la *complejidad temporal* del método es $O(pn^2)$, *n: cardinalidad del conjunto de datos*. En la actualidad, se han presentado variantes optimizadas de este método que tienen *complejidad temporal lineal* con respecto a *n*, aunque es exponencial con respecto a *p* (Prayote, 2007).

Algunos enfoques basados en distancias aplican la técnica de los *k vecinos más cercanos (K-NN)* desde la siguiente perspectiva: Dados los números enteros *k* y *n* (*k*<*n*). Se consideran *outliers* los *n* objetos con la mayor distancia a sus *k-ésimos vecinos más cercanos*, o sea, los de mayor distancia a los que más cerca tiene (Ramaswamy *et al.*, 2000)

Técnicas basadas en *Densidades*

Este tipo de técnica, por lo general, se centra en la *densidad* de los datos en la *vecindad local* de un objeto y resuelve el problema de la detección de *outliers* a partir del establecimiento de un conjunto arbitrario de *clusters*. En cada *cluster* se agrupa una determinada cantidad de datos

y se establece la *densidad* del mismo. Los métodos que siguen este enfoque, en vez de decidir de forma binaria (*verdadero* o *falso)* si un objeto es un *outlier* o no; asignan un *factor de ruido* a cada objeto. Mientras más *ruidoso* sea un objeto, mayor es su probabilidad de ser *outlier*. Los métodos de detección basados en *densidades* fueron desarrollados especialmente para la detección de *outliers* en *conjuntos de datos* multidimensionales, en los que es más significativo el establecimiento de este tipo de *factores* en vez de aplicar el enfoque binario en la detección (Jin *et al.*, 2001), (Chiu & Fu, 2003), (Hu & Sung, 2003), (Ren *et al.*, 2004).

A modo de ejemplo, podemos hacer referencia a los trabajos de (Markus *et al.*, 2000). En ellos se define para cada objeto en el *conjunto de datos* un *LOF* (*local outlier factor*), que juega un papel decisivo en la detección de *outliers*. El valor del *LOF* se determina en función de la *densidad* de la *vecindad local* de cada objeto. El *LOF* es local en el sentido de que solo una *vecindad* restringida de cada objeto es tenida en consideración. Ésta se define en función de los *k-ésimos vecinos más cercanos* a dicho objeto (*k* es un valor mínimo convenientemente establecido). Los objetos con un valor alto de su *LOF* son considerados *outliers*. El *LOF* es una medida para establecer la diferencia, en cuanto a *densidad*, entre un objeto y los de su *vecindad*.

En otros casos, este tipo de métodos están basados en la Teoría de Grafos, donde juega un papel importante la conectividad entre los *nodos* (datos) del grafo. En tales casos, se tienen en cuenta tanto los valores de los atributos de los datos como las relaciones espaciales entre los mismos.

Métodos basados en Técnicas de *Clusters*

Se basan en diversos algoritmos de *clustering*. Dichos algoritmos se distinguen unos de otros, a partir de las técnicas usadas para el establecimiento de los *clusters*: particionales (*partition clustering),* jerárquicos (*hierarchical clustering)*, basados en *densidades* (*density-based clustering),* basados en *distancias* (*distance-based clustering*) y basados en *cuadrículas (grid-based clustering),* entre otros (Kaufman & Rousseeuw, 1990), (Sudipto *et al.*, 2000), (Jiang *et al.*, 2001), (Yu *et al.*, 2002), (He *et al.*, 2002), (He *et al.*, 2003).

La esencia de la técnica es agrupar objetos tratando de minimizar la distancia de los mismos a un supuesto *centro del cluster* al cual pertenecen.

Un objetivo que se trata de alcanzar en los métodos de detección de *outliers* basados en *clustering* es intentar determinar los *outliers* a la vez que se van conformando los *clusters,* para posteriormente eliminarlos del *conjunto de datos* original y hacer así más fiable dicho proceso.

Los objetos que se encuentran aislados y que forman un *cluster* de tamaño 1 son considerados *outliers*. De igual modo, los objetos que están en *clusters* considerados muy pequeños, pero con cardinalidad >1, son también considerados como tal.

En una de las técnicas más referenciadas, el *clustering* se aplica estableciendo *particiones* del *conjunto de datos*. En estos casos, el usuario proporciona el número de *clusters* que desea crear y el número de variables que se usan para crear la *partición*. Algunos algoritmos muy conocidos que se basan en este enfoque son *k-means* (Hautamäki *et al.*, 2005), PAM y CLARA (Kaufman & Rousseeuw, 1990), (Han & Kamber, 2000). *K-means clustering* es uno de los más

conocidos por su efectividad al ser aplicado a datos en *espacios euclidianos*. *Clustering Outliers Removal* es un método basado en *k-means clustering* (Hautamäki *et al.*, 2005).

Técnicas basadas en el Uso de *Redes Neuronales*

Los métodos de detección basados en *redes neuronales* son generalmente *no paramétricos* y permiten hacer generalizaciones de *patrones* desconocidos (Hodge & Austin, 2004). Las *redes neuronales* se han usado, en diversos contextos, como *clasificadores*. Antes de que estén en condiciones de clasificar nuevos *conjuntos de datos*, algunas requieren de *entrenamiento* y *validación* para alcanzar el refinamiento necesario (Simon *et al.*, 2002), (Hawkins *et al.*, 2002), (He, *et al.*, 2002), (He *et al.*, 2004), (Toth & Gosztolya, 2004).

Los métodos de detección basados en redes neuronales se clasifican en *supervisados* y *no supervisados*, según la *clasificación* que en este sentido tenga la red que se utilice.

Universitat d'Alacant
Universidad de Alicante

Métodos basados en *Redes Supervisadas*

Las *redes supervisadas* usan la *clasificación* de los datos para guiar el proceso de *aprendizaje* de la misma. En ellas se establecen mecanismos a partir de los cuales se ajustan los *pesos* y los *umbrales* de forma tal que se asegure una correcta *clasificación* de los datos de entrada. Requieren de un proceso de *entrenamiento* previo de la red para garantizar el *aprendizaje*. El entrenamiento previo comienza a partir de un *conjunto de datos* clasificado (cada dato de entrada se ha clasificado en función a su *respuesta objetivo*) y, por lo general, este proceso conlleva a que dicho conjunto tenga que recorrerse varias veces para que la red

logre clasificar y modelar los datos correctamente. Como es obvio, esto representa un gasto de tiempo adicional.

Entre las *redes supervisadas* más usadas en la detección de *outliers* puede mencionarse la *Multi-Layer Perceptron (MLP).*

En (Hodge & Austin, 2004) se citan los trabajos de Bishop y Nairac como ejemplos del uso de *MLP* en la detección de *outliers*. En (Nairac *et al.*, 1999) se describe una aplicación en este sentido para el diagnóstico de fallos de motores en naves aéreas. Por su parte, en (Bishop, 1994) se describe una aplicación asociada a la monitorización de procesos, como por ejemplo; el flujo de petróleo en los oleoductos.

La red neuronal *Replicator Neural Network (RNN)* es una red inspirada en el *MLP.* La forma flexible y *no-paramétrica* de representar *clusters* que presenta hace que, en sí misma, constituya un potente método de detección de *outliers*. Esto justifica el hecho de que en muchos de los trabajos publicados se cita su uso. (Graham *et al.*, 2002) y (Hawkins *et al.*, 2002) constituyen dos ejemplos de ello.

Universitat d'Alacant
Universidad de Alicante

Métodos basados en *Redes No Supervisadas*

Como acabamos de comentar, los métodos basados en *redes supervisadas* necesitan que un conjunto de datos de entrada sea clasificado previamente para posibilitar el *aprendizaje* de la red. Cuando no se dispone de tal *clasificación*, entonces se recomienda el uso de *redes no supervisadas*.

Entre las *redes neuronales no supervisadas* más usadas en los métodos de detección de *outliers*, puede citarse la *Self Organising Maps (SOM)* (Kohonen, 1996).

En general, la efectividad de varios enfoques basados en *redes neuronales* se ve afectada por la dimensionalidad del

conjunto de datos, aunque tal afectación es en menor medida que la que sufren las *técnicas estadísticas de detección*. Por tal motivo, la mayoría de estos enfoques reducen, automáticamente, los atributos de entrada para centrarse en los atributos esenciales para el análisis. Con ello logran proyecciones de los datos a partir de las cuales se reduce la dimensionalidad de los mismos.

En la funcionalidad de algunos métodos de detección basados en *redes neuronales,* la *densidad* de los datos juega un papel importante y por tanto, incide en la eficiencia del método.

Técnicas basadas en *Subespacios*

Como ya se ha señalado, el tamaño y la dimensión del *conjunto de datos* puede tener una incidencia negativa en la efectividad de los algoritmos de detección de *outliers*. Teniendo esto en cuenta, se han diseñado técnicas de detección que se basan en el establecimiento de determinadas *proyecciones* de los datos y en la *distribución de densidad* de los mismos dentro de ellas. Es decir, *proyectar* los datos que son más susceptibles al análisis en *subespacios* de menor dimensión de manera tal que esto facilite el proceso de detección (Aggarwal & Yu, 2001), (Wei *et al.*, 2003), (Aggarwal & Yu, 2005).

Técnicas basadas en *Soporte Vectorial* (*Support Vector*)

En (Tax & Duin, 1999) se desarrolla por primera vez el *support vector novelty detector* (*SVND*). Algunas aproximaciones al *SVND* se basan en la estimación de una esfera en la cual se encuentran inmersos los datos. Aquellos con patrones normales tienen un pequeño radio

con respecto al centro de la esfera. Los *outliers*, sin embargo, serán los datos con radios considerablemente grandes con respecto al mismo (Scholkopf *et al.*, 2001), (Cao *et al.*, 2003), (Petrovsky, 2003).

En otros casos, lo que se estima es una función que separa la región donde se encuentran los datos con patrones normales, de la región donde se encuentran los que tienen una máxima marginalidad.

Otras Técnicas de Detección de *Outliers*

Otras técnicas usadas en la detección de *outliers* son las siguientes: Técnicas basadas en *ejemplos* (Zhu *et al.*, 2005); Técnicas basadas en *geometría computacional* (Jhonson *et al.*, 1998); Técnicas basadas en *enlaces* (*link-based*) (Ghoting *et al.*, 2004); Técnicas basadas en la *lógica matemática* (Angiulli *et al.*, 2006) (Angiulli *et al.*, 2007); Técnicas basadas en *patrones frecuentes* (Zengyou *et al.*, 2005); Métodos *espaciales* (*Spatial Methods*) (Shekhar *et al.*, 2003), (Lu *et al.*, 2003); Técnicas basadas en *conjuntos difusos* (*Fuzzy Sets*) (Last & Kandel, 2001); Técnicas basadas en *modelos de optimización* (Xu *et al.*, 2005); Técnicas basadas en *métodos de Kernel* (Shawne-Taylor & Cristianini, 2004); Técnicas basadas en el *análisis de fractales* (Cramer *et al.*, 2004); Técnicas basadas en *grafos* (Zhong *et al.*, 2009); Técnicas basadas en *árboles de decisión* (Takeshi & Einoshin, 2002), (Reif *et al.*, 2008); y Métodos basados en *técnicas de paralelismo* (Nguyen *et al.*, 2006).

Consideraciones para la Concepción de Métodos de Detección de *Outliers*

Para lograr una buena efectividad en la detección de *outliers,* se requiere la construcción de un modelo de detección que se ajuste y represente con precisión los datos sobre los cuales será aplicado. Sin embargo, los complejos y disímiles *conjuntos de datos* de hoy en día, donde estas técnicas se aplican, hacen que surjan diversas dificultades que limitan la efectividad de ciertos métodos. A modo de resumen, puede señalarse que entre los problemas más sobresalientes en este sentido se encuentran los que a continuación se relacionan (Ben-Gal, 2005), (Otey *et al.*, 2005c):

- El *dinamismo* de los *conjuntos de datos*.
- La dimensión y el tamaño de los *conjunto de datos.*
- La posibilidad real que existe hoy en día, a partir de los avances en las nuevas tecnologías de la información, de que un mismo *conjunto de datos* esté *distribuido* en varios sitios.
- La mezcla de diferentes tipos de atributos en los datos.
- La naturaleza o el tipo de *outlier* que se espera que se genere.
- La proporción de *outliers* que exista en el *conjunto de datos*.

Por la importancia que reviste, especialmente en las aplicaciones vinculadas al *DM,* las consideraciones en cuanto a la alta dimensionalidad de los *conjuntos de datos*, es importante tener en cuenta los aspectos señalados en (Aggarwal & Yu, 2001) en relación a las características que

deben tener los algoritmos de detección de *outliers* para que sean eficientes en tales contextos:

- Deben ser diseñados en base a técnicas que resuelvan de forma eficiente el problema que entraña el hecho de que en los *conjuntos de datos* de alta dimensionalidad los datos, por lo general, se encuentran muy *esparcidos*.
- Deben facilitar el razonamiento a partir del cual se explique de dónde procede la *excepcionalidad*.
- Los algoritmos de detección deben ser diseñados de manera tal que su aplicación no se vuelva impracticable a medida que crezca el tamaño y la dimensión del *conjunto de datos* (*escalabilidad*).
- Los algoritmos deben conceder importancia al comportamiento local de los datos a la hora de su *clasificación* como *outliers*.

A modo de conclusión podemos plantear que en el contexto del *DM:*

- el *dinamismo*
- el tamaño
- la dimensión
- la posibilidad de estar distribuidos

son aspectos, en relación al *conjunto de datos*, a los que se debe prestar especial atención a la hora de diseñar o aplicar los métodos de detección de *outliers*. Otro aspecto a tener en cuenta, es la diversidad en los tipos de los valores de los atributos asociados a los datos. Cada vez con mayor frecuencia, aparecen datos *no ordinales* expresados de forma *categórica*. De igual forma, deben ser tenidas en cuenta las particularidades que distinguen a los *outliers* que pueden ser generados en uno u otro contexto.

Universitat d'Alacant
Universidad de Alicante

Consideraciones sobre Comparaciones de Métodos

En (Ben-Gal, 2005) se argumenta que debido a que una buena parte de los algoritmos de detección de *outliers* se basan, asumen, abordan y tienen en consideración un conjunto disjunto de aspectos, así como que se aplican a diversos tipos de datos dentro de entornos diferentes, no es posible establecer una comparación directa entre todos ellos. En varios casos, las estructuras de datos usadas y los mecanismos de generación de *outliers* en los que se basa la investigación, determinan el método de detección que debe aplicarse.

En la revisión bibliográfica realizada se han encontrado algunos trabajos donde se hacen comparaciones parciales entre ciertos métodos de detección de *outliers*. Sólo por citar algunos ejemplos en este sentido, se mencionan los siguientes:

- En (Penny & Jolliffe, 2001) se comparan seis métodos de detección de *outliers multivariados*. Uno aspecto esencial que se resalta en ese trabajo es que no se puede considerar que una técnica de detección es superior al resto. Hay que tener en cuenta las particularidades del contexto donde será aplicada y tratar que sea eficiente en dicho contexto.
- En (Graham *et al.*, 2002) se hace una comparación entre un método de detección de *outliers* basado en el *replicator neural network* (*RNN*) (Simon *et al.*, 2002), dos métodos *paramétricos* estadísticos y un método *no paramétrico* de detección usado en *minería de datos*. En este trabajo se resume el estudio comparativo indicando que en los problemas de detección de *outliers* no pueden aplicarse, fácilmente, criterios sencillos de ejecución.

- En (Shekhar *et al.*, 2003) se caracterizan los algoritmos *espaciales* de detección de *outliers* y se hacen comparaciones entre varios de ellos. Un aspecto importante que concluyen es que la aplicación de este tipo de algoritmos reduce considerablemente el riesgo de determinar *outliers falso – positivos*.

- En (Hodge & Austin, 2004) se hace un recuento de algunas de las metodologías de detección de *outliers* más recientes. En dicho trabajo se comparan cinco metodologías diferentes: modelos estadísticos (enfoques basados en *proximidades*, métodos *paramétricos*, no *paramétricos* y *semiparamétricos*), *redes neuronales* (*supervisadas* y *no supervisadas*), otras técnicas de *Machine Learning* y los *modelos híbridos*, usados especialmente para resolver algunas deficiencias de algoritmos específicos de *clasificación*, explotando las ventajas de otras técnicas.

- *En* (Otey *et al.*, 2005c) se comparan tres métodos en cuanto a calidad de la detección, *complejidad espacial* y *temporal*. A saber:

  ORCA (Bay & Schwabacher, 2003) pretende resolver los problemas asociados al orden cuadrático de la mayoría de los métodos basados en *distancias*. La idea central del método es establecer un mecanismo de control a partir del cual se descartan datos a tener en consideración en el análisis. Para el *caso peor*, cuando no existen *outliers* en el *conjunto de datos*, el método se comporta de forma *cuadrática*, pero cuando esto no es así, lo cual es el caso promedio, se comporta *casi lineal*. Trabaja sobre *conjuntos de datos* donde hay mezcla de atributos, utilizando diferentes criterios de *distancia* para cada caso. Hace un uso racional de la memoria.

Universitat d'Alacant
Universidad de Alicante

LOADED (Ghoting *et al.*, 2004) se diseñó explícitamente para ser usado tanto en *conjuntos de datos estáticos* como *dinámicos* con atributos de diverso tipo. Versiones recientes del mismo, extienden su uso a *conjuntos de datos distribuidos*. Usa como estructura de datos fundamental para modelar los datos el *lattice* (*malla* o *entramado*) *ampliado*. En él, se representan los conjuntos de valores asociados a cada uno de los atributos *categóricos* de los datos. Este método es el que mejor funciona, de los tres que se comparan, con respecto a los índices de detección. Su principal limitación es que no hace un uso eficiente de la memoria.

RELOADED (Otey *et al.*, 2005a), (Otey *et al.*, 2005b) fue diseñado para dar solución a los problemas de memoria que presenta *LOADED*. Descarta el uso del *lattice* y usa varios conjuntos de clasificadores para modelar las dependencias entre los distintos atributos categóricos.

Financiación da las Investigaciones

Un aspecto que puede señalarse como argumento para destacar el interés que se presta a nivel internacional a las investigaciones referidas al problema de la detección de *outliers*, es el hecho de que varias de las mismas son financiadas por programas e instituciones, gubernamentales y no gubernamentales, prestigiosas. A modo de ejemplo se citan los siguientes casos:

- Los importantes resultados de la investigación expuestos en (Otey *et al.*, 2005a), (Otey *et al.*, 2005b), (Otey *et al.*, 2005c) por un grupo de investigadores del *Department of Computer Science and Engineering* de la Universidad de Ohio están soportados por un *grant*

otorgado por la *US NSF National Science Foundation* de los Estados Unidos. (CAREER-IIS-0347662) y (NGS-CNS-0406386).

- La investigación reportada en (Ziarko, 2001) por el autor del Modelo de *Rough Sets* de Precisión Variable (*VPRSM*) está financiada por un *research grant awarded* otorgado por el *Natural Sciencies and Engineering Research Council* de Canadá.
- El resultado publicado en (Bing-Zhen *et al.*, 2004) se enmarca en un proyecto de investigación financiado por el *National Natural Science* de China (40235053), el *Natural Science Foundation* de la provincia de Gansu (NWNU-KJCXGC-212) y el *Key Project Fund* del NWNU.
- El proyecto de investigación en el cual se inserta la investigación expuesta en este documento tiene sus orígenes en el contexto de un proyecto de investigación financiado por la Universidad Politécnica de Madrid, España. (AL-1005-2002/2005).
- La investigación cuyos resultados son expuestos en (Stomoimenova *et al.*, 2005) es financiada por una beca *Marie Curie* del programa *Knowledge Transfer for Digitalization of Cultural and Scientific Heritage in Bulgaria* de la Comunidad Económica Europea, con número de contrato: MTKD-CT-2004-509754.
- (Zhu *et al.*, 2005) es un resultado que esta soportado, desde el punto de vista económico, por *Japan-U.S. Cooperative Science Program of JSPS* y *The Grant-in-Aid for Scientific Research from JSPS and MEXT* (#15300227).
- El trabajo que se expone en (Last & Kandel, 2001) está financiado por *The National Institute for Systems Test and Productivity* en la *University of South Florida (USF)*

bajo los auspicios del *USA Space and Naval Warfare Systems Command. Grant No.:* N00039-01-1-2248, además por el *Center for Software Testing* de la USF. *Grant* No.: 2108-004-00.

- (Xu *et al.*, 2005) es un proyecto financiado por *The High Technology Research and Development Program of China* (No. 2003AA4Z2170, No. 2003AA413021), *The National Nature Science Foundation of China* (No. 40301038) e *IBM SUR Research Fund*.

Teoría de *Rough Sets* aplicada a la Detección de *Outliers*

Como ya hemos afirmado, la *minería de datos* emerge, cada vez con más fuerza, como un área de la Inteligencia Artificial que brinda nuevas técnicas para el análisis de datos en los complejos *conjuntos de datos* de hoy en día. La Teoría de *RS,* en este ámbito, incide también decisivamente en el empeño por alcanzar tales metas.

Desde finales de la década de los 80 ya se refieren resultados importantes de la aplicación de la misma en el contexto del *DM*. En el trabajo titulado *The State of Rough Sets for Database Mining Applications* publicado en 1995 (Raghavan & Sever, 1995) por dos prestigiosos investigadores de la *University of Southwestern Louisiana* de los Estados Unidos, puede apreciarse una revisión del *estado del arte* en relación a este aspecto.

La aplicación en los últimos años de la Teoría de *RS* en múltiples contextos investigativos pone de manifiesto su efectividad en la solución de disímiles problemas. Esto se evidencia a partir de un gran número de trabajos que han sido presentados en congresos celebrados recientemente o

reportados en publicaciones prestigiosas a nivel internacional. Todo ello demuestra la versatilidad de dicha teoría y sus disímiles entornos de aplicación.

Entre los eventos científicos más importantes, celebrados en años recientes, donde en sus *proceedings* se han publicado varios trabajos que constituyen aplicaciones de la Teoría de *RS* en diversos ámbitos investigativos de gran impacto, pueden señalarse los siguientes:

(En cada caso se cita una referencia representativa del total de aplicaciones de la Teoría de *RS* presentadas en dichos congresos)

- 2009-*International Conference on Management of e-Commerce and e-Government*.

  (Gang, 2009) Aplicación de la Teoría de *RS* al proceso de evaluación de Sistemas de Información complejos.

- 2009-*International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*.

  (Liu *et al.*, 2009) Modelo de red neuronal basado en *RS*.

- 2009-*2nd IEEE International Conference on Computer Science and Information Technology*.

  (Xiaowen *et al.*, 2009) Análisis de Sistemas de Fabricación Reconfigurados (*RMS*) basados en *RS.*

- 2009-*Ninth International Conference on Hybrid Intelligent Systems*.

  (Zhao *et al.*, 2009) Aplicación de la Teoría de *RS* a la solución de problemas en el ámbito de las inversiones.

- 2009-*International Conference on Computational Science and Engineering*.

  (Bouyer *et al.*, 2009) Aplicación de la Teoría de *RS* al *Grid scheduling process.*

Universitat d'Alacant
Universidad de Alicante

- 2009-*International Conference on Business Intelligence and Financial Engineering*.

  (Yao, 2009) Aplicación de la Teoría de *RS* al proceso de adjudicación de créditos bancarios.

- 2009-*International Conference on Environmental Science and Information Application Technology*.

  (Junding & Suxia, 2009) Aplicación de la Teoría de *RS* al procesamiento de imágenes.

- 2009-*Asia-Pacific Conference on Information Processing*.

  (Yin *et al.*, 2009) Aplicación de la Teoría de *RS* a métodos de evaluación de la capacidad de innovación empresarial.

- 2009-*International Conference on Computational Intelligence and Natural Computing*.

  (Wu *et al.*, 2009) Aplicación de la Teoría de *RS* a problemas de seguridad vial.

- 2009-*International Conference on Electronic Commerce and Business Intelligence*.

  (Qingkui & Junhu, 2009) Aplicación de la Teoría de *RS* a sistemas para minimizar el impacto medioambiental negativo de las nuevas tecnologías en el contexto empresarial.

- 2009-*Eigth IEEE/ACIS International Conference on Computer and Information Science (icis 2009)*.

  (Zeng *et al.*, 2009) Aplicación de la Teoría de *RS* a la detección de anomalías en las redes de computadoras.

En algunos de estos congresos, entre los temas centrales de los mismos están las aplicaciones de la Teoría de *RS*. En tal caso pueden citarse las Conferencias Internacionales *RSCTC* (*Rough Sets and Current Trends in Computing)* y

RSFDGrC (*Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing)* que se celebran de manera bianual en diferentes países.

A partir del estudio bibliográfico realizado, el primer reporte de la aplicación de la Teoría de *Rough Sets* en el campo de la detección de *outliers* es una comunicación presentada en *RSFDGrC'05* bajo el título *Outlier detection using Rough Sets Theory* (Jiang *et al.*, 2005). Los estudios previos y los resultados de F.Jiang, Y.Sui y C.Cao (Jiang *et al.*, 2006), en gran medida, sentaron las bases y constituyeron el punto de partida del presente trabajo de investigación. En el siguiente capítulo se hace un análisis crítico de dicha propuesta.

Conclusiones

Finalmente, a partir de la revisión del *estado del arte* realizada podemos arribar a algunas conclusiones en relación al *problema de la detección de outliers* y en tal sentido puede afirmarse lo siguiente:

- El impacto del problema alcanza a diferentes contextos de la vida social, económica y cultural del mundo actual.
- La comunidad científica internacional reconoce que las investigaciones en relación a este tema están entre las que se encuentran a la *cabeza* del desarrollo de la ciencia a nivel internacional. Se reconoce también, el papel que juega en ello el vertiginoso desarrollo alcanzado en los últimos años por las nuevas tecnologías de la información. Por tanto, se reconoce además, la existencia del problema y se destinan esfuerzos y recursos a su solución. El hecho de que diferentes instituciones gubernamentales, sociales,

Universitat d'Alacant
Universidad de Alicante

académicas y científicas destinen parte de los fondos dedicados al desarrollo científico a financiar investigaciones relacionadas con este tema es un ejemplo elocuente de ello.

- Hay conciencia, entre los investigadores, del interés y la necesidad de resolver el *problema de la detección de outliers* con la efectividad requerida, así como de la importancia de su solución para el desarrollo de investigaciones científicas en diversos ámbitos.
- La existencia de un amplio conjunto de técnicas, todas las cuales han demostrado ser lo suficientemente *maduras* para abordar el problema de forma efectiva en dominios específicos de aplicación, han demostrado que la resolución del problema es viable y los resultados previos alcanzados sirven de base teórico-práctica para el desarrollo de nuevas técnicas. Estas intentan dar solución a los problemas abiertos que aún están por resolver tratando de obtener resultados más eficientes e integradores. Esto garantiza la continuidad de las investigaciones en esta dirección.
- Desde nuestro punto de vista, el aspecto principal que hace que el *problema de la detección de outliers,* en general, sea aún un *problema abierto*, es el hecho de que aún no se cuenta con una aproximación universalmente aplicable al mismo y quedan aún un conjunto de subproblemas particulares sin resolver.
- La Teoría de Conjuntos Aproximados (*RS*) ha resultado especialmente útil para realizar intuitivas y novedosas propuestas a partir de las cuales se puede construir algoritmos eficientes para la detección de *outliers*.

Teniendo en cuenta que la principal hipótesis de la investigación radica en la aplicación de la Teoría de *RS* al

problema de la detección de *outliers*, así como de la propuesta realizada en (Jiang *et al.*, 2005), se ha considerado especialmente relevante realizar un análisis de esta propuesta por constituir los antecedentes fundamentales sobre los que se sustenta este trabajo.

El objetivo del análisis es justificar la hipótesis de partida demostrando la idoneidad de las propuestas existentes, al tiempo que queden identificados los problemas abiertos en la actualidad y las posibles vías de solución.

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Capítulo 3

Antecedentes de la Investigación

Teniendo en cuenta la hipótesis global y el objetivo general planteados en la investigación, en el estudio del *estado del arte* realizado llamó nuestra atención la propuesta de un método de detección de *outliers* basado en la Teoría de Conjuntos Aproximados (*Rough Sets*) (Jiang *et al.*, 2005). Dicha propuesta resultó especialmente actractiva gracias, fundamentalmente, a tres motivos:

- Este resultado constituye el primer antecedente de la utilización de dicho modelo en el campo de la detección de *outliers.*
- El marco teórico planteado resultaba muy cercano a nuestro interés.
- La simplicidad del planteamiento formal del método de detección propuesto y lo original de su enfoque, basado en una teoría de bases matemáticas simples y solidas: la Teoría de *Rough Sets*.

Universidad de Alicante

Teniendo en cuenta los elementos anteriormente expresados, se valora como positivo realizar un estudio de esta propuesta, así como, un análisis crítico de la misma.

Para familiarizar al lector con los elementos fundamentales de la Teoría de *RS*, presentamos a continuación los elementos esenciales de la misma. En (Pawlak, 1982), (Pawlak, 1991), se puede encontrar una explicación más completa y detallada de sus fundamentos matemáticos.

La Teoría de *Rough Sets* es una extensión de la Teoría de Conjuntos para su aplicación al caso de información incompleta y/o insuficiente. Esta teoría surge a partir de la necesidad práctica de resolver problemas de *clasificación* y en ella se supone que junto a cualquier *objeto* del *universo* hay asociada una cierta información: el conocimiento que se tiene acerca de dicho *objeto,* que se expresa mediante valores asociados a un conjunto de atributos (propiedades) que describen a dicho *objeto*.

La Teoría de *Rough Sets* tiene el atractivo de contar con una base matemática simple y sólida: La *teoría de las relaciones de equivalencia,* que permite describir particiones constituidas por clases indiscernibles que agrupan a *objetos* con atributos idénticos. Es una *metodología de clasificación de datos* cuya aplicación conduce a la extracción de reglas muy fiables, pero, en algunos casos, poco representativas (escaso soporte). Esto supone, para dicho modelo, un carácter determinista de la *clasificación* obtenida. Una generalización del *Modelo Básico de RS* es el *Modelo de Conjuntos Aproximados de Precisión Variable,* propuesto por W. Ziarko (*Variable Precision Rough Sets Model - VPRSM*), (Ziarko, 1993) que subsana el inconveniente descrito partiendo de una idea muy simple: La *relajación* del concepto de *inclusión de conjuntos*, manejando unos umbrales definidos por el usuario.

Universitat d'Alacant
Universidad de Alicante

Conceptos Fundamentales de la Teoría de *RS*

A continuación se exponen los conceptos fundamentales de la Teoría de *RS* que son usados en el método de detección propuesto.

Sean $U \neq \phi$ el *universo* (finito) y $r \subseteq UXU$, una *relación de equivalencia* definida sobre *U*. Sea $X \subseteq U$, un *concepto*. *X* se describe mediante dos aproximaciones:

Aproximación superior: $\bar{r}(X) = \cup\{Y \in U / r : Y \cap X \neq \phi\}$. La unión de todas las *clases de equivalencia* inducidas por *r* en *U* cuya intersección con *X* es no vacía.

Aproximación inferior: $\underline{r}(X) = \cup\{Y \in U / r : Y \subseteq X\}$. La unión de todas las *clases de equivalencia* inducidas por *r* en *U* que están contenidas en *X*.

La Teoría de *RS*, define el concepto de *frontera* de este modo:

Frontera: $BN(X) = \bar{r}(X) - \underline{r}(X)$

En la **Figura 3-1** se puede observar una representación gráfica de estos tres conceptos.

Análisis Crítico de un Método de Detección de *Outliers* basado en *Rough Sets*

La caracterización matemática de *outliers* propuesta en (Jiang *et al.*, 2005) se basa en los elementos anteriormente enunciados a partir de los cuales definen un nuevo concepto, el de *frontera interna*, del que a su vez derivan otras definiciones:

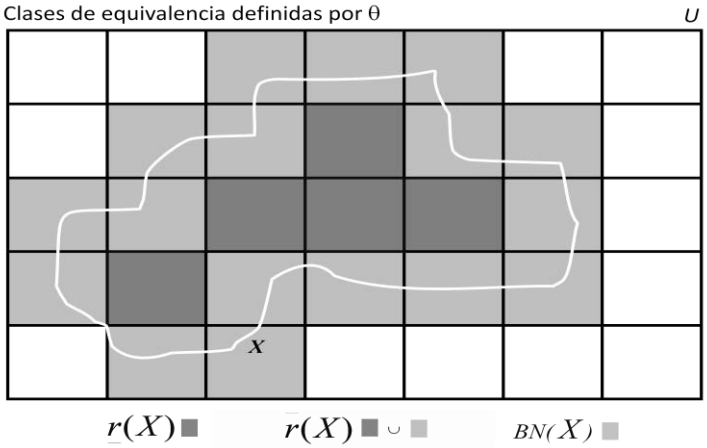


Figura 3-1 Aproximación Inferior, Superior y Frontera

Definición 1 - *Frontera interna:*

Sean $\Re = \{r_1, r_2, ..., r_m\}$, *m relaciones de equivalencia* definidas sobre el *universo U*. La *frontera interna* de *X* con respecto a r_i, se define en la forma siguiente: $B_i(X) = BN_i(X) \cap X = X - \underline{r_i}(X)$

Definición 2 - *Conjunto excepcional:*

Sea $e \subseteq X$ tal que: $\forall r_i \in \Re, \forall B_i(X) \neq \phi$, se cumple $e \cap B_i(X) \neq \phi$. Entonces, al conjunto *e* se le llama *conjunto excepcional* de *X* con respecto a $\Re$.

Definición 3 - ***Elemento dispensable:***

Sea *e* un *conjunto excepcional* de *X* con respecto a $\Re$ y sea $x \in e$ tal que *e-{x}* es también *excepcional* con respecto a $\Re$. Decimos entonces que *x* es *dispensable* en *e* con respecto a $\Re$.

En caso contrario, *x* es ***indispensable*** en *e* con respecto a $\Re$.

Definición 4 - *Conjunto excepcional no redundante:*

Se dice que un *conjunto excepcional* es *no redundante* si todos sus elementos son *indispensables*.

Definición 5 - *Grado de Marginalidad*:

Sea *x* un elemento arbitrario de *X*. El *grado de marginalidad* de *x* con respecto a $\Re$ es la cantidad de *fronteras internas* diferentes de *X* con respecto a $\Re$, que contienen a *x*: $BD(X) = \left|\{B_i(X), i = 1,2,...,m : x \in B_i(X)\}\right|$.

Definición 6 - ***Grado de excepcionalidad:***

El *grado de excepcionalidad* de *x* se define en la forma siguiente: $OD(x)=BD(x) \,/\, |\Re|$

Definición 7 - *Outlier:*

Un *outlier* en *X* con respecto a $\Re$ es un objeto *x* que pertenece a algún *conjunto excepcional no redundante* de *X* con respecto a $\Re$ y que tiene un *grado de excepcionalidad* mayor que un umbral μ dado.

Universitat d'Alacant
Universidad de Alicante

Una aproximación simplificada al método de detección propuesto puede resumirse en dos pasos:

Primer paso:

Determinar todos los *conjuntos excepcionales no redundantes* sobre el conjunto *X* (*un concepto*) dado.

Segundo paso:

Detectar los *outliers* a partir de los elementos de los *conjuntos excepcionales no redundantes*:

«*Todo aquel elemento, cuyo grado de excepcionalidad sea mayor que un umbral μ dado, será un outlier*»

Si se supone que todos los *outliers* en *X* tienen que pertenecer a algún *conjunto excepcional no redundante*, entonces, si algún elemento en *X* no cumple esta condición, podemos afirmar que dicho elemento no será *outlier*.

En el referido trabajo se presenta un algoritmo para determinar el *conjunto excepcional no redundante minimal*, entendiéndose por ello**,** el *conjunto excepcional no redundante,* de *X* con respecto a $\Re$, de menor cardinalidad. En primer lugar, consideramos que este término es irrelevante desde el punto de vista práctico, pues no incide directamente en la esencia del método de detección propuesto, pero además, se encontró un contraejemplo que ilustra que dicho algoritmo no funciona correctamente según el propósito para el cual fue concebido.

El algoritmo que se propone es el siguiente:

Algoritmo para obtener el *conjunto excepcional no redundante minimal* de *X* respecto a $\Re$

Entrada:

$B = \{B_1, B_2, \ldots, B_m\}$ es el conjunto de todas las *fronteras internas* de *X* con respecto a cada *clase de equivalencia* en $\Re$.

Salida:

El *conjunto excepcional no redundante minimal* (*CENRMín*) de *X* respecto a $\Re$.

(1) ***CENRMín*** $= \phi$ // se inicializa y se va formando
// paulatinamente

(2) **while** (B ≠ ϕ) do

{

(3) **for** cada $B_i \in B$

(4) **for** cada $x \in B_i$

(5) Determinar **BD(x);**

(6) Determinar el ***y*** con mayor **BD** en todos los $B_i \in B$;

(si existe más de uno, escoger cualquiera aleatoriamente)

(7) ***CENRMín*** = ***CENRMín*** $\cup \{y\}$;

(8) Eliminar de *B* todos los *Bi* que contengan a ***y;***

}

(9) **return** ***CENRMín;***

Algoritmo 1 Algoritmo de formación del Conjunto *CENRmím*

A partir del análisis que se hace a continuación, del funcionamiento de este algoritmo para un *universo* de datos dado, se determina el contraejemplo.

La **Tabla 3-1** representa, de modo simplificado, un *universo U*. El conjunto de estudio *X* serán los elementos de *U* que son *estudiantes*.

Sean las siguientes *relaciones de equivalencia* definidas sobre *U:*

***r*₁**: edad {1, 2,..., 100}

***r*₂**: país de nacimiento {Brasil, Cuba, Dominicana, ..., Venezuela}

r_3: color del pelo {negro, rubio, rojo, castaño}

r_4: color de los ojos {negros, verdes, azules}

r_5: estado civil {soltero, casado, divorciado, viudo}

r_6: estatura (*pies*) {1, 2,..., 8}

Tabla 3-1 Ejemplo de un *universo* de datos sobre *personas*

ID	Nombre	País	Color cabello	Estado civil	Altura (feets)	Ocup.	Color ojos
1	Carlos	Dom.	negro	soltero	5	estud.	azul
2	Luis	Cuba	rojo	casado	6	estud.	azul
3	Miguel	Cuba	negro	soltero	7	estud.	verde
4	Lorenzo	P. Rico	rubio	divorciado	6	estud.	negro
5	Mario	Cuba	negro	casado	5	desoc..	negro
a	Ernesto	Nicar.	castaño	soltero	7	estud.	negro
b	Pedro	Venez.	castaño	casado	6	estud.	verde
c	Enrique	Brasil	castaño	soltero	5	estud.	verde

A partir de las *relaciones de equivalencia* dadas, se obtienen las siguientes *fronteras* para cada una de ellas:

$BN_1 = \{3, 4, 5\}$ *con frontera interna* $B_1 = \{3, 4\}$

$BN_2 = \{3, 2, 5\}$ *con frontera interna* $B_2 = \{3, 2\}$

$BN_3 = \{3, 1, 5\}$ *con frontera interna* $B_3 = \{3, 1\}$

$BN_4 = \{4, a, 5\}$ *con frontera interna* $B_4 = \{4, a\}$

$BN_5 = \{2, b, 5\}$ *con frontera interna* $B_5 = \{2, b\}$

$BN_6 = \{1, c, 5\}$ *con frontera interna* $B_6 = \{1, c\}$

El conjunto B que sería la entrada del algoritmo es el siguiente:

$B = \{B_1, B_2, B_3, B_4, B_5, B_6\}$

Siendo consecuente con los pasos establecidos en el algoritmo propuesto, se selecciona el elemento con mayor *grado de marginalidad*, o sea, el que aparece en el mayor número de *fronteras internas*. El 3 en este caso y se inserta en el conjunto *CENRMín* que se comienza a formar y que según el algoritmo, será *minimal* entre todos los *conjuntos excepcionales no redundantes* existentes. Tras este primer análisis, dicho conjunto tendrá al 3 como único elemento. Posteriormente y según el paso (8) del algoritmo, se eliminan de *B*, todos los conjuntos donde aparezca este elemento. Tras ello, el conjunto *B* quedaría de la siguiente manera:

$B = \{B_4, B_5, B_6\}$

Posteriormente, se repite de nuevo el proceso. Como en las *fronteras internas* que quedan representadas en *B*, todos los elementos tienen igual número de ocurrencias y como el algoritmo no expresa ningún criterio de selección para tal caso, se puede escoger, aleatoriamente, cualquier elemento. Supongamos que se escoge el 4. Se inserta dicho elemento en *CENRMín*. A partir de esto, los elementos de dicho conjunto serían ahora los siguientes, *CENRMín* = {3, 4} y nuevamente, según el paso (8) del algoritmo, es eliminada la *frontera interna* B_4 de *B*. A partir de ello, dicho conjunto queda de la siguiente manera:

$B = \{B_5, B_6\}$

En la siguiente iteración, todos los elementos de las *fronteras internas* que quedan en *B* tienen igual número de ocurrencias, por tanto, se escoge, aleatoriamente, un elemento de cualquiera de ellas. Se selecciona el 2 de B_5. Tras ello, *CENRMín* = {3, 4, 2} y se debe eliminar de *B* a B_5, quedando

$B = \{B_6\}$

En el siguiente paso sucederá algo similar a lo descrito y se escoge en este caso el 1 de B_6. Se inserta en *CENRMín* y así, finalmente, dicho conjunto queda de la siguiente forma:

CENRMín = {3, 4, 2, 1}

B_6 es eliminado de B y así éste queda vacío y se alcanza la condición de parada del algoritmo.

El *CENRMín* resultante cumple con la definición dada. Posee un elemento por cada *frontera interna* y fue hallado siguiendo los pasos descritos en el algoritmo propuesto.

CONTRAEJEMPLO:

El *conjunto excepcional no redundante* {4, 2, 1} cumple también con la definición y sin embargo, es de menor cardinalidad que el hallado siguiendo los pasos del algoritmo.

Conclusiones del Análisis

Hasta aquí el análisis crítico sobre la propuesta teórica de (Jiang *et al.*, 2005). A partir del mismo se llega a las siguientes conclusiones:

- Los resultados aportan un marco teórico sin materializar una solución, lo cual permite desarrollar la idea.
- El enfoque propuesto es original pues no existen, al menos en la bibliografía revisada, antecedentes de otro con un planteamiento similar. No cae dentro de ninguna de las categorías que la bibliografía recoge para clasificar los métodos de detección según el principio en el cual se basan los mismos.

Universitat d'Alacant
Universidad de Alicante

- El método de detección propuesto es simple en cuanto al planteamiento teórico en que se basa, pero su implementación computacional a partir de la definición de *outlier* dada conduce a un problema no tratable:

  «Se *sabe que, dado un conjunto C, con |C|= n, la cantidad total de posibles subconjuntos de C (El Conjunto Potencia de C) es* 2^n*. Un algoritmo de detección según la definición de outliers dada, tendría que hallar siempre el Conjunto Potencia de X para, posteriormente, seleccionar a partir del mismo los conjuntos excepcionales no redundantes, de los cuales saldrían finalmente los outliers. Por ello, la complejidad temporal de dicho algoritmo sería* $\Omega(2^n)$»

- El modelo *Rough Set* ha sido aplicado con éxito a la solución de un gran número de problemas. En consecuencia, un método basado en dicho modelo, se espera que también sea eficaz al ser usado dentro de contextos en los que sea factible su aplicación, pero, como el procedimiento de detección propuesto se basa en el Modelo Básico de *RS*, hay que prestar especial atención a las limitaciones que tiene el mismo: Incapacidad para modelar información incierta. La *clasificación* con un *grado controlado de incertidumbre* o un posible *error de clasificación*, está fuera del alcance de este modelo. Sin embargo, en la práctica, poder admitir algún nivel de incertidumbre en el *proceso de clasificación* puede llevar a una comprensión más profunda y a una mejor utilización de las propiedades de los datos analizados. La definición estándar de *inclusión de conjunto* tenida en cuenta en el Modelo Básico de *RS* es demasiado rigurosa para modelar una inclusión de conjuntos *casi* completa. Por tanto, estas limitaciones le restan soporte al resultado.

Universidad de Alicante

- Uno de los elementos que puede limitar la efectividad de un método de detección de *outliers* es la naturaleza de los datos sobre los cuales se aplica el mismo. Sin embargo, el método propuesto es aplicable tanto a datos *continuos* como *discretos* (*categóricos*), lo cual es una ventaja con respecto a los métodos que presentan la limitación planteada.
- En la propuesta del método no se aprecia ningún elemento que pueda limitar su aplicación en *conjuntos de datos* de gran tamaño y alta dimensionalidad. Sin embargo, este es otro de los aspectos que limita la aplicación efectiva de varios métodos de detección recogidos en la bibliografía revisada.
- El método es aplicable a datos en forma tabular, con atributos monovaluados, para que exista la posibilidad de establecer, a partir de ellos, *relaciones de equivalencia*.

Tomando como antecedentes los elementos extraídos del estudio del *estado del arte* junto con el análisis realizado en este capítulo sobre la aplicación de la Teoría de *RS* a la detección de *outliers*, en el siguiente capítulo se propone una ampliación del marco teórico existente a partir de la cual es posible establecer un método de detección de *outliers* basado en el modelo básico de la Teoría de *RS* y computacionalmente viable.

Capítulo 4

Detección Eficiente de *Outliers* basada en *RSBM*

En este capítulo se aborda el **primer objetivo parcial** declarado como meta de la investigación:

> *«Ampliar el marco teórico de la teoría Rough Sets, de forma tal que pueda ser aplicable al problema de la detección de outliers, tomando como punto de partida las propuestas previamente realizadas en este ámbito; a partir de soluciones computacionalmente viables a dicho problema».*

O lo que es lo mismo, establecer un marco teórico para la consecución de un algoritmo computacionalmente eficiente para la detección de *casos excepcionales* basado en el *modelo básico de conjuntos aproximados* (*RSBM*).

Dicho objetivo plantea un *problema abierto* en la actualidad, pues tal y como se desprende del estudio del *estado actual de la técnica*, así como del análisis realizado en el **Capítulo 3** sobre los antecedentes previos, no existe,

en la actualidad, un marco teórico que permita proponer un método de detección de *outliers* basado en la Teoría de *Rough Sets* que sea computacionalmente viable. Sin embargo, ese mismo análisis nos permite establecer una nueva **subhipótesis de trabajo** para resolver el problema:

> *«La aplicación de la Teoría de Conjuntos Aproximados (Rough Sets Theory), (Pawlak, 1982), (Pawlak, 1991), al problema de la detección de outliers permite aprovechar elementos conceptuales inherentes a la misma para mejorar aspectos de los métodos actuales de detección de outliers, tales como: complejidad temporal, aplicación efectiva de los mismos en conjuntos de datos de gran tamaño y alta dimensionalidad, posibilidad de trabajar con atributos de valores heterogéneos, fundamentalmente».*

Por lo tanto, a partir de este planteamiento y de acuerdo al objetivo parcial establecido, en este capítulo se ofrece una **propuesta de solución** que consiste en:

> *«Una ampliación del marco teórico formal existente, de manera tal que, la extensión del mismo permita establecer un método de detección de outliers basado en RS que sea computacionalmente viable».*

La aplicación de este marco teórico permite establecer un nuevo método para la detección de *outliers* (basado en *RS*). Este método constituye uno de los principales aportes de esta investigación y su implementación consiste en un algoritmo, a partir del cual, se comprueba que la propuesta es computacionalmente viable.

El resto del capítulo se estructura de la siguiente forma: se propone y se demuestra un conjunto de resultados matemáticos que constituyen nuestra propuesta de

ampliación del marco teórico, a partir de la cual se establece un método de detección de *outliers* basado en *RS* y computacionalmente viable. La aplicación del marco formal alcanzado, desde el punto de vista del método científico, se concreta en un algoritmo para la detección de *outliers* basado en el *Modelo Básico de Conjuntos Aproximados* (*RSBM*). Sobre dicho algoritmo se ofrecen detalles en relación a su implementación computacional y se ilustra su funcionalidad a partir de un ejemplo concreto. La validación de los resultados con *conjuntos de datos* del mundo real nos permite dar criterios objetivos en cuanto al comportamiento del algoritmo respecto a su *complejidad temporal* y a la calidad de la detección. Finalmente, se establecen las conclusiones del capítulo.

Propuesta de Ampliación del Marco Teórico

Los siguientes *Lemas* y *Proposiciones* permiten ampliar el marco teórico existente y constituyen el fundamento matemático para la propuesta de un método de detección de *outliers,* computacionalmente viable, basado en *RSBM*.

Universitat d'Alacant
Universidad de Alicante

Preliminares:

Sea C el *concepto* y sea $\Re = \{r_1, r_2, ..., r_m\}$ un conjunto de *relaciones de equivalencia* definidas sobre un *universo* finto de datos U. Sea $X \subseteq U$, el conjunto de elementos de U que cumplen C y B_i la *frontera interna* de X con respecto a r_i.

En las demostraciones de los resultados teóricos que se exponen a continuación, cuando se hace referencia al complemento de cualquier conjunto, dicho complemento es con referencia al conjunto X.

Proposición 8:

$\forall e,e' \subseteq X$, si e es *excepcional* y $e \subseteq e'$, entonces e' es *excepcional* también.

Demostración: *trivial*

Lema 9:

Sea f un *conjunto excepcional no redundante* y sea a: $a \in X \wedge a \in f$, $\exists\ f \Leftrightarrow \exists i,\ 1 \leq i \leq m$, tal que $B_i^c \cup \{a\}$ es un *conjunto excepcional*.

Demostración:

($\Leftarrow$)

B_i^c , no es *excepcional*, pues no posee ningún elemento que pertenezca a la *frontera interna* B_i (por definición de complemento de un conjunto). Sin embargo, por hipótesis, $B_i^c \cup \{a\}$ es un *conjunto excepcional*, entonces a es un *elemento indispensable* en dicho conjunto. Bastaría con extraer de $B_i^c \cup \{a\}$ todos los *elementos dispensables* y obtener así un conjunto, al cual pertenece a, donde todos sus elementos son *indispensables* y por tanto, es un *conjunto excepcional* no redundante.

($\Rightarrow$)

Por hipótesis, $a \in f$ y f es *excepcional no redundante*. Entonces, a es *indispensable* en f, lo cual implica que $\exists i,\ 1 \leq i \leq m$, tal que $f \cap B_i = \{a\}$ (por definición de elemento *indispensable*), o sea, a sería el único representante en f de la *frontera interna* B_i, por tanto, $\forall y$ tal que $y \in f\text{-}\{a\}$ se cumple que $y \notin B_i$, lo cual a su vez implica que $y \in B_i^c$, por tanto, $f\text{-}\{a\} \subseteq B_i^c$.

Hagamos la unión de ambos conjuntos con el conjunto $\{a\}$:

$(f\text{-}\{a\}) \cup \{a\} \subseteq B_i^c \cup \{a\}$

$f \subseteq B_i^c \cup \{a\}$.

Como f es un *conjunto excepcional,* entonces $B_i^c \cup \{a\}$ lo será también por la **Proposición 8**.

Lema 10:

Si $\exists a \in B_i$, $1 \leq i \leq m$, tal que $B_i^c \cup \{a\}$ no es un *conjunto excepcional,* entonces, $\exists\, j$, $1 \leq j \leq m$, $j \neq i$, tal que $B_j \subseteq B_i$.

Demostración:

Si $B_i^c \cup \{a\}$ no es un *conjunto excepcional*, entonces existirá una j, $1 \leq j \leq m$, tal que: $\forall y \in B_i^c \cup \{a\}$, $y \notin B_j$.

Como $a \in B_i$, la *frontera interna* ausente no podrá ser B_i, por tanto, $j \neq i$. Si $\forall y \in B_i^c \cup \{a\}$, $y \notin B_j$, entonces $y \in B_j^c$.

Luego como $\forall y \in B_i^c \cup \{a\}$, $y \in B_j^c$, entonces,

$B_i^c \cup \{a\} \subseteq B_j^c \Rightarrow B_i^c \subseteq B_j^c \Rightarrow B_j \subseteq B_i$.

Aplicando el contrarecíproco del **Lema 10**, se enuncia el siguiente *Corolario*:

Corolario 11:

Si $\forall j$, $1 \leq i$, $j \leq m$, $j \neq i$, se cumple $B_j \not\subset B_i$ (lo cual quiere decir que la *frontera interna* B_i no contiene completamente a ninguna otra *frontera interna*), entonces, $\forall a \in B_i$, $B_i^c \cup \{a\}$ es un *conjunto excepcional*.

Universitat d'Alacant
Universidad de Alicante

Lema 12:

Sea $a \in X$. Si $\exists j, j \neq i, 1 \leq i, j \leq m$, tal que $B_j \subseteq B_i$ y $B_i^c \cup \{a\}$ es un *conjunto excepcional*, entonces, $B_j^c \cup \{a\}$ es también un *conjunto excepcional*.

Demostración:

$B_j \subseteq B_i \Rightarrow B_i^c \subseteq B_j^c \Rightarrow (B_i^c \cup \{a\}) \subseteq (B_j^c \cup \{a\})$

Luego, como $B_i^c \cup \{a\}$ es un *conjunto excepcional*, aplicando la **Proposición 8**, podemos concluir que: $B_j^c \cup \{a\}$ es también un *conjunto excepcional*.

Definición 13:

Sea f cualquier *conjunto excepcional no redundante*. Se define el conjunto E_i, $1 \leq i \leq m$, de la siguiente forma: $E_i = \{a: a \in X, a \in f, f \cap B_i = \{a\}\}$.

Resulta importante resaltar que a es el único elemento del conjunto f que pertenece a la *frontera interna* B_i.

El conjunto E_i contendrá a todos los elementos de X que pertenecen a algún *conjunto excepcional no redundante* (tenidos en cuenta todos ellos) y además son los únicos miembros de la *frontera interna* B_i en dichos conjuntos.

A partir de lo anterior, se llega a la siguiente conclusión:

$E = \bigcup_{i=1}^{m} E_i$, es el conjunto de todos los elementos de X que pertenecen a algún *conjunto excepcional no redundante*.

De igual forma, como consecuencia lógica de la secuencia de resultados teóricos presentados, podemos caracterizar a los elementos que no pertenecen a ningún *conjunto excepcional no redundante*: Supongamos que e es un

elemento cualquiera del *universo* que cumple el *concepto*. Entonces, *si* $e \in B_i$ y $\exists B_j$, tal que $B_j \subset B_i$ y $e \notin B_j$, $1 \leq i,j \leq m$, $i \neq j$, entonces *e* no podrá pertenecer a ningún *conjunto excepcional no redundante* en el cual él sea el representante de B_i.

Lema 14:

$\forall i$, $1 \leq i \leq m$, se cumple que $E_i \subseteq B_i$.

Esto quiere decir, en otras palabras, que todos los elementos de un E_i particular son elementos de la *frontera interna* B_i.

Demostración:

$\forall a \in E_i$, entonces, por definición de E_i, existe un *conjunto excepcional no redundante e tal que* $e \cap B_i = \{a\}$, *por tanto,* $a \in B_i$.

Lema 15:

Sea $a \in X$, $1 \leq i \leq m$, $B_i^c \cup \{a\}$ es un *conjunto excepcional,* si y sólo si, $a \in E_i$.

Demostración:

($\Rightarrow$)

Como $B_i^c \cup \{a\}$ es un *conjunto excepcional* en el cual *a* (ver demostración **Lema 9**) es un *elemento indispensable* (si se elimina al elemento *a* de dicho conjunto nos quedaría el conjunto B_i^c que se sabe que no es *excepcional* pues no posee ningún elemento de B_i), se puede obtener un conjunto $f \subseteq B_i^c \cup \{a\}$ tal que *f* sea un *conjunto excepcional no redundante* (eliminando de $B_i^c \cup \{a\}$ a los *elementos dispensables*) y en dicho conjunto, el único representante

de la *frontera interna* B_i es a (se infiere de los argumentos dados anteriormente), o sea $a \in E_i$.

($\Leftarrow$)

Si $a \in E_i$, entonces existirá un conjunto e, *excepcional no redundante*, que contiene al elemento a y $B_i \cap e = \{a\}$.

Uniendo el conjunto B_i^c a los conjuntos que están en ambos miembros de la igualdad, se tiene lo siguiente:

$$B_i^c \cup (B_i \cap e) = B_i^c \cup \{a\}$$

$$(B_i^c \cup B_i) \cap (B_i^c \cup e) = B_i^c \cup \{a\}$$

$$X \cap (B_i^c \cup e) = B_i^c \cup \{a\}$$

$$B_i^c \cup e = B_i^c \cup \{a\}$$

Luego, como e es *excepcional* y $e \subseteq (B_i^c \cup e)$, por la **Proposición 8**, $B_i^c \cup e$ es *excepcional* también y como $B_i^c \cup e = B_i^c \cup \{a\}$, entonces, $B_i^c \cup \{a\}$ es un *conjunto excepcional*.

Universitat d'Alacant
Universidad de Alicante

A partir del marco formal alcanzado, se puede proponer un método de detección, computacionalmente viable, basado en *RSBM*. Lo anteriormente expresado se valida mediante un algoritmo que implementa el método y el cual tiene *complejidad temporal no exponencial.* La descripción de ambos —método y algoritmo— se aborda en los siguientes apartados.

Método de Detección de *Outliers* basado en *RSBM*

Antes de proponer el método de detección, propiamente dicho, se precisan algunos aspectos esenciales que intervienen en su concepción:

Sea C el *concepto* a tener en cuenta y sea $\Re = \{r_1, r_2, ..., r_m\}$ un conjunto de *relaciones de equivalencia* (*criterios*) definidas sobre un *universo* finto de datos U. Sea $X \subseteq U$ el conjunto de elementos de U que cumplen C y sea $0 \leq \mu \leq 1$ un *umbral de excepcionalidad* establecido.

Pseudo-código del Algoritmo *RSBM*

El método de detección consta de dos fases fundamentales:

FASE 1. Construcción de *Fronteras Internas*

- Calcular las *fronteras internas* B_i, $1 \leq i \leq m$, de X (elementos de U que cumplen C) con respecto a cada elemento de $\Re$.
- Para cada elemento que cumple C y además pertenece a alguna *frontera interna* B_i, calcular su *grado de excepcionalidad*.

Como suposición teórica, sólo se tienen en cuenta *fronteras internas* diferentes y *no vacías*. Esta consideración se hace, sin pérdida de generalidad, pues si hay dos *fronteras internas* iguales, los elementos de una, que pertenecen a algún *conjunto excepcional no redundante*, serían los mismos en ambos casos y, por tanto, tenerlas duplicadas no aporta nada relevante en relación a los resultados que se obtienen. Por definición, las *fronteras internas vacías* no

Universitat d'Alacant
Universidad de Alicante

son tenidas en cuenta (Ver **Definición 2** - *conjunto excepcional*).

La *primera fase* del método se concreta en un algoritmo que implementa su funcionalidad —cálculo de *fronteras internas* y determinación del *grado de excepcionalidad*—. A continuación, se presenta una versión en *pseudo*-código del mismo.

(1) **for** *r* **in** $\Re$: // para cada *relación de equivalencia*:
(2) - clasificar cada elemento de *U* según *r* y definir la partición P_r
(3) **for** clase **in** P_r // por cada clase de equiv. en P_r
(4) **if** clase $\subseteq X$
(5) **then**
- por **Definición *Aproximación Inferior***
clase $\in \underline{X}$
(6) **else if** clase $\subseteq X^c$
(7) **then**
- por **Definición *Aproximación Superior***
clase $\notin \overline{X}$
(8) **else**
- por **Definición *frontera interna***:
(clase $\cap$ X) $\subseteq B_r$
// agregar los elementos de la *clase* que
// cumplen el *concepto* a la *frontera*
// *interna* relativa a *r*
(9) $B_r = B_r \cup$(clase $\cap$ X)
(10) **for** e **in** X:

Universitat d'Alacant
Universidad de Alicante

```
(11)     for r in ℜ :
(12)         if e∈ B_r
(13)         then
(14)             OD(e) = OD(e) + 1
(15)     OD(e) = OD(e) / m //calcula grado de excep.
```

Algoritmo 2 Cálculo de fronteras internas y grado de excepcionalidad – Algoritmo *RSBM*

FASE 2. Construcción del conjunto *E* y Detección de *Outliers*

Paso 1. Construcción del conjunto E

Construir el conjunto E —conjunto de todos los elementos de U que cumplen C y pertenecen a algún *conjunto excepcional no redundante*— a partir de los elementos de los conjuntos E_i:

- Analizar todas las *fronteras internas* y, en función de la relación de inclusión entre ellas, tomar diferentes decisiones y realizar determinadas acciones que no son más que la aplicación directa de algunos de los *Lemas* y *Corolarios* que fueron demostrados y que sirven de marco teórico a la concepción del método.

  Dos resultados trascendentes que se derivan de lo anteriormente expresado, son los siguientes:

  > *«Si para alguna frontera interna B_i se determina que no existe una frontera interna B_j, tal que $B_j \subset B_i$, entonces todos los elementos de B_i pertenecen a E_i, $i \neq j$, $1 \leq i,j \leq m$»*

  > *«Si para alguna frontera interna B_i se determina que existe una frontera interna B_j, tal que $B_j \subset B_i$, $i \neq j$, $1 \leq i,j$*

≤*m, entonces* $E_i \subseteq E_j$ *y la inclusión de los elementos de* E_i en el conjunto E depende de E_j »

La secuencia de deducciones teóricas que justifican estos planteamientos se expresa en detalle en el **Algoritmo 3** que se ilustrará seguidamente.

- Si en el análisis de un E_i particular se detecta algún elemento de X que con anterioridad ya había sido identificado como miembro de otro E_j y, por tanto, ya había sido incluido en E_j, éste no se tiene en consideración.

Paso 2. Detección de outliers

Una vez obtenido el conjunto E, determinar los elementos del mismo cuyo *grado de excepcionalidad* es mayor que el *umbral de excepcionalidad* (μ) previamente establecido. Todos los elementos que cumplan esta condición, se consideran *outliers*.

La *segunda fase* del método se concreta a partir del siguiente algoritmo —versión *pseudo*-código del mismo— donde se conforma el conjunto E y se establece el proceso de detección de *outliers*.

Universidad de Alicante

```
(1) for i:=1 to m
//iterar por todas las fronteras internas

(2)          if ∀j, 1≤j≠i≤m : Bj ⊄ Bi
(3)          then
// no existe j: Bj es subconjunto propio de Bi
                - por Corolario 11:
                  (∀j, 1≤j≠i≤m : Bj ⊄ Bi) ⇒
```

$\forall a \in Bi$, $Bi^{C} \cup \{a\}$ *conj. excep.*

- por ***Lema 15***:

$(\forall a \in Bi, Bi^{C} \cup \{a\}$ *conj. excep.*$) \Leftrightarrow$

$a \in Ei \Rightarrow B_i \subseteq Ei$ (I)

- por ***Lema 14***:

$E_i \subseteq B_i$ (II)

- por (I) y (II):

$Ei = Bi$

// si ninguna *frontera interna* es *subconjunto propio* de

// la *frontera interna* B_i, entonces todos los elementos

// de B_i conforman el conjunto E_i

(4) else $\exists j, 1 \leq j \neq i \leq m$, tal que $Bj \subset Bi$

// existe al menos una *frontera interna* B_j tal que

// B_j es *subconjunto propio* de B_i

(5) for a in Bi:

(6) if $Bi^{C} \cup \{a\}$ es un *conj. excep.*

(7) then

// si $Bj \subset Bi$ y para $a \in Bi$ se cumple que $B_i^{C} \cup \{a\}$ es un

// *conjunto excepcional*, entonces todo *elemento* de E_i

// estará incluido también en E_j

- por ***Lema 15***:

$(Bi^{C} \cup \{a\}$ conj. excep.$) \Leftrightarrow a \in Ei$ (III)

- por ***Lema 12***:

Universitat d'Alacant
Universidad de Alicante

$(Bi^C \cup \{a\}$ conj. excep.)$\wedge(Bj \subset Bi) \Rightarrow$

$Bj^C \cup \{a\}$ conj. excep.

- por ***Lema 15***:

$(Bj^C \cup \{a\}$ conj. excep.) $\Leftrightarrow a \in Ej$ (IV)

(8) else

// si $B_i{}^C \cup \{a\}$ no es un *conjunto excepcional*, entonces $a \notin E_i$

$\neg(Bi^C \cup \{a\}$ conj. excep.)

- por ***contrarecíproco del Lema 15***:

$\neg(Bi^C \cup \{a\}$ conj. excep.)$\Leftrightarrow a \notin Ei$ (V)

(9) - por (III), (IV) y (V): $(\forall a \in Ei \Rightarrow a \in Ej) \Rightarrow Ei \subseteq Ej$

// por tanto, la inclusión de los elementos del conjunto E_i

// en E se realizará cuando se conforme el conjunto E_j

// ya que todos los elementos de E_i pertenecen también a E_j

// IMPORTANTE: Esta situación no genera ciclos porque queda

// garantizado que $B_i \not\subset B_j$ (B_j es *subconjunto propio* de B_i)

(10) E= $\bigcup_{i=1}^{m} E_i$

(11) for e in E:

if $OD(e) \geq \mu$

then e es *outlier*

// una vez construido E, detectar los elementos del mismo

// cuyo *grado de excepcionalidad* sea mayor o igual que el

Universitat d'Alacant
Universidad de Alicante

```
// umbral μ previamente establecido. Dichos elementos, son
// los outliers
```

Algoritmo 3 Construcción del conjunto *E* y detección de *outliers*- Algoritmo *RSBM*

En el **Algoritmo 3** se presenta la justificación teórica de todas las alternativas previstas en el mismo. En cada paso se detallan los elementos del marco teórico que justifican cada una de las acciones o decisiones tomadas. No obstante, teniendo en cuenta los momentos del algoritmo en que realmente hay aportes a la *construcción* del conjunto *E,* y considerando estos los más trascendentes dentro del mismo, se presenta a continuación el **Algoritmo 3.1** como versión simplificada del inicial.

(1) **for each** *frontera interna* B_i, $1 \le i \le m$

(2) **if** *ninguna frontera interna* B_j, $j \neq i$, *es subconjunto propio de* B_i

(3) **then**

Todos los elementos de la frontera interna B_i *pertenecen a algún conjunto excepcional no redundante. Siendo cada elemento, en cada caso, el único representante de la frontera interna i en el conjunto excepcional no redundante al cual pertenece. Por todo lo anterior,* $E_i = B_i$

(4) $E = \bigcup_{i=1}^{m} E_i$

(5) *Todo elemento en E cuyo grado de excepcionalidad sea mayor o igual que* μ*, es un outlier en U*

Algoritmo 3.1 Versión simplificada del algoritmo de formación del conjunto *E* y detección de *outliers*- Algoritmo *RSBM*

Universitat d'Alacant
Universidad de Alicante

Los dos algoritmos que implementan las dos fases del método, constituyen el cuerpo principal del algoritmo de detección de *outliers* basado en *RSBM*. A continuación se dan detalles en relación a la implementación computacional del mismo.

Implementación Computacional. Algoritmo *RSBM*

Los parámetros de entrada del algoritmo son los siguientes:

- El universo U
- El *concepto* C
- Las *relaciones de equivalencia* r_i , $1 \leq i \leq m$, $\Re = \{r_1, r_2, ..., r_m\}$
- El *umbral de excepcionalidad* (μ) establecido

Las entradas del algoritmo se almacenan en *listas*. La *lista* es un tipo básico, por lo que no es necesario dar detalles con respecto a su funcionalidad.

La estructura de datos fundamental que se utiliza en el algoritmo es la de *diccionario*, entendiendo por esto un conjunto de pares (c*lave*, *valor*), donde *clave* es un objeto cualquiera al cual se le asocia un único objeto de tipo *valor*.

En el algoritmo, las *claves* se obtienen como resultado de aplicar un *clasificador* a un elemento cualquiera del *universo*. Dicho *clasificador* está asociado a una *relación de equivalencia* r_i particular, $1 \leq i \leq m$ y permite clasificar los miembros de las *clases de equivalencia* definidas por dicha relación. Los *valores* asociados a las *claves* son *listas* de elementos que pertenecen a la *clase de equivalencia* identificada por la *clave* asociada a dicho *valor*.

En general, para cada *relación de equivalencia* se construye un *diccionario* y a partir de todos ellos se construye una lista *m*-dimensional de diccionarios (**Figura 4-1**).

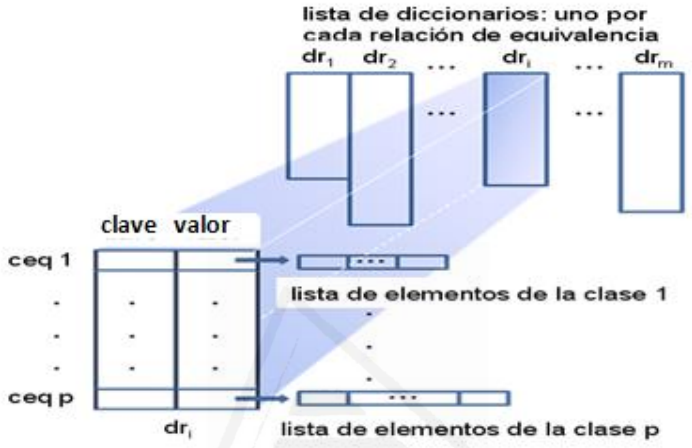


Figura 4-1 Estructuras de datos utilizadas

Complejidad Espacial

De acuerdo a las estructuras de datos utilizadas, puede decirse que la *complejidad espacial* del algoritmo, para el *caso peor*, es $O(n \times m)$, *n: cardinalidad del universo, m: número de relaciones de equivalencia tenidas en cuenta en el análisis*. Para establecer esta cota se tuvo en cuenta que cada *diccionario*, a lo sumo, puede contener a todos los elementos del *universo*. La memoria que ocupan las restantes estructuras de datos no supera el orden establecido.

Complejidad Temporal

El análisis de la *complejidad temporal* del algoritmo general, para el *caso peor*, se establece a partir de la *complejidad*

temporal de cada una de las dos fases anteriormente descritas:

Fase 1. Construcción de *Fronteras Internas*

- Aplicar los clasificadores (uno por cada *relación de equivalencia*) a todos los elementos del *conjunto de datos* con el objetivo de formar las *fronteras internas* con respecto a cada una de las *relaciones de equivalencia* tenidas en cuenta en el análisis: $O(n \times m \times c)$.

- Una vez concluido este proceso, calcular el *grado de excepcionalidad* de cada elemento del *universo* que cumple el *concepto* y además pertenece a alguna *frontera interna*: $O(n \times m)$.

La *complejidad temporal* de la **Fase 1**, para el *caso peor*, es: $O(n \times m \times c)$, *c:* costo de clasificar cada elemento.

Fase 2. Construcción del conjunto *E* y Detección de *Outliers*

Es una implementación concreta del **Algoritmo 3**.

La *complejidad temporal*, para el *caso peor*, es: $O(n \times m^2)$.

Teniendo en cuenta las **Fases 1** y **2**, la *complejidad temporal* general del algoritmo, para el *caso peor*, es $O(máx(O(Fase\ 1), O(Fase\ 2)))= O(Fase\ 2)=O(n \times m^2)$.

En general, el número de *relaciones de equivalencia* que intervienen en el análisis, en la inmensa mayoría de los casos, no es muy grande en relación al número de filas de la tabla (*conjunto de datos*). Por tanto, la dependencia cuadrática del tiempo de ejecución con respecto a la cantidad de *relaciones de equivalencia* no afecta en gran medida el tiempo de ejecución del algoritmo. Como se verá en los resultados obtenidos, esta dependencia cuadrática es *casi lineal* para valores pequeños de m ($m \leq 40$).

Universitat d'Alacant
Universidad de Alicante

El ejemplo que se muestra en el siguiente apartado ilustra la funcionalidad del algoritmo descrito.

Localización de *Outliers* dentro de un Conjunto de Datos

A continuación se muestra un ejemplo donde se ilustran aspectos esenciales de la funcionalidad del algoritmo. En él se considera un *universo U* que representa a 21 *pacientes* (**Tabla 4-1**). En la tabla, por cada *paciente*, en función de su *temperatura* y a partir de la existencia o no de *dolor de cabeza*, se establece un diagnóstico en cuanto al padecimiento de una *gripe*.

Tabla 4-1 Datos sobre *pacientes* que representan un *universo* dado U

ID	Dolor Cabeza	Temperatura	Diagnóstico
1	si	normal	desconocido
2	no	muy alta	gripe
3	si	alta	gripe
4	no	normal	desconocido
5	si	muy alta	gripe
6	no	alta	desconocido
7	no	alta	insolación
8	no	muy alta	gripe
9	si	normal	desconocido
10	si	normal	insolación
11	si	muy alta	gripe
12	no	normal	desconocido
13	si	normal	cefalea
14	si	normal	cefalea
15	no	alta	insolación
16	no	muy alta	gripe
17	no	muy alta	gripe
18	no	normal	desconocido
19	no	muy alta	gripe
20	si	alta	gripe
21	si	alta	desconocido

Universitat d'Alacant
Universidad de Alicante

Se definen dos *criterios* (*relaciones de equivalencia* tenidas en cuenta en el análisis). Cada uno de los cuales *particiona* a *U* en un número determinado de *clases de equivalencia.*

$$r_1 = \left\{ x \in U : \begin{Bmatrix} 1\_si\_dolor\_de\_cabeza(x) \\ 0\_no\_dolor\_de\_cabeza(x) \end{Bmatrix} \right\}$$

$$r_2 = \left\{ x \in U : \begin{Bmatrix} 0\_si\_temperatura\_Normal(x) \\ 1\_si\_temperatura\_Alta(x) \\ 2\_si\_temperatura\_MuyAlta(x) \end{Bmatrix} \right\}$$

concepto C = {x∈U ^ gripe(x)}

En la **Figura 4-2** se muestran las *clases de equivalencia* que forman parte de la partición de *U* que se crea a partir de *r1*. En ambas hay elementos que cumplen el *concepto* y elementos que no lo cumplen. Por tanto, ambas clases quedan dentro de la *frontera* de *C* respecto a r_1. Los elementos de ambas clases que cumplen el *concepto* constituyen la *frontera interna* respecto a r_1.

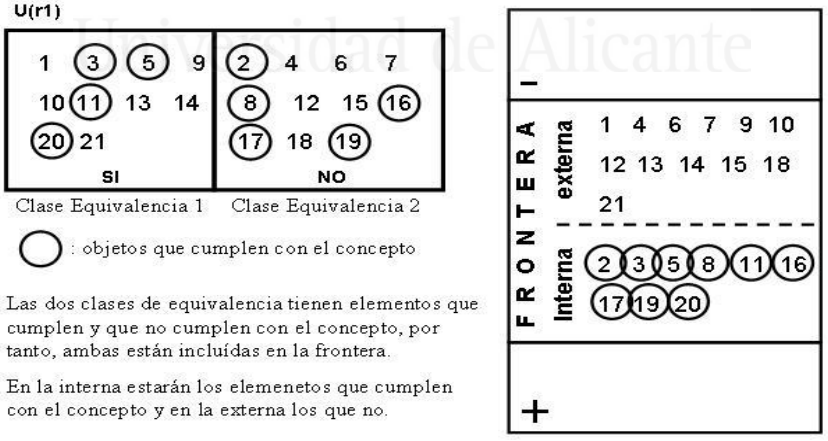


Figura 4-2 Partición que establece r_1 sobre *U* y frontera de *X* respecto a r_1

En la **Figura 4-3** se muestran las *clases de equivalencia* que forman parte de la partición de *U* que se crea a partir de r_2. En este caso, la *clase 1* no tiene elementos que cumplen el *concepto*, por tanto, dicha clase es irrelevante para el análisis. La *clase 2* tiene elementos que cumplen el *concepto* y elementos que no lo cumplen. Los que lo cumplen, constituyen la *frontera interna* respecto a r_2. Por su parte, todos los elementos de la *clase 3* cumplen el *concepto* y por ello dicha clase está incluida, completamente, en la *aproximación inferior* de r_2.

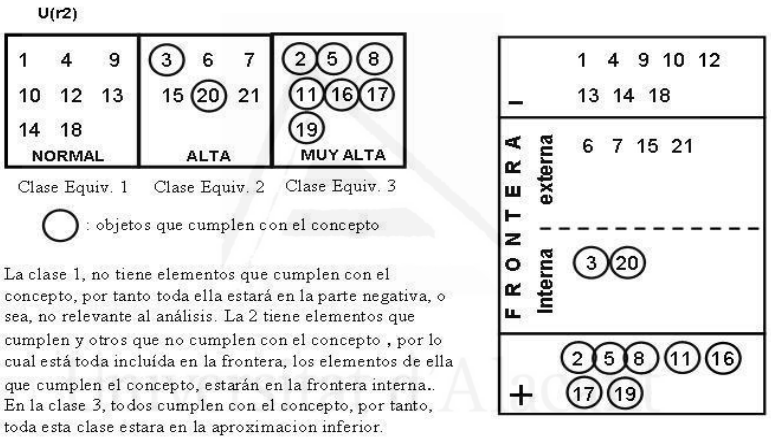


Figura 4.3 Partición que estable r_2 sobre *U* y *frontera* de *X* respecto a r_2

Aplicando el algoritmo, se obtienen las siguientes *fronteras internas*:

$B_1 = \{2, 3, 5, 8, 11, 16, 17, 19, 20\}$

$B_2 = \{3, 20\}$

El conjunto *E* calculado, que contiene todos los elementos de *U* que pertenecen a algún *conjunto excepcional no redundante,* sería el siguiente:

$E = \{3, 20\}$

El *grado de excepcionalidad* para los elementos de *E* sería:

Grado_de_excepcionalidad (3) = 1

Grado_de_excepcionalidad (20) = 1

Teniendo en cuenta que el valor del *grado de excepcionalidad* es un valor entre 0 y 1, se puede afirmar que ambos elementos serían considerados *outliers* para cualquier umbral μ dado.

Una interpretación de este hecho es que ambos elementos son contradictorios con el elemento 21 de la tabla, que también tiene los mismos síntomas que ellos y al que sin embargo, no se le diagnosticó *gripe*.

Validación de los Resultados

Prueba 4-1

Objetivo de la prueba: validar *orden de complejidad temporal* del algoritmo *RSBM*

***Conjunto de datos* utilizado:** *Adult Data Set*

Descripción del *conjunto de datos*: contiene datos extraídos del *Census Bureau Database of USA* (*CENSUS*, 2009)

Fuente: *UCI Machine Learning Repository*. *Center for Machine Learning and Intelligent Systems*. Universidad de California, Irvine (*UCI*, 2009)

El *UCI Machine Learning Repository* ofrece una colección de *conjuntos de datos* que son usados en las investigaciones relacionadas con *machine learning* y *data mining* para el análisis empírico de los algoritmos relacionados con estos temas.

Tipo del *conjunto de datos*: *multivariado.*

Tipo de los atributos: *categóricos* y *enteros.*

Cantidad de filas: 48.842

Cantidad de atributos (columnas): 14

Dispositivo de cálculo utilizado en la prueba:

INTEL Pentium 4, CPU 1.5 Ghz, 256 MB de RAM. Plataforma: Windows XP SP3

Descripción de la prueba: en las pruebas realizadas se tuvo en cuenta la variación de todos los parámetros que definen el tamaño de la entrada del algoritmo, es decir, tamaño (# de filas) y dimensionalidad (# de columnas) del *conjunto de datos*, así como el número de *relaciones de equivalencia* tenidas en cuenta en el análisis.

Los resultados que se muestran en la **Figura 4-4** permiten concluir que la dimensionalidad del *conjunto de datos* no influye en el tiempo de ejecución. Un aumento de la misma no representa un problema para que el algoritmo se ejecute correctamente. Los resultados alcanzados corroboran lo expresado en el análisis teórico de este aspecto.

Universitat d'Alacant
Universidad de Alicante

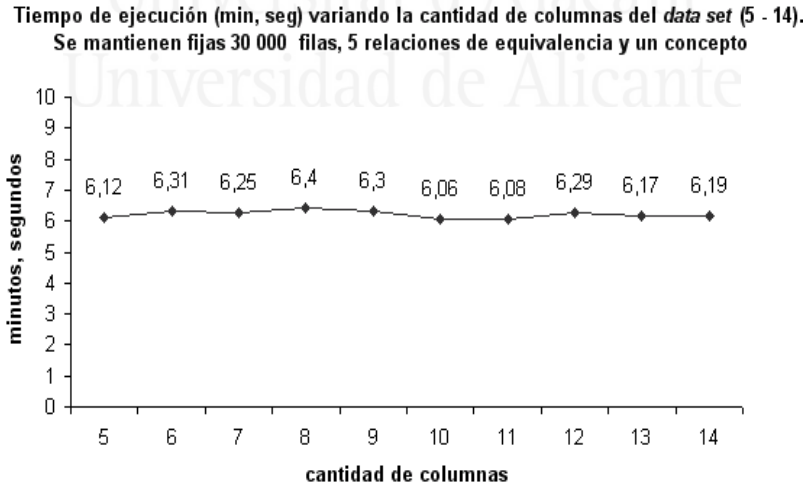


Figura 4-4 Variando número de columnas del conjunto de datos

Los resultados mostrados en la gráfica de la **Figura 4-5** reflejan los tiempos de ejecución alcanzados por el algoritmo haciendo variar la cantidad de filas del *conjunto de datos*. Se puede apreciar que la variación considerada oscila entre 5.000 y 30.000 filas, obteniéndose en todos los casos tiempos de ejecución razonables para el volumen de información procesado.

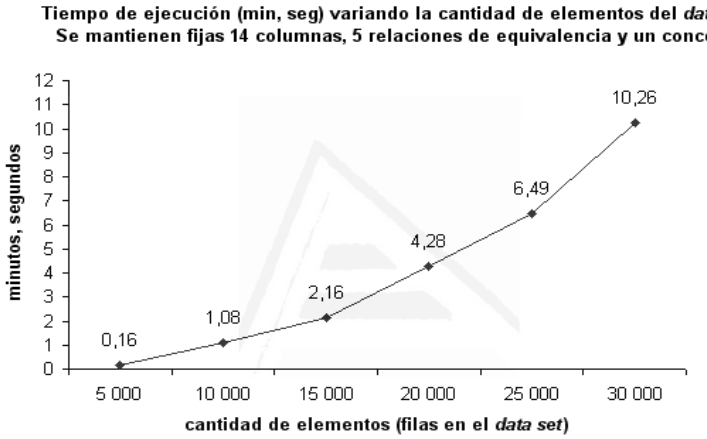


Figura 4-5 Variando la cardinalidad del conjunto de datos

Universitat d'Alacant
Universidad de Alicante

En la **Figura 4-6** se muestra la dependencia del tiempo de ejecución con respecto a la cantidad de *relaciones de equivalencia*. Teóricamente se analizó que dicha dependencia es cuadrática y en esta gráfica podemos ver que para valores pequeños de *m*, dicha dependencia se comporta *casi* lineal. Es decir, todo indica que las constantes definen una parábola muy abierta. Por tanto, para valores no muy grandes de *m*, que es lo más usual, se garantiza la casi linealidad de la *complejidad temporal* del algoritmo con respecto a dicho parámetro.

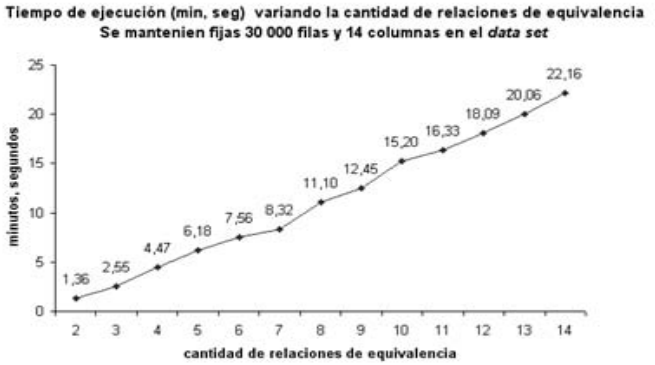


Figura 4-6 Variando el número de *relaciones de equivalencia*

Prueba 4-2

Objetivo de la prueba: validar *calidad de la detección* del algoritmo *RSBM*

***Conjunto de datos* utilizado:** *Adult Data Set*

Descripción del *conjunto de datos*: contiene datos extraídos del *Census Bureau Database of USA* (*CENSUS*, 2009)

Fuente: *UCI Machine Learning Repository. Center for Machine Learning and Intelligent Systems*. Universidad de California, Irvine (*UCI*, 2009)

Tipo del *conjunto de datos*: *multivariado.*

Tipo de los atributos: *categóricos* y *valores enteros.*

Cantidad de filas: 48.842

Cantidad de atributos (columnas): 14

Dispositivo de cálculo utilizado en la prueba: INTEL Pentium 4, CPU 1.5 Ghz, 256 MB de RAM. Plataforma: Windows XP SP3

Universitat d'Alacant
Universidad de Alicante

Descripción de la prueba: para esta prueba se seleccionaron los siguientes parámetros:

- Los individuos de la tabla que cumplían con el siguiente *concepto*: $1 \leq$ *personas_con_edad* ≤ 10, fueron la muestra estudiada.
- Los criterios a partir de los cuales se hizo el análisis quedaron establecidos por las siguientes *relaciones de equivalencia*:

r1: definida a partir del atributo categórico *workclass*

- c1.1: workclass = ['private' OR 'self-emp-not-inc' OR 'self-emp-inc' OR 'federal-gov local-gov' OR 'state-gov without-pay']
- c1.2: workclass = ['never-worked']

r2: definida a partir del atributo categórico *education*

- c2.1: education = ['bachelors' OR 'some-college' OR '11th' OR '9th' OR '7th-8th' OR '12th' OR '10th' OR 'HS-grad' OR 'prof-school' OR 'assoc-acdm' OR 'assoc-voc' OR 'masters' OR 'doctorate']
- c2.2: education = ['preschool' OR '1st-4th' OR '5th-6th']

r3: definida a partir del atributo categórico *marital-status*

- c3.1: marital-status = ['married-civ-spouse' OR 'divorced' OR 'separated' OR 'widowed' OR 'married-spouse-absent' OR 'married-AF-spouse']
- c3.2: marital-status = ['never-married']

r4: definida a partir del atributo categórico *occupation*

- c4.1: occupation = ['tech-support' OR 'craft-repair' OR 'other-service' OR 'sales' OR 'exec-managerial' OR 'prof-specialty' OR 'handlers-cleaners' OR 'machine-op-inspct' OR 'adm-clerical' OR 'farming-fishing' OR 'transport-moving' OR 'priv-house-serv' OR 'protective-serv' OR 'armed-Forces']
- c4.2: occupation = ['student']

Universitat d'Alacant
Universidad de Alicante

Cualquier elemento que cumpla el *concepto* y pertenezca a clase c*x*.1, con *x*: 1, 2, 3, 4, es contradictorio por la relación r_x, teniendo en cuenta que los individuos sujetos al análisis, por su *edad*, son *niños* entre 1 y 10 años.

Los elementos que aparecen en la **Tabla 4-2** representan un conjunto de *outliers* que, de forma intencional, se creó para *bombardear* el *conjunto de datos*.

Tabla 4-2 *outliers* introducidos en el conjunto de datos

Age	WorkClass	Education	Marital-Status	Occupation
7	***self-emp-inc***	1st-4th	never-married	student
6	never-worked	***masters***	never-married	student
9	never-worked	***doctorate***	never-married	student
9	never-worked	5th-6th	never-married	***Armed-Forces***
7	never-worked	1st-4th	never-married	***Adm-clerical***
8	***self-emp-inc***	***masters***	never-married	Student
8	never-worked	***doctorate***	***married-civ-spouse***	Student
6	never-worked	1st-4th	***divorced***	***Armed-Forces***
9	***federal-gov***	5th-6th	never-married	***Adm-clerical***
3	***self-emp-inc***	***masters***	***married-civ-spouse***	Student
7	never-worked	***doctorate***	***divorced***	***Adm-clerical***
2	***federal-gov***	***masters***	***divorced***	***Armed-Forces***
8	***self-emp-inc***	***doctorate***	***married-civ-spouse***	***Armed-Forces***

Universitat d'Alacant
Universidad de Alicante

Los valores de los restantes atributos para estos elementos son irrelevantes para el análisis, debido a que las *relaciones de equivalencia* con las que se trabaja sólo toman en consideración los atributos *workclass*, *education*, *marital status* y *occupation*. Los valores de los atributos resaltados

en **negrita** e *itálica* son contradictorios para niños con edades entre 1 y 10 años. Nótese que en la tabla el nivel de contradicción de los individuos varía. En algunos casos son contradictorios por uno o dos atributos, mientras que en otros, lo son por tres o por cuatro y éstos son, precisamente, los elementos más contradictorios.

La **Figura 4-7** muestra la cantidad de *outliers* detectados para diferentes valores del *umbral de excepcionalidad* μ a partir de diferentes pruebas en las que varió la cantidad de *outliers* introducidos en el *conjunto de datos*.

Interpretación de los resultados: como aspecto a destacar en los resultados alcanzados, cabe señalar que siempre en el conjunto de *outliers* detectados se encontraron los *outliers* introducidos. Esto se cumple tanto cuando la cantidad detectada fue mayor que la cantidad introducida así como cuando el número de *outliers* detectados fue menor que dicha cantidad.

Lo anterior refleja el nivel de eficiencia que puede alcanzar el algoritmo en cuánto a detección. Por otro lado, la variación del valor del umbral μ conduce a un cierto refinamiento en la detección, aunque en algunos casos no se logra. Esto se manifiesta en algunos estancamientos que se aprecian en las gráficas. En otras ocasiones, se deja bruscamente de detectar *outliers*, lo cual se evidencia en los pronunciados descensos a *cero* que ocurren. La causa de este escaso refinamiento parece estar provocada por el carácter determinista del método en lo que respecta a *clasificación*. Este hecho hace suponer que al permitirse un cierto *grado de desclasificación*, se puede lograr, en algunos casos, una mayor calidad de la detección. En consecuencia, se podrán obtener, finalmente, como *outliers* los elementos más contradictorios.

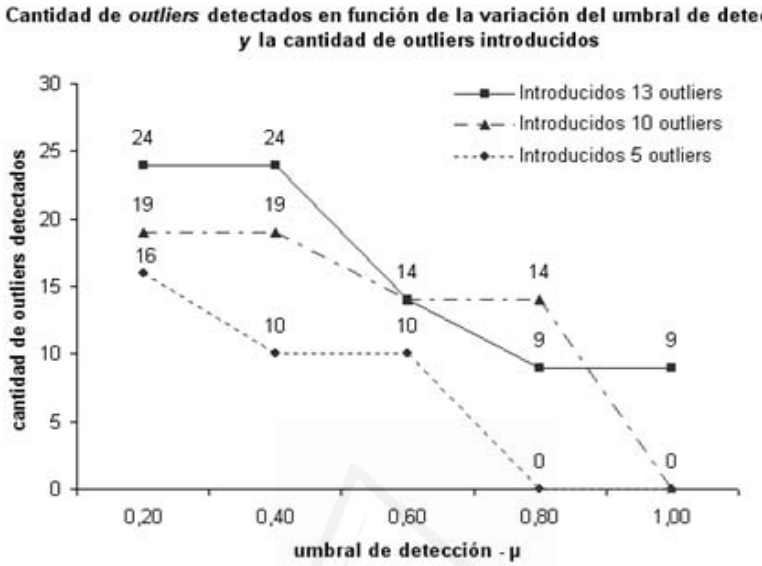


Figura 4-7 Resultados de las pruebas de detección

A continuación se enumeran las conclusiones parciales en relación al resultado alcanzado y expuesto en el presente capítulo.

Universitat d'Alacant
Universidad de Alicante

Conclusiones

Los resultados alcanzados pueden resumirse de la siguiente forma:

- Un marco teórico basado en elementos conceptuales del modelo básico de la Teoría de *RS* y en propuestas anteriores (Jiang *et al.*, 2005) que permitió establecer un método de detección de *outliers,* computacionalmente viable, basado en dichos elementos.
  - El enfoque del método es original, pues no existen antecedentes de otro, computacionalmente viable, con un planteamiento similar.

  - El método es aplicable a datos en forma tabular. En las tablas no debe haber redundancias. Sus atributos deben ser monovaluados, de lo contrario, entrarían en contradicción con la esencia del método pues no existiría la posibilidad de establecer a partir de ellos *relaciones de equivalencia*. Lo que acabamos de expresar delimita el ámbito de aplicación del problema.

- Un algoritmo —algoritmo *RSBM*— que permitió validar la viabilidad computacional del método propuesto. El orden de *complejidad temporal*, para el *caso peor*, de dicho algoritmo es *lineal* con respecto a la cardinalidad del *universo* de datos sobre el cual se aplica y *cuadrática* respecto al número de *relaciones de equivalencia* usadas. Este último parámetro representa una constante y su valor suele ser significativamente menor que la cardinalidad del *universo*, por lo cual no afecta significativamente a la *complejidad temporal* del algoritmo.

Las pruebas de validación realizadas al algoritmo con *conjuntos de datos* del mundo real verificaron que la existencia de valores heterogéneos (*continuos* y *discretos*) en los atributos, así como el tamaño y la dimensionalidad del *conjunto de datos*, no constituyen obstáculos para la ejecución del mismo. Esto supone ventajas del método propuesto con respecto a otros métodos existentes.

Los resultados generales alcanzados constituyen una solución al problema planteado a partir del objetivo parcial propuesto y han permitido verificar la subhipótesis de partida.

La principal limitación del método propuesto es que se basa en el modelo básico de *Rough Sets,* al cual se le critica su incapacidad para modelar información incierta. En este

modelo, no se permite la *clasificación* con un *grado controlado de incertidumbre*. En la práctica, poder admitir algún nivel de *incertidumbre* en el *proceso de clasificación*, conduce a una comprensión más profunda y a una mejor utilización de las propiedades de los datos analizados.

Como propuesta de solución para esta limitación, se propone extender el marco teórico existente, incorporando al mismo elementos del *Modelo de Conjuntos Aproximados de Precisión Variable VPRSM* (Ziarko, 1993), a partir de lo cual se elimine el carácter determinista de la propuesta actual.

En el siguiente capítulo se exponen los resultados alcanzados en esta dirección.

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Capítulo 5

Detección No Determinista de *Outliers* basada en *VPRSM*

Las limitaciones que se señalan en la primera versión del método de detección propuesto en el capítulo anterior, motivaron el planteamiento de un nuevo **objetivo parcial** en la investigación:

> *«Ampliar el marco teórico obtenido a partir del cumplimiento del objetivo parcial precedente, de manera tal que sea posible establecer un método de detección no determinista basado en el Modelo de Rough Sets de Precisión Variable —VPRSM— (Ziarko, 1993). El nuevo método debe superar las limitaciones de carácter determinista del modelo básico de RS, en lo relativo a la clasificación, y al mismo tiempo, debe seguir siendo computacionalmente viable.»*

El análisis de los elementos conceptuales que caracterizan a *VPRSM* realizado en el estudio del *estado del arte* permitió

establecer una nueva **subhipótesis de trabajo** en función del objetivo parcial propuesto:

> *«La introducción en el marco teórico existente de elementos conceptuales de VPRSM, nos permite ampliar la aplicación del método de detección basado en RSBM a contextos en los que sea necesaria una clasificación con un cierto grado de incertidumbre».*

A partir de este planteamiento y en correspondencia con el objetivo establecido, en el presente capítulo se ofrece una **propuesta de solución** que consiste en:

> *«Una ampliación del marco formal existente, de manera tal que la extensión del mismo permita establecer un método de detección de outliers no determinista —en cuanto a la clasificación—, basado en VPRSM y que mantenga la viabilidad computacional del método basado en RSBM —este nuevo método constituye un nuevo aporte de esta investigación—. La viabilidad computacional del método se valida a partir de la propuesta de un algoritmo que mantiene la complejidad temporal y espacial del algoritmo RSBM».*

Universidad de Alicante

De acuerdo con lo expuesto, el resto del capítulo se estructura de la siguiente forma: comparación general entre *RSBM* y *VPRSM*, que permite introducir los conceptos fundamentales que distinguen a *VPRSM* y los cuales sirven de base a la ampliación del marco formal propuesta. A su vez, el marco resultante constituye la base teórica-matemática de un nuevo método de detección no determinista de *outliers*, basado en *VPRSM*. La aplicación del mismo, desde el punto de vista del método científico, se concreta en un algoritmo que permite validar la viabilidad

computacional del método propuesto. Con posterioridad a la presentación del método y del algoritmo que lo implementa, se ofrecen detalles sobre la implementación computacional del mismo y se ilustra su funcionalidad a partir de un ejemplo concreto. La validación de los resultados con *conjuntos de datos* del mundo real permite dar criterios objetivos en cuanto al comportamiento del algoritmo en relación a su *complejidad temporal* y a la calidad de la detección. Finalmente, se presentan las conclusiones del capítulo.

RSBM frente a *VPRSM*

La *minería de datos* emerge cada vez con más fuerza, como un área de la Inteligencia Artificial que brinda técnicas, teorías y herramientas para el análisis de datos en los complejos *conjuntos de datos* de hoy en día. La Teoría de *RS* en este sentido ha incidido también, decisivamente, en el empeño por alcanzar tales metas. Desde finales de la década de los 80 ya se refieren resultados importantes de la aplicación de la misma en este campo y en los últimos años su aplicación en múltiples contextos de investigación pone de manifiesto su efectividad en la solución de problemas diversos. En el *estado del arte* expuesto en el **Capítulo 2** del presente trabajo, se han señalado varios ejemplos de aplicaciones recientes de esta teoría en diversos ámbitos.

El problema fundamental de la Teoría de *RS* es facilitar el análisis de la *clasificación*. La *aproximación* (*superior* e *inferior*) se hace necesaria ante la incapacidad de establecer, con el conocimiento disponible, clasificaciones completas de objetos que pertenezcan a una cierta *categoría* o *concepto*.

Con cierta frecuencia, la información disponible permite sólo hacer clasificaciones parciales. En tales casos, la teoría de *RS* puede utilizarse con efectividad para modelar este tipo de *clasificación* pero, a partir de este modelo, dicha *clasificación* debe ser completamente *correcta* o *cierta*. Esto limita la posibilidad de concebir, bajo el mismo, una *clasificación* con un *grado controlado de incertidumbre*, es decir, la posibilidad de que exista un cierto *error en la clasificación*. En la práctica, en muchos casos, resulta conveniente admitir algún *nivel de incertidumbre* en el *proceso de clasificación*, esto propicia una mejor comprensión y utilización de las propiedades de los datos que se están analizando. *RSBM* permite obtener hipótesis basadas sólo en *reglas de clasificación libres de errores* (las cuales se expresan en la *aproximación inferior*: $\underline{X}$) que se obtienen del análisis de los datos tenidos en consideración (U).

Los aspectos antes señalados, establecen el carácter determinista de *RSBM*. Sin embargo, hay múltiples situaciones en el mundo real que avalan la necesidad de tener también en cuenta *clasificaciones parcialmente incorrectas*. Una *regla de clasificación parcialmente incorrecta* proporciona también información útil. Puede establecer la tendencia de los valores si la mayoría de los datos disponibles a los que se aplica la regla pueden clasificarse correctamente. Precisamente, *VPRSM* brinda la posibilidad de detectar o expresar esta tendencia de la información y, a partir de ella, realizar determinados análisis sobre un cierto *universo de datos* (Ziarko, 2001), (Ziarko, 2002).

En el siguiente apartado se destacan los aspectos más relevantes del *Modelo de Conjuntos Aproximados de Precisión Variable —VPRSM—* que se orientan,

fundamentalmente, a resolver las limitaciones del modelo *RSBM* que hemos comentado.

VPRSM. Notaciones Básicas y Propiedades

VPRSM es una generalización de *RSBM*. Se deriva del mismo, sin supuestos adicionales. A partir de esta generalización, se permite el manejo de información con un *grado controlado de incertidumbre*. De igual forma a como se ilustró en relación a *RSBM*, se reportan diversos resultados investigativos relacionadas con la aplicación de *VPRSM*. Por solo citar algunos ejemplos, pueden señalarse los expuestos en: (Maheswari *et al.*, 2001), (Śle¸zak & Ziarko, 2002), (Bing-Zhen *et al.*, 2004), (Gong *et al.*, 2004), (Beynon & Driffield, 2005), (Beynon, 2006), (Su & Hsu, 2006), etc.

La definición estándar de *inclusión de conjuntos* tenida en cuenta en *RSBM* es demasiado rigurosa para modelar una *inclusión de conjuntos casi completa*. Sin embargo, el *Modelo de Conjuntos Aproximados de Precisión Variable* (*VPRSM*) subsana el carácter determinista de dicho modelo, estableciendo una nueva concepción de dicha relación: la *relación de inclusión mayoritaria*. Con ello, se permite establecer un cierto *grado de error en la clasificación*.

Primeramente, recordemos la definición de *relación de inclusión estándar*:

Definición 16 - Relación de inclusión estándar:

Sea U un *universo* finito de objetos. Sean X, $Y \subseteq U$, $X \neq \phi$, $Y \neq \phi$. Decimos que X está *incluido* en Y, o $X \subseteq Y$, si $\forall x \in X$, entonces, $x \in Y$.

La **Figura 5-1** ilustra esta definición.

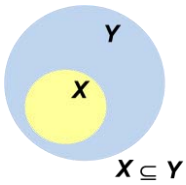


Figura 5-1 Inclusión de conjuntos estándar

Resulta evidente que de acuerdo a esta definición, no existe la posibilidad de ningún tipo de *desclasificación*. Antes de establecer una definición más general para esta relación, definamos lo que en *VPRSM* se denomina *medida del grado de desclasificación relativo del conjunto X con respecto al conjunto Y*, *c(X, Y)*.

$$c(X,Y)=\begin{cases}1-|X\cap Y|/|X| & \text{si } |X|\neq 0\\ 0 & \text{si } |X|=0\end{cases} \qquad (9)$$

De (9) pueden derivarse las siguientes interpretaciones:

- si $X\subseteq Y \Rightarrow |X\cap Y|=|X| \Rightarrow c(X, Y) = 0 \Rightarrow$
  no hay *error en la clasificación*.
- si $c(X, Y) \approx 1 \Rightarrow X, Y$ se acercan a ser *disjuntos*.
- si $c(X, Y) = 1 \Rightarrow |X\cap Y|=0 \Rightarrow X, Y$ son *disjuntos*.

La expresión numérica *c(X, Y)* es un indicativo del *error relativo de clasificación*. El producto $c(X, Y)\times|X|$ indica el *error absoluto de clasificación*, o sea, la cantidad de objetos *mal clasificados*.

Si se toma como base la *medida de desclasificación relativa*, se puede definir la *relación de inclusión*, obviando poner de

forma explícita el *cuantificador general*, de la siguiente forma:

$X \subseteq Y \Leftrightarrow c(X, Y) = 0$

Los elementos anteriormente expuestos sirven de base al establecimiento de una nueva definición:

Definición 17 - Relación de inclusión mayoritaria:

Sea U un *universo* finito de objetos. Sea β $(0 \leq \beta < 0{,}5)$ el *error de clasificación* admisible. Sean $X, Y \subseteq U$, $X \neq \phi$, $Y \neq \phi$. Decimos que *X está incluido mayoritariamente en Y*, o que *X está incluido en Y* con un *β-error*, $X \overset{\beta}{\subseteq} Y$ si y solo si $c(X, Y) \leq \beta$.

De la misma definición se deduce que $\beta=0$ expresa una *relación de inclusión estándar*, a la cual se le llama en este modelo, *inclusión total*.

A modo de ejemplo, en la **Figura 5-2** se muestra la *relación de inclusión mayoritaria* entre los cuatro conjuntos siguientes:

$X_1 = \{x_1, x_2, x_3, x_4\}$

$X_2 = \{x_1, x_2, x_5\}$

$X_3 = \{x_1, x_6, x_7\}$

$Y = \{x_1, x_2, x_3, x_8\}$

En dicha figura puede observarse el *grado de desclasificación* existente entre X_1, X_2, X_3 y el conjunto Y. Podemos apreciar cómo, a partir de la definición de *inclusión mayoritaria*, no se cumple $X_3 \overset{\beta}{\subseteq} Y$; pues entre esos dos conjuntos el *error de clasificación* es $\beta > 0{,}5$.

Universitat d'Alacant
Universidad de Alicante

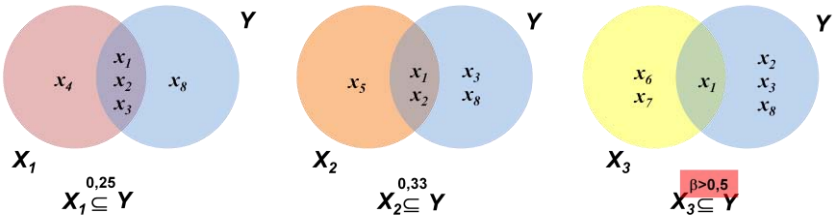


Figura 5-2 Ejemplo de inclusión mayoritaria

Se debe puntualizar que la relación de *inclusión mayoritaria* no es *transitiva,* como podemos ver en el ejemplo que se muestra en la **Figura 5-3**.

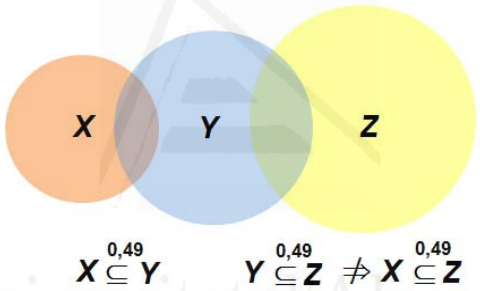


Figura 5-3 La relación de inclusión mayoritaria no es transitiva

A partir de la nueva definición para la *relación de inclusión*, se redefinen los conceptos más representativos de *RSBM* y de la propuesta de (Jiang *et al.*, 2005):

Definición 18:

Sea $X \subseteq U$ y $\theta \subseteq U \times U$ una *relación de equivalencia* que particiona U en un conjunto finito de *clases de equivalencia* $\langle x \rangle_\theta$. Se definen:

a) $\underline{X}_\beta = \cup \{\langle x \rangle_\theta : \langle x \rangle_\theta \overset{\beta}{\subseteq} X\}$ y se sabe que,

$$\langle x \rangle_\theta \overset{\beta}{\subseteq} X \Leftrightarrow c(\langle x \rangle_\theta, X) \leq \beta$$

b) $\overline{X_\beta} = \cup\{\langle x\rangle_\theta : \langle x\rangle_\theta \overset{\beta}{\not\subset} X^c\}$

A partir de b) puede deducirse que:

$\langle x\rangle_\theta \overset{\beta}{\not\subset} X^c \Leftrightarrow c(\langle x\rangle_\theta, X) < 1-\beta$

Demostración:

Si $\langle x\rangle_\theta \overset{\beta}{\subseteq} X^c \Leftrightarrow c(\langle x\rangle_\theta, X^c) \leq \beta$, por **definición 17**

$\forall A \subseteq U$, calculemos el valor de la suma $c(A, X^c) + c(A, X)$

$$\begin{aligned} c(A, X^c) + c(A, X) &= (1-|A\cap X^c|/|A|) + (1-|A\cap X|/|A|) \\ &= 2 - (|A\cap X^c| + |A\cap X|) / |A| \\ &= 2 - (|A| / |A|) = 1 \end{aligned}$$

Luego, $c(\langle x\rangle_\theta, X^c) + c(\langle x\rangle_\theta, X) = 1 \Leftrightarrow c(\langle x\rangle_\theta, X^c) = 1 - c(\langle x\rangle_\theta, X)$

Por lo que: $\langle x\rangle_\theta \overset{\beta}{\subseteq} X^c \Leftrightarrow 1 - c(\langle x\rangle_\theta, X) \leq \beta \Leftrightarrow 1 - \beta \leq c(\langle x\rangle_\theta, X)$

Por tanto aplicando el contra recíproco, tenemos

$\langle x\rangle_\theta \overset{\beta}{\not\subset} X^c \Leftrightarrow c(\langle x\rangle_\theta, X) < 1-\beta$

c) BN_β (región *β-frontera*) = $\overline{X_\beta} - \underline{X}_\beta$

d) B^β (región *β-frontera interna*) = $X \cap BN_\beta$

e) NEG_β (región *β-negativa*) = $\cup\{\langle x\rangle_\theta : \langle x\rangle_\theta \overset{\beta}{\subseteq} X^c\}$ y se sabe que, $\langle x\rangle_\theta \overset{\beta}{\subseteq} X^c \Leftrightarrow c(\langle x\rangle_\theta, X^c) \leq \beta \Leftrightarrow 1 - \beta \leq c(\langle x\rangle_\theta, X)$

Interpretando algunos de los elementos que se acaban de definir, podría decirse que:

Universitat d'Alacant
Universidad de Alicante

- La *aproximación β-inferior*, $\underline{X}_\beta$, de X incluye a todos los elementos de U que pertenecen a las *clases de equivalencia* $\langle x\rangle_\theta$, tal que $\underline{X}_\beta = \cup\{\langle x\rangle_\theta : \langle x\rangle_\theta \overset{\beta}{\subseteq} X\}$.

- La región *β-frontera*, BN_β, de X estará formada por todos aquellos elementos de U que no pueden ser clasificados ni dentro de X, ni dentro de X^c con un *error de clasificación* menor que β. O sea, $BN_\beta = \cup\{\langle x\rangle_\theta : \langle x\rangle_\theta \overset{\beta}{\not\subset} X^c \wedge \langle x\rangle_\theta \overset{\beta}{\not\subset} X\}$.

- La *aproximación β-superior*, $\overline{X_\beta}$, de X incluye a todos los elementos de U que no pueden ser clasificados dentro de X^c con un error menor que β. O sea, $\overline{X_\beta} = BN_\beta \cup \underline{X}_\beta$.

- La región *β-frontera interna*, B^β, de X incluye a todos los elementos de U que pertenecen X y a BN_β. O sea, $B^\beta = \cup\{\langle x\rangle_\theta \cap X : \langle x\rangle_\theta \in BN_\beta\}$.

- La región *β-negativa*, NEG_β, de X incluye a todos los elementos de U que pueden ser clasificados dentro de X^c con un *error* < β. O sea, $NEG_\beta = \cup\{\langle x\rangle_\theta : \langle x\rangle_\theta \overset{\beta}{\subseteq} X^c\}$

Universidad de Alicante

En la **Figura 5-4** puede apreciarse cómo *RSBM* es un caso particular de *VPRSM*. En ella se muestran las regiones representativas de *RSBM* para un *error de clasificación β=0*. En tal situación, *VPRSM* se corresponde con *RSBM*.

En la **Figura 5-5** se aprecia cómo varían las regiones significativas si se permite un cierto *error de clasificación*. En este caso, por ejemplo, se asumió *β=0,1*. Observe que la región *β-negativa* de X es la unión de todas las *clases de equivalencia* que pueden clasificarse dentro de X^c, con un *error de clasificación* no mayor que β.

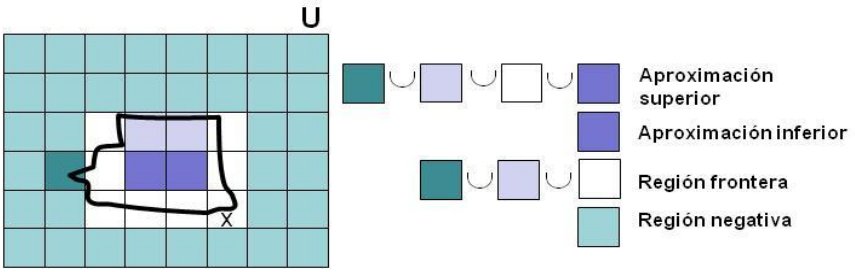


Figura 5-4 Regiones representativas para $\beta=0$. Correspondencia con *RSBM*

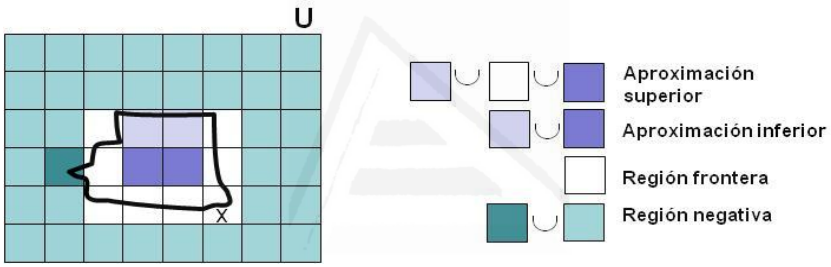


Figura 5-5 Variación de las regiones significativas para un *error de clasificación* $\beta=0,1$

RSBM es un caso particular de *VPRSM*, cuando $\beta=0$. A partir de esto, se establece la siguiente proposición que expresa la equivalencia entre los dos modelos.

Proposición 19:

Sea X un subconjunto arbitrario del *universo* U. Sea $\theta \subseteq U \times U$ una *relación de equivalencia* que particiona U en un conjunto finito de *clases de equivalencia* $\langle x \rangle_\theta$. Se tiene que:

- $\underline{X}_{\beta=0} = \underline{X}$, donde $\underline{X}$ es la *aproximación inferior* definida en *RSBM*, o sea, $\underline{X} = \cup \{\langle x \rangle_\theta : \langle x \rangle_\theta \subseteq X\}$.

- $\overline{X_{\beta=0}} = \overline{X}$, donde $\overline{X}$ es la *aproximación superior* definida en *RSBM*, o sea, $\overline{X} = \cup\{\langle x\rangle_\theta : \langle x\rangle_\theta \cap X \neq \phi\}$.
- $BN_{\beta=0} = BN$, donde *BN* expresa el concepto de *frontera* definido en *RSBM*, o sea, $BN = \overline{X} - \underline{X}$.
- $B^{\beta=0}$ (región *β-frontera interna*) = *B*, donde *B* expresa el concepto de *frontera* interna definido en *RSBM*, o sea, $B = X \cap BN$.
- $NEG_{\beta=0} = NEG$, donde *NEG* expresa el concepto de *región negativa* definido en *RSBM*, o sea, $NEG = U - \overline{X}$.

Además de la **Proposición 19** enunciada, para valores de β: $0 \leq \beta < 0{,}5$ se satisfacen las relaciones que se expresan en la siguiente proposición:

Proposición 20:

- a) $\underline{X} \subseteq \underline{X}_\beta$
- b) $\overline{X}_\beta \subseteq \overline{X}$
- c) $BN_\beta \subseteq BN$
- d) $NEG \subseteq NEG_\beta$

Intuitivamente, podemos darnos cuenta que cuando aumenta el *error de clasificación* β, el tamaño de la *región positiva* y de la *región negativa* de *X* aumenta, mientras que la región *frontera* disminuye. La **Figura 5-6** muestra esta variación de las *regiones significativas* tomando en consideración el β-error. Además, ilustra y resume muchas de las propiedades que han sido enunciadas.

Universitat d'Alacant
Universidad de Alicante

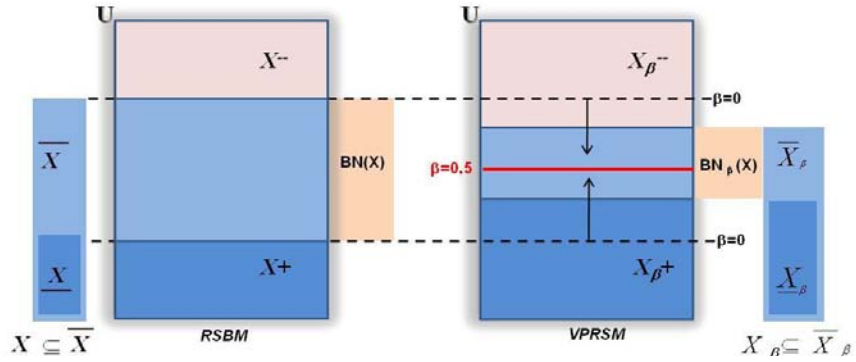


Figura 5-6 Variación de las regiones significativas a partir de la variación del *β-error*

Los aspectos teóricos expuestos constituyen un marco formal que permite establecer un *método de detección no determinista de outliers* basado en *VPRSM*, computacionalmente viable. Dicho método se describe en el siguiente apartado.

Método de Detección No Determinista de *Outliers* basado en *VPRSM*

La principal modificación hecha al diseño del método basado en *RSBM* radica en el cálculo de las *regiones significativas* según el marco teórico propuesto en *VPRSM*. En especial, en lo que respecta a la determinación de las *β-fronteras internas*. Como ya se ha señalado, en dicho modelo se permite un cierto *β-error* en la *clasificación*, a partir de lo cual se flexibilizan las relaciones de inclusión a la hora de establecer las *regiones significativas* del modelo en el marco del análisis. Con ello, se da la posibilidad de establecer una *clasificación casi completa*. Esto permite eliminar el carácter determinista de la misma, propio de *RSBM*.

Se establecen los siguientes elementos para el diseño del método:

Sea *C* el *concepto* a tener en cuenta; $\Re = \{r_1, r_2, ..., r_m\}$ un conjunto de *relaciones de equivalencia* (*criterios*) definidas sobre *U*; $0 \leq \mu \leq 1$ un *umbral de excepcionalidad* y $0 \leq \beta < 0,5$ el *error de clasificación* permitido.

Se mantienen los supuestos teóricos de base de *RSBM*: Sólo se tienen en cuenta *fronteras internas* diferentes y *no nulas*.

Pseudo-código del Algoritmo *VPRSM*

El nuevo método mantiene las dos fases del método de detección basado en *RSBM*:

Fase 1. Construcción de *β-Fronteras Internas*

- Se aplican los *clasificadores* (uno por cada *relación de equivalencia*) a los elementos de *U* para construir las *β-fronteras internas* (una por cada *relación de equivalencia* tenida en cuenta en el análisis).
- Para cada elemento del *universo* que cumple el *concepto* y pertenece a alguna *β-frontera interna,* se calcula su *grado de excepcionalidad.*

Esta *primera fase* se ejecuta mediante un algoritmo del que a continuación, se presenta una versión en *pseudo*-código, donde se justifican las decisiones o acciones tomadas a partir de los elementos del marco formal establecido.

```
(1) for r in ℜ: // para cada relación de equivalencia:
(2)      - clasificar cada elemento de U según r y definir
           la partición Pr
(3)      for clase in Pr: // por cada clase equiv. en Pr
(4)           if c(clase, X) ≤ β
```

(5) **then**

- por **Definición 18-a)**:

$\text{clase} \overset{\beta}{\subseteq} X \Rightarrow \text{clase} \in \underline{X}_{\beta}$

(6) **else if** c(clase, X) ≥ 1-β

(7) **then**

- por **Definición 18-b)**:

$\text{clase} \overset{\beta}{\not\subset} \overline{X} \Rightarrow \text{clase} \in \text{NEG}_{\beta}$

(8) **else**

- por **Definición 18-d)**:

$(\text{clase} \cap X) \subseteq B_r^{\beta}$

//agregar los elementos de clase que cumplen

//el concepto, a la β-frontera interna

//relativa a r

(9) $B_r^{\beta} = B_r^{\beta} \cup (\text{clase} \cap X)$

(10) **for** e in X:

(11) **for** r **in** $\Re$:

(12) **if** e∈ B_r^{β}

(13) **then**

(14) *OD(e) = OD(e) + 1*

(15) *OD(e) = OD(e) / m //calcula grado de excep.*

Algoritmo 4 Formación de las *β-fronteras internas* – Algoritmo *VPRSM*

Universitat d'Alacant
Universidad de Alicante

Fase 2. Construcción del Conjunto *E* y Detección de *Outliers*

Se construye el conjunto que contendrá todos los elementos que cumplen el *concepto* y estarán en algún *conjunto excepcional no redundante*, los cuales serán los candidatos a ser *outliers*. De ellos, todos aquellos cuyo *grado de excepcionalidad* sea mayor que el *umbral de excepcionalidad* μ establecido, serán clasificados como tal.

Esta fase del método se concreta a partir del siguiente algoritmo —versión *pseudo*-código del mismo—:

(1) E=ϕ

(2) Construir las β-*fronteras internas*

(3) **for** i:= 1 **to** m // por cada frontera interna

(4) **if** $\forall j,\ 1 \leq j \neq i < m : B_j^\beta \not\subset B_i^\beta$

// si ninguna β-*frontera interna* es

// subconjunto de la *i*, entonces,

// todos los

// elementos de la β-*frontera interna* B_i

// conforman el conjunto E_i

(5) **then**

(6) $E = E \cup B_i^\beta$

(7) *Outliers* = {x: x∈E ^ *OD(x)*≥ μ}

Algoritmo 5 Formación del conjunto *E* y detección de *outliers* – Algoritmo *VPRSM*

Universitat d'Alacant
Universidad de Alicante

Implementación Computacional. Algoritmo *VPRSM*

A las entradas del algoritmo *RSBM* se le añade el *β- error*, por lo que las entradas del algoritmo *VPRSM* son las siguientes:

- El *universo U.*
- El *concepto C.*
- Los criterios que distinguen a las *relaciones de equivalencia* tenidas en cuenta en el análisis (r_i, $1 \leq i \leq m$).
- El valor del umbral μ establecido.
- El *β-error*.

Se mantienen las mismas *estructuras de datos* descritas para el algoritmo *RSBM,* por tanto, el algoritmo *VPRSM* mantiene la *complejidad espacial* del anterior: $O(n \times m)$.

Teniendo en cuenta que las dos fases descritas constituyen el cuerpo principal del algoritmo *VPRSM*, su *complejidad temporal*, para el *caso peor*, queda definida en función de la *complejidad temporal* de cada una de ellas:

Fase 1 - $O(n \times m \times c)$

Fase 2 - $O(n \times m^2)$

Complejidad temporal, para el *caso peor*, del algoritmo *VPRSM*:

$O(máx(O(Fase\ 1), O(Fase\ 2)))= O(Fase\ 2)=O(n \times m^2)$

Donde:

c: costo de clasificar a cada elemento de *U*.

n: cardinalidad del *universo*.

m: número de *relaciones de equivalencia* tenidas en cuenta en el análisis.

De nuevo, consideramos oportuno señalar que el número de *relaciones de equivalencia* que intervienen en el análisis, en la inmensa mayoría de los casos, no es muy grande en relación al número de elementos del *conjunto de datos*. Por tal motivo, la dependencia cuadrática del tiempo de ejecución con respecto a la cantidad de *relaciones de equivalencia* no afecta, en gran medida, al *tiempo de ejecución* del algoritmo. Como se verá en los resultados obtenidos, esta dependencia cuadrática es *casi lineal* para valores pequeños ($m \leq 40$).

El ejemplo que se muestra a continuación ilustra la variación de las *regiones significativas* al permitirse un cierto *β-error* y, por tanto, cómo se flexibiliza la *clasificación*.

Localización de *Outliers* dentro de un Conjunto de Datos

Universitat d'Alacant
Universidad de Alicante

Se considera un *universo U* que representa a 25 *pacientes* (**Tabla 5-1**). En la tabla, por cada *paciente*, en función de su *temperatura* y a partir de la existencia o no de *dolor de cabeza*, se establece un diagnóstico relativo al padecimiento de una *gripe*.

Se definen dos criterios. Cada uno de ellos particiona *U* en un número determinado de *clases de equivalencia*.

$$r_1 = \left\{ x \in U : \left\{ \begin{matrix} 1\_si\_dolor\_de\_cabeza(x) \\ 0\_no\_dolor\_de\_cabeza(x) \end{matrix} \right\} \right\}$$

$$r_2 = \left\{ x \in U : \begin{Bmatrix} 0\_si\_temperatura\_Normal(x) \\ 1\_si\_temperatura\_Alta(x) \\ 2\_si\_temperatura\_MuyAlta(x) \end{Bmatrix} \right\}$$

concepto C = {x∈U ^ gripe(x)}

En la **Figura 5-7** se muestran las *clases de equivalencia* que forman parte de la partición de *U* que se crea a partir de r_1. En ambas, hay elementos que cumplen el *concepto C* y elementos que no lo cumplen, por tanto, ambas *clases* quedan dentro de la *frontera* respecto a r_1. Los elementos de ambas *clases* que cumplen *C* son los que forman la *frontera interna*. En la figura puede observarse, además, cómo queda la *clasificación* para *β=0* (*RSBM*) y permitiendo un *error de clasificación β=0,25*. Obsérvese que para r_1 todo se mantiene igual, en ambos casos.

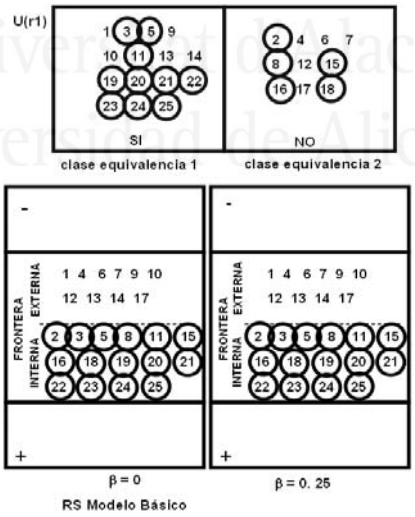


Figura 5-7 Partición que establece r_1 sobre *U* y *frontera* de *X* respecto a r_1. *β=0*; *β=0,25*

Tabla 5-1 Datos del ejemplo que representa al *universo* U

ID	Dolor Cabeza	Temperatura	Diagnóstico
1	SI	NORMAL	DESCONOCIDO
2	NO	MUY ALTA	GRIPE
3	SI	ALTA	GRIPE
4	NO	NORMAL	DESCONOCIDO
5	SI	MUY ALTA	GRIPE
6	NO	ALTA	DESCONOCIDO
7	NO	ALTA	INSOLACION
8	NO	MUY ALTA	GRIPE
9	SI	NORMAL	DESCONOCIDO
10	SI	NORMAL	INSOLACION
11	SI	MUY ALTA	GRIPE
12	NO	NORMAL	DESCONOCIDO
13	SI	NORMAL	CEFALEA
14	SI	NORMAL	CEFALEA
15	NO	MUY ALTA	GRIPE
16	NO	MUY ALTA	GRIPE
17	NO	NORMAL	DESCONOCIDO
18	NO	MUY ALTA	GRIPE
19	SI	ALTA	GRIPE
20	SI	ALTA	GRIPE
21	SI	ALTA	GRIPE
22	SI	ALTA	GRIPE
23	SI	ALTA	GRIPE
24	SI	ALTA	GRIPE
25	SI	ALTA	GRIPE

Universitat d'Alacant
Universidad de Alicante

Sin embargo, cuando se analiza lo que sucede con relación a r_2 (situación ilustrada en la **Figura 5-8**), se observa que la *clase de equivalencia* 2 —que, al trabajar con el modelo básico, quedaba en la *frontera* y no clasificaba dentro de la *región positiva,* aun cuando más del 80% de sus elementos cumplían con el *concepto*— al permitirse un *error de clasificación* $\beta=0,25$ pasa a la *región positiva*. Esto tiene

mucho más sentido teniendo en cuenta el porcentaje de elementos de la misma que cumplen *C*.

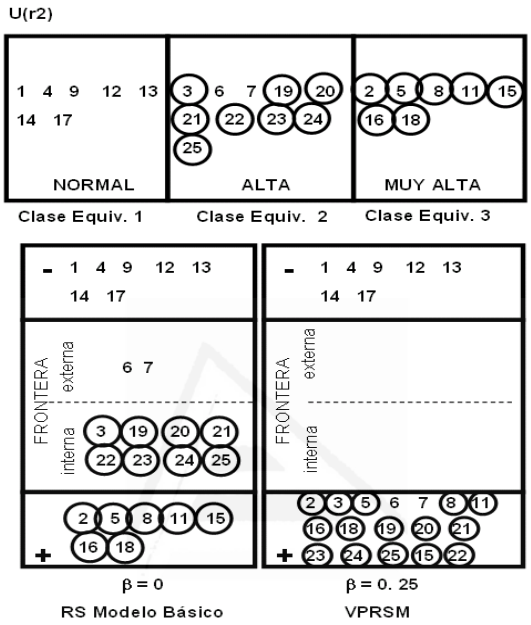


Figura 5-8 Partición que establece r_2 sobre *U* y *frontera* de *X* respecto a r_2. $\beta=0$; $\beta=0,25$

Validación de los Resultados

Prueba 5-1

Objetivo de la prueba: comparar *tiempos de ejecución* entre los algoritmos *RSBM* y *VPRSM*

***Conjunto de datos* utilizado:** *Adult Data Set*

Descripción del *conjunto de datos*: contiene datos extraídos del *Census Bureau Database of USA* (*CENSUS*, 2009)

Fuente: *UCI Machine Learning Repository. Center for Machine Learning and Intelligent Systems*. Universidad de California, Irvine (*UCI*, 2009)

Tipo del *conjunto de datos*: *multivariado.*

Tipo de los atributos: *categóricos* y *enteros.*

Cantidad de filas: 48.842

Cantidad de atributos (columnas): 14

Dispositivo de cálculo utilizado en la prueba: INTEL Pentium 4, CPU 1.5 Ghz, 256 MB de RAM. Plataforma: Windows XP SP3

Descripción de la prueba: la prueba tuvo como objetivo fundamental establecer una comparación entre el algoritmo *RSBM* y el *VPRSM* en cuanto a *tiempo de ejecución*.

Las **Figura 5-9**, **Figura 5-10**, y **Figura 5-11** reflejan los *tiempos de ejecución alcanzados* por los algoritmos *VPRSM* y *RSBM* en tres situaciones diferentes:

Figura 5-9: se mantuvieron fijos el número de filas del *conjunto de datos* (30.000), la cantidad de *relaciones de equivalencia* consideradas en el análisis (5) y el *concepto*. En este caso, se varió la cantidad de columnas del *conjunto de datos*.

Figura 5-10: se mantuvieron fijas las filas (30.000) y las columnas (14) del *conjunto de datos*, variándose el número de *relaciones de equivalencia* consideradas en el análisis.

Figura 5-11: la variación se hizo con respecto a la cardinalidad del *conjunto de datos.*

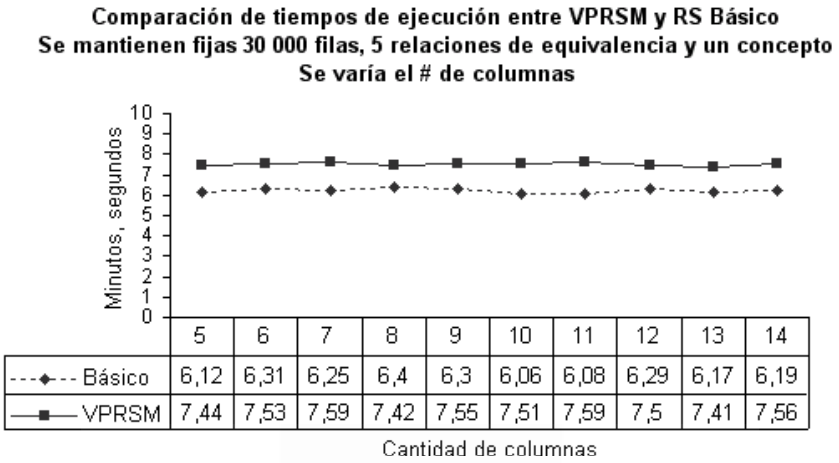


Figura 5-9 *RS vs. VPRSM* en cuanto a tiempo de ejecución. Variando número de columnas

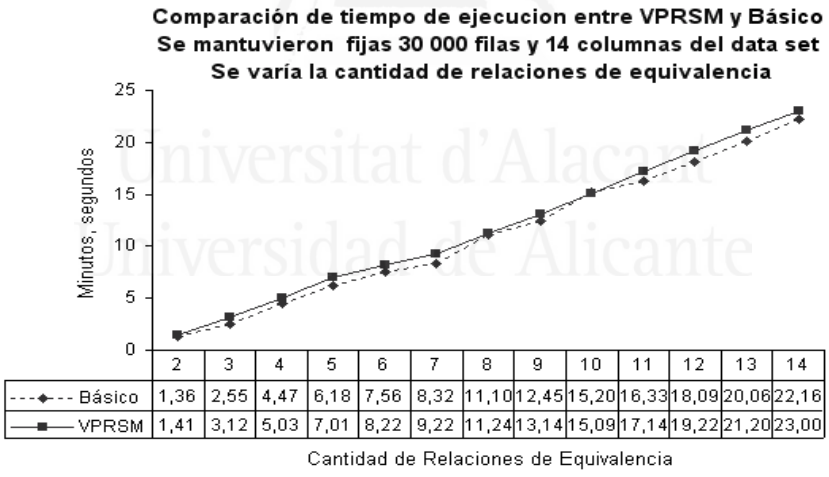


Figura 5-10 *RS vs. VPRSM* en cuanto a tiempo de ejecución. Variando número de relaciones

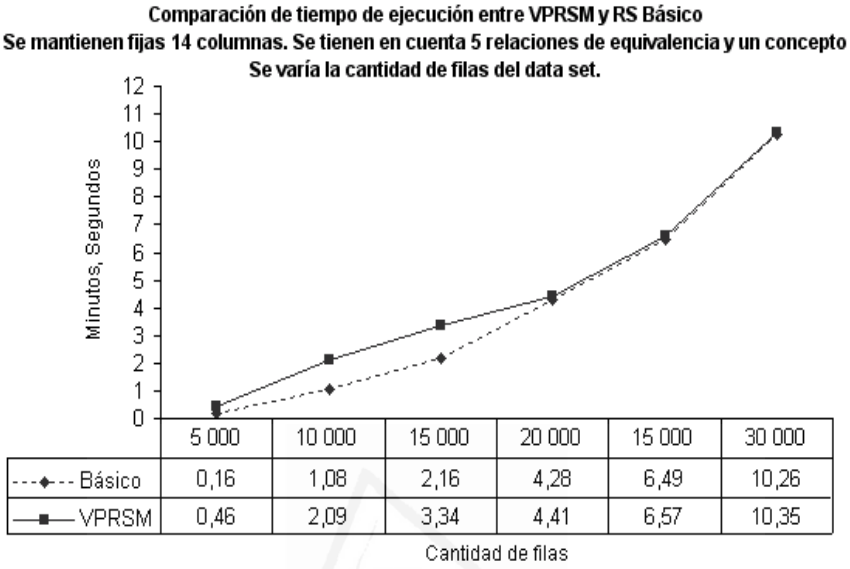


Figura 5-11 *RS vs. VPRSM* en cuanto a tiempo de ejecución. Variando número de filas

Interpretación de los resultados: los *tiempos de ejecución* entre los dos algoritmos fueron similares en todas las pruebas realizadas. Por tanto, se corrobora que el algoritmo *VPRSM* mantiene el mismo *orden de complejidad temporal* que el algoritmo *RSBM.*

Prueba 5-2

Objetivo de la prueba: validar *tiempo de ejecución* del algoritmo *VPRSM* al trabajar con *conjuntos de datos* de *gran tamaño* y *alta dimensionalidad*

***Conjunto de datos* utilizado:** *conjunto de datos sintético*

Descripción del *conjunto de datos*: *conjunto de datos aleatorio generado automáticamente* a partir del uso de técnicas estadísticas

Tipo del *conjunto de datos*: *multivariado.*

Tipo de los atributos: *categóricos* y *continuos*

Cantidad de filas: 500.000

Cantidad de atributos (columnas): 100

Dispositivo de cálculo utilizado en la prueba: procesador: *Intel(R) Core(TM)2 Quad CPU Q6600 @ 2.40Ghz 2.39Ghz*

Memoria: *3.25GB.* Sistema Operativo: *Windows 7 Ultimate*

Descripción de la prueba: la prueba tuvo como objetivo fundamental medir tiempo de ejecución del algoritmo *VPRSM* con un *conjunto de datos* de *gran tamaño* y *alta dimensionalidad*. En esta ocasión, la misma fue realizada con un *conjunto de datos sintético*, es decir, un *conjunto de datos aleatorio generado automáticamente*. A continuación se describe el proceso de generación de dicho conjunto.

Proceso de generación del *conjunto de datos sintético*

El proceso parte de saber cómo generar un *valor aleatorio* para cada uno de los atributos que definen el *conjunto de datos*. Un *valor aleatorio* para cada atributo, establece entonces una *fila aleatoria*, y muchas *filas aleatorias* constituyen un *conjunto de datos aleatorio*.

Universitat d'Alacant
Universidad de Alicante

El proceso seguido para la generación de *valores aleatorios* para los atributos es el siguiente:

- Si el atributo es *categórico*:

  Llamémosle *AC* a dicho atributo y sean $value_1$, $value_2$,..., $value_n$, los valores posibles para el mismo. Usando cualquier *herramienta de generación de números aleatorios*, se puede generar un número aleatorio *x;* tal que $x \in [1; n+1)$, con *distribución uniforme* —una *distribución* es *uniforme* cuando todos los posibles valores tienen la misma probabilidad de ser elegidos—, entonces, a partir de *x* se determina el valor aleatorio de *AC* de la siguiente forma:

- calcular y=*[x], —parte entera de x—*
- *AC=valor*$_y$.

- Si el atributo es *numérico*:

  Llamémosle *AN* a dicho atributo, donde *mín* ≤ *AN* ≤ *máx*. Usando cualquier *herramienta de generación de números aleatorios*, se puede generar un número aleatorio *x,* tal que *mín* ≤ *x* ≤ *máx,* con *distribución uniforme*, entonces, a partir de *x* se determina el valor aleatorio de *AN* de la siguiente forma: *AN=x.*

  En este proceso de generación del *conjunto aleatorio de datos*, debe establecerse el *dominio* de cada atributo. Es decir, definir *n* si el atributo es *categórico* y definir *mín* y *máx* si el atributo es *numérico*. A tales efectos, en esta ocasión, se establecieron los siguientes valores para dichos dominios:

  - *n* = 100 para todos los atributos *categóricos*.
  - *mín* = 0 y *máx* = 500 para todos los atributos *numéricos*.

De igual forma, se deben establecer cuántos atributos serán *categóricos* y cuántos *numéricos*. En este caso, se realizó el proceso de generación usando 50 atributos *categóricos* y 50 atributos *numéricos*.

Teniendo en cuenta que el objetivo de la prueba era medir *tiempo de ejecución* y no *calidad de la detección*, no resulta trascendente tener en cuenta aspectos semánticos en los atributos generados.

Los restantes parámetros que constituyen entradas del algoritmo fueron generados siguiendo los siguientes criterios:

- *relaciones de equivalencia* consideradas en el análisis.

Universitat d'Alacant
Universidad de Alicante

Por la incidencia que tiene el número de *relaciones* consideradas en la funcionalidad del algoritmo, se hizo que dicho número coincidiera con el número de columnas del *conjunto de datos* generado, es decir, 100.

Las *relaciones de equivalencia* fueron generadas a partir del siguiente criterio:

- Si el atributo asociado a la columna es *categórico*, entonces cada posible valor del mismo establece una *clase de equivalencia*.
- Si el atributo asociado a la columna es *numérico*, entonces cada número entero *n; mín=0 ≤ n ≤ máx=500*, establece una *clase de equivalencia* a la que pertenecen todos los valores en el intervalo [*n*; *n+1*).

- *concepto*

Para generar el *concepto C*, se buscó que, aproximadamente, la mitad de los elementos del *conjunto de datos* cumpliera el mismo. Teniendo en cuenta que la *herramienta* usada para la generación de valores aleatorios *distribuye uniforme*, lo anterior se pudo lograr de la siguiente forma:

Se seleccionó, de manera aleatoria, un *atributo categórico A* dentro de todo el conjunto de atributos.

A partir de dicho atributo, se definió que un elemento *x* cumple *C* si el valor de su atributo *A* es uno de los primeros *n/2* valores posibles de dicho atributo.

Veamos un ejemplo que ilustra esto:

Suponiendo que el atributo *A* tiene 100 valores posibles, sean $value_1$, $value_2$, … , $value_{100}$, entonces, se dice que el elemento *x* cumple *C* si en su atributo *A* tiene cualquiera de los siguientes valores: $value_1$, $value_2$, $value_3$,…, ó $value_{50}$.

El número de filas del *conjunto de datos* generado inicialmente fue 100.000 y, posteriormente, se fue incrementando en 100.000, hasta llegar a 500.000.

La **Figura 5-12** muestra los resultados conseguidos en esta prueba.

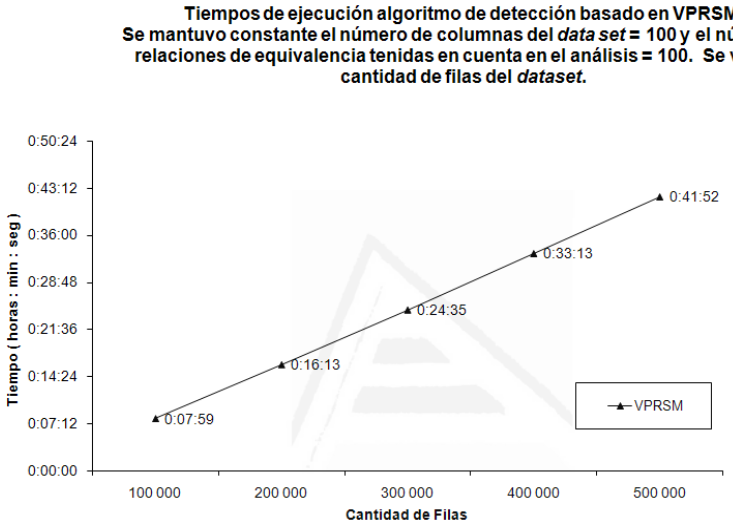


Figura 5-12 Tiempo de ejecución del algoritmo *VPRSM* sobre un *conjunto de datos* generado de forma *sintética*

Interpretación de los resultados: los resultados demuestran que el algoritmo es computacionalmente eficiente al ejecutarse sobre un *conjunto de datos* de *gran tamaño* y *alta dimensionalidad*. Se aprecia la *linealidad* en los tiempos de ejecución alcanzados por el mismo para dicho *conjunto de datos*.

Prueba 5-3

Objetivo de la prueba: validar *calidad de la detección* del algoritmo *VPRSM*

***Conjunto de datos* utilizado:** *Adult Data Set*

Descripción del *conjunto de datos*: contiene datos extraídos del *Census Bureau Database of USA* (*CENSUS*, 2009)

Fuente: *UCI Machine Learning Repository*. *Center for Machine Learning and Intelligent Systems*. Universidad de California, Irvine (*UCI*, 2009)

Tipo del *conjunto de datos*: *multivariado.*

Tipo de los atributos: *categóricos* y *enteros.*

Cantidad de filas: 48.842

Cantidad de atributos (columnas): 14

Dispositivo de cálculo utilizado en la prueba: INTEL Pentium 4, CPU 1.5 Ghz, 256 MB de RAM. Plataforma: Windows XP SP3

Descripción de la prueba: los individuos del *conjunto de datos* que fueron objeto de estudio son los que cumplían el siguiente *concepto*: $1 \leq personas\_con\_edad \leq 10$.

Los criterios a partir de los cuales se hizo el análisis quedaron establecidos por las siguientes *relaciones de equivalencia*:

r1: definida a partir del atributo categórico "workclass"

-c1.1: workclass =['private' OR 'self-emp-not-inc' OR 'self-emp-inc' OR 'federal-gov local-gov' OR 'state-gov without-pay']

-c1.2: workclass = ['never-worked']

r2: definida a partir del atributo categórico "education"

-c2.1: education = ['bachelors' OR 'some-college' OR '11th' OR '9th' OR '7th-8th' OR '12th' OR '10th' OR 'HS-grad' OR 'prof-school' OR 'assoc-acdm' OR 'assoc-voc' OR 'masters' OR 'doctorate']

-c2.2: education = ['preschool' OR '1st-4th' OR '5th-6th']

r3: definida a partir del atributo categórico "marital-status"

Universitat d'Alacant
Universidad de Alicante

```
    -c3.1: marital-status = ['married-civ-spouse' OR 'divorced'
        OR 'separated' OR 'widowed' OR 'married-spouse-absent'
        OR 'married-AF-spouse']

    -c3.2: marital-status = ['never-married']

  r4: definida a partir del atributo categórico "occupation"

    -c4.1: occupation = ['tech-support' OR 'craft-repair' OR
        'other-service' OR 'sales' OR 'exec-managerial' OR
        'prof-specialty' OR 'handlers-cleaners' OR 'machine-op-
        inspct' OR 'adm-clerical' OR 'farming-fishing' OR
        'transport-moving' OR 'priv-house-serv' OR 'protective-
        serv' OR 'armed-Forces']

    -c4.2: occupation = ['student']
```

Cualquier elemento que cumpla el *concepto* y pertenezca a la clase c*x*.1, *x*: 1, 2, 3, 4, es contradictorio por la relación *rx*, pues los individuos sujetos al análisis son *niños entre 1 y 10 años*.

En el *conjunto de datos* se introdujo intencionalmente un conjunto de *outliers,* que se muestra en la **Tabla 5-2**. En ella, sólo se reflejan los valores de los atributos que son relevantes para el análisis. Los restantes atributos son obviados por ser irrelevantes para el mismo. En esta tabla, además, puede observarse que hay valores de atributos resaltados en **negrita** e *itálica,* lo que significa que dichos valores son contradictorios para *niños con edades entre 1 y 10 años*. En el conjunto de *outliers* introducido, el nivel de contradicción de los individuos varía. En algunos casos, son contradictorios por uno o dos atributos, mientras que, en otros, lo son por tres o por cuatro y éstos son, precisamente, los elementos más contradictorios.

La gráfica que se presenta en la **Figura 5-13** muestra la cantidad de *outliers* detectados para diferentes valores de los umbrales β y μ. Los resultados que corresponden a *RSBM* son exactamente los alcanzados para $\beta=0$. Los valores $\beta=0,1;\ 0,2;\ 0,3;\ 0,4$ y $0,5$ establecen los diferentes valores del *error admitido en la clasificación*.

Tabla 5-2 *Outliers* introducidos en el *conjunto de datos*: *Census Bureau Database*

Age	WorkClass	Education	Marital-Status	Occupation
7	***self-emp-inc***	1st-4th	never-married	student
6	never-worked	***masters***	never-married	student
9	never-worked	***doctorate***	never-married	student
9	never-worked	5th-6th	never-married	***Armed-Forces***
7	never-worked	1st-4th	never-married	***Adm-clerical***
8	***self-emp-inc***	***masters***	never-married	Student
8	never-worked	***doctorate***	***married-civ-spouse***	Student
6	never-worked	1st-4th	***divorced***	***Armed-Forces***
9	***federal-gov***	5th-6th	never-married	***Adm-clerical***
3	***self-emp-inc***	***masters***	***married-civ-spouse***	Student
7	never-worked	***doctorate***	***divorced***	***Adm-clerical***
2	***federal-gov***	***masters***	***divorced***	***Armed-Forces***
8	***self-emp-inc***	***doctorate***	***married-civ-spouse***	***Armed-Forces***

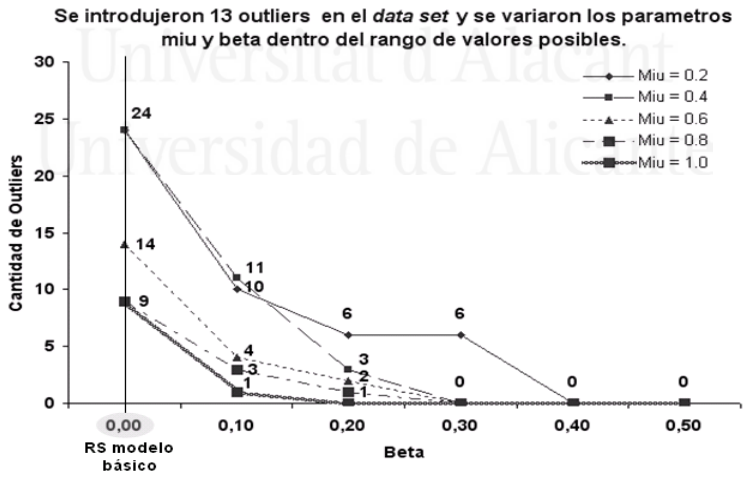


Figura 5-13 *RS* modelo básico *(RSBM) vs. VPRSM* en cuanto a detección de *outliers*

Interpretación de los resultados: el objetivo fundamental de esta prueba era mostrar la variación que experimenta la cantidad de *outliers* detectados a medida que se varía el valor de los umbrales β y μ. Permite además comparar lo que sucede en tal sentido cuando se trabaja con *RSBM* ($\beta=0$) y lo que sucede cuando se trabaja con *VPRSM* ($\beta\neq0$).

Tras interpretar los resultados se debe destacar que en todos los casos, dentro del conjunto de *outliers* detectados, siempre se encontraron algunos de los que fueron introducidos intencionalmente en el *conjunto de datos*:

- Cuando la cantidad de *outliers* detectados fue mayor que la cantidad de *outliers* introducidos, entonces dentro de los detectados estaban todos los introducidos.
- Cuando el número de *outliers* detectados fue menor que la cantidad de introducidos, entonces, de ellos, los que resultaron detectados fueron siempre los más contradictorios.

Los ejemplos siguientes ilustran lo expresado:

- $\mu = 0,2$; $\beta = 0$: Se detectaron 24 *outliers*, entre ellos, estaban los 13 introducidos.
- $\mu = 0,6$; $\beta = 0,2$: Se detectaron sólo 2 *outliers* que coinciden con dos de los 13 introducidos y especialmente los dos más contradictorios, pues lo eran por los cuatro atributos tomados en consideración.

La interpretación de las pruebas realizadas nos permite, además, llegar a las siguientes conclusiones:

- Una adecuada elección de las *relaciones de equivalencia* o *criterios de clasificación* garantiza, en gran medida, la efectividad en la detección.
- Para valores pequeños de los parámetros μ y β, en ocasiones, el número de *outliers* detectados es alto y se detectan como tales elementos que realmente no lo son.

Por ejemplo, para $\mu = 0,2$ y $\beta = 0$ se detectaron 24 *outliers*.

Esto reafirma un aspecto importante de la visión estadística del problema de la *detección de outliers* para la designación final de un caso como *excepcional*: cuando las observaciones candidatas a ser consideradas como tal han sido identificadas por algún método de detección entonces, posteriormente, el investigador debe hacer un análisis de estos resultados y seleccionar aquellas observaciones que muestran una contradicción real con respecto a la muestra estudiada.

- Al ir aumentando sucesivamente el valor del *umbral de excepcionalidad* (μ), se logra un refinamiento en la detección. Por lo general, una vez que el valor de dicho parámetro aumenta, disminuye el número de *outliers* detectados. Puede observarse que los que van quedando en cada caso, son los que son contradictorios respecto de una mayor cantidad de atributos. Sin embargo, en algunos casos y para ciertas variaciones del valor de μ, no se logra tal refinamiento.

  Por ejemplo:

  - El número de *outliers* detectados, al variar el valor de μ de *0,2* a *0,4* con $\beta = 0$, es 24, en ambos casos.
  - Cuando se varia μ de *0,8* a *1,0* con $\beta = 0$, en ambos casos también, la cantidad de *outliers* detectados fue la misma (9).

  Puede observarse que:

  - En los dos ejemplos, el valor de $\beta = 0$, lo cual implica que no se ha permitido ningún *grado de desclasificación*, por tanto, son resultados referidos a *RSBM.*

Universitat d'Alacant
Universidad de Alicante

- Al ir permitiendo un cierto grado de desclasificación (valores de $\beta \neq 0$) para las mismas variaciones de los valores de μ que las referidas en los ejemplos anteriores, la cantidad de *outlier*s detectados es diferente.
- Una vez que μ alcanza el mayor valor posible, $\mu=1$, el número de *outliers* detectados es 9, sin embargo, nuevamente, al hacer variaciones en el valor de β, se alcanza un mayor refinamiento en la detección, detectándose finalmente como *outliers* los elementos más contradictorios.

Lo anteriormente expresado demuestra que:

- Al permitirse un *grado de desclasificación* controlado e irse variando este progresivamente, se mejora la calidad de la detección.
- A pesar de lo anteriormente dicho, se debe ser cauteloso en la variación del valor de β. Permitir un alto *grado de desclasificación* puede implicar que se *limpien* completamente las *fronteras internas* y, por tanto, no se detecte ningún *outlier*.

  Por ejemplo:

  En las pruebas realizadas se pone de manifiesto que esto sucede a partir de $\beta = 0,3$.

Prueba 5-4

Objetivo de la prueba: validar *calidad de la detección* del algoritmo *VPRSM*

***Conjunto de datos* utilizado:** *Arrhythmia Data Set*

Descripción del *conjunto de datos*: datos de *pacientes* con problemas cardiovasculares

Fuente: *UCI Machine Learning Data Repository*

Universitat d'Alacant
Universidad de Alicante

Tipo del *conjunto de datos*: *multivariado*

Tipo de los atributos: *reales, enteros* y *categóricos.*

Cantidad de filas: 452

Cantidad de atributos (columnas): 279

Dispositivo de cálculo utilizado en la prueba: INTEL(R) Core(TM) 2 Duo, CPU T5450 @ 1.66 Ghz (2 CPUs), 2046 MB de RAM. Plataforma: Windows Vista

Sobre este *conjunto de datos se* hicieron dos pruebas diferentes, por tanto, la descripción de cada una de ellas se identificará de manera diferente también.

Descripción de la Prueba 5-4.1

A continuación se expresa el *concepto* y los criterios de clasificación que han sido considerados:

Concepto

personas con *peso* ≤40 *kg* (*weigth* ≤ 40)

Relaciones de Equivalencia

- *relación de equivalencia-1:*

  Se estableció a partir del atributo *heart rate*: cantidad promedio de *latidos por minuto* del corazón de las personas.

  A partir de los posibles valores este atributo, la relación de equivalencia particiona al *conjunto de datos* en dos clases de equivalencia:

  [44, 61] y [62, 163]

- *relación de equivalencia-2:*

Universitat d'Alacant
Universidad de Alicante

Se estableció a partir del atributo *number of intrinsic deflections:* cantidad de *desvíos arteriales* inherentes a cada persona.

A partir de los posibles valores este atributo, la relación de equivalencia particiona al *conjunto de datos* en dos clases de equivalencia:

[0, 59] y [60, 100]

- *relación de equivalencia-3:*

  Se estableció a partir del atributo *height: altura* de una persona expresada en centímetros.

  A partir de los posibles valores este atributo, la relación de equivalencia particiona al *conjunto de datos* en dos clases de equivalencia:

  [60, 175] y [176, 190]

Las *personas* que clasifican dentro del *concepto* se consideran de *bajo peso*. Teniendo en cuenta los valores habituales de las mismas para los atributos implicados en las relaciones de equivalencia descritas, se introdujeron intencionalmente en el *conjunto de datos*, 12 *outliers*: *personas de bajo peso con valores contradictorios para varios de dichos atributos*.

Universidad de Alicante

En tal sentido, vale señalar que los valores *normales* de los atributos tenidos en consideración en las *relaciones de equivalencia* para las *personas de bajo peso*, son los siguientes:

heart rate: más de 65 latidos promedio por minuto

intrinsic deflections: menos de 50 desvíos arteriales

height : menor que 170 cm

La **Tabla 5-3** refleja los *outliers* introducidos. En ella, los valores en **negrita** e *itálica* representan valores

contradictorios, mientras que los demás representan valores normales, para los atributos a los que están asociados.

En la tabla se destaca solo el valor de los atributos que son relevantes al análisis. Note que los elementos 5, 6, 7, 8 y 11 son los más contradictorios, por serlo para TODOS los atributos que intervienen en el análisis.

Tabla 5-3 *Outliers* introducidos Prueba 5-4.1: *Arrhythmia DS*

Id	*weight (Kg)*	*heart rate*	*number of intrinsic deflections*	*height (cm)*
1	15	***60***	17	***180***
2	31	93	***68***	***178***
3	39	***50***	***82***	130
4	10	***53***	16	***188***
5	19	***45***	***90***	***190***
6	20	***48***	***86***	***183***
7	25	***50***	***71***	***180***
8	29	***55***	***75***	***179***
9	33	90	***60***	***176***
10	40	***61***	20	***186***
11	26	***50***	***99***	***180***
12	38	92	***100***	***178***

Universitat d'Alacant
Universidad de Alicante

En la prueba, los valores de μ tenidos en consideración fueron los siguientes: 0,2; 0,4; 0,6; 0,8 y 1. Para cada uno de estos valores, se varió el valor de β a partir de la siguiente secuencia de valores: 0; 0,1; 0,2 y 0,3. Los valores 0,4 y 0,5 que pudieran haber sido también representativos no son mencionados pues a partir de β=0,3 la cantidad de *outliers* detectados se mantuvo en 0.

Interpretación de los Resultados de la Prueba 5-4.1

En esta prueba se pone de manifiesto, nuevamente, que al permitirse un cierto *grado de desclasificación* en el cálculo

de las regiones significativas se refina el proceso de detección.

Significativamente, para todos los valores de μ tenidos en consideración y el valor β=0 —correspondencia con *RSBM*— se detectó siempre la misma cantidad de *outliers*: 30. Solo al comenzar, en cada caso, la variación de los valores de β fue que entonces el número de *outliers* detectados comenzó a experimentar una variación. A partir de ella, se alcanzó un refinamiento en el proceso de detección. Este resultado es un ejemplo característico que evidencia la importancia de permitir un cierto *error de clasificación* en el proceso de detección y de cómo la variación del mismo influye en la calidad de dicho proceso.

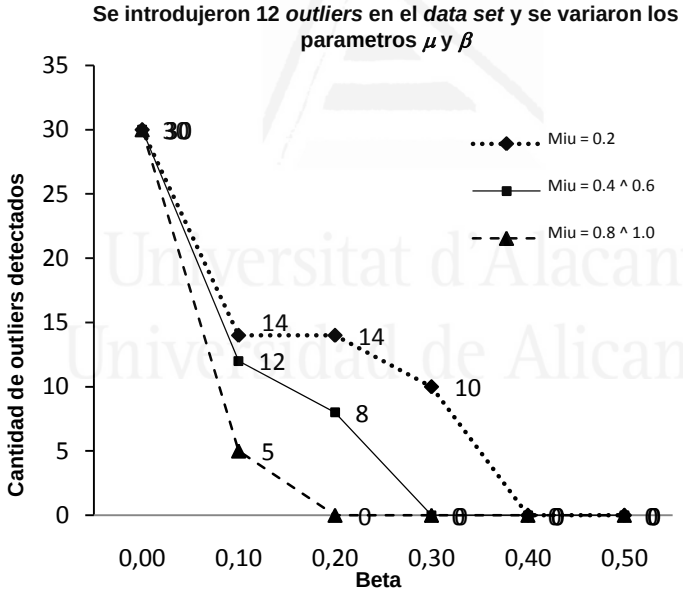


Figura 5-14 Detección de *outliers* Prueba 5-4.1*: Arrhythmia DS*

La **Tabla 5-4** refleja, para cada valor de μ, la variación que experimenta la cantidad de *outliers* detectados en función de la variación de los valores de β. La **Figura 5-14**, por su

parte, muestra una interpretación gráfica de los resultados alcanzados.

Tabla 5-4 Detección de *outliers* Prueba 5-4.1: *Arrhythmia DS*

β	μ	**cantidad de *outliers* detectados**	***Del total de detectados, cuántos eran outliers insertados***
0	0.2	30	12
0	0.4	30	12
0	0.6	30	12
0	0.8	30	12
0	1.0	30	12
0.1	0.2	14	12
0.1	0.4	12	12
0.1	0.6	12	12
0.1	0.8	5	5
0.1	1.0	5	5
0.2	0.2	14	12
0.2	0.4	8	8
0.2	0.6	8	8
0.2	0.8	0	0
0.2	1.0	0	0
0.3	0.2	10	9
0.3	0.4	0	0
0.3	0.6	0	0
0.3	0.8	0	0
0.3	1.0	0	0

Universitat d'Alacant
Universidad de Alicante

En todos los casos, los *outliers* introducidos quedaron dentro del conjunto de *outliers* detectados. Cuando el refinamiento alcanzó su máxima expresión, los *outliers* detectados fueron los más contradictorios introducidos. Por ejemplo, para μ = 0,8 con β = 0,1 los 5 *outliers* detectados fueron los más contradictorios que se introdujeron en el *conjunto de datos*.

Observación: Para los valores de μ = 0,4 y 0,6; así como para los valores 0,8 y 1 del mismo umbral, hubo coincidencia en los resultados alcanzados.

Descripción de la Prueba 5-4.2

Introducción

Teniendo en cuenta que los atributos a partir de los cuales se establecen el *concepto* y las *relaciones de equivalencia* están asociados a términos médicos dentro de la especialidad de *cardiología*, se hace necesario, previamente, brindar una breve explicación técnica de su significado para garantizar el entendimiento de la prueba. En especial, resulta necesario conocer aspectos esenciales referidos al funcionamiento de un *electrocardiograma*. Los elementos que se aportan fueron suministrados por especialistas en el área y están expresados de forma simple para garantizar la comprensión por parte de un lector no especializado.

Como es sabido, el *corazón* es el órgano encargado de *bombear* la sangre oxigenada al resto del cuerpo humano. Cuando el *corazón* realiza este proceso se *despolariza* —emite *energía eléctrica*— y esto se traduce en la emisión de una serie de *ondas electromagnéticas* que son las que se reflejan en un *electrocardiograma*. En tal sentido, se identifican 5 ondas principales: *P*, *Q*, *R*, *S* y *T*. Cada una de ellas está asociada a un determinado *proceso* realizado por el *corazón*.

A continuación se definen el *concepto* y las *relaciones de equivalencia* tenidas en cuenta en esta prueba. Con anterioridad a la definición de cada uno de estos aspectos, se describe el significado, desde el punto de vista médico, de los atributos que intervienen en la definición de los mismos.

- Atributo *QRS duration:*

  Se denomina *complejo QRS* a la emisión de las ondas *Q, R* y *S*, —una detrás de la otra— antes de que se contraiga el *ventrículo izquierdo* del *corazón* para comenzar a *bombear* la sangre ya oxigenada al resto del organismo. Este atributo mide el tiempo promedio que dura la secuencia de emisión de las ondas antes mencionadas. En una *persona normal*, oscila entre 60 y 80 milisegundos.

  En el *concepto* que se definirá resultaron de interés aquellas personas para las cuales el valor de dicho atributo era *normal*: 60≤*QRS duration* ≤80.

- Atributo *P-R interval*

  El *intervalo P-R* es el lapsus de tiempo que media entre la detección de la onda *P* y la detección de la onda *R* por un *electrocardiograma*. El tiempo promedio para *personas normales* está entre los 160 y 165 milisegundos.

  Al igual que en el parámetro anterior, en la definición del *concepto* resultaban de interés las *personas* con valores *normales* para este indicador, o sea: 160≤*P-R interval* ≤165.

Universitat d'Alacant
Universidad de Alicante

Concepto

El *concepto* se establece a partir de los *pacientes* de la muestra que cumplen:

[60≤*QRS duration* ≤80] *and* [160≤*P-R interval* ≤165]

Relaciones de equivalencia

- *Relación de equivalencia No. 1: a*tributo *Q-T interval*

El *intervalo Q-T* es el lapsus de tiempo que media entre la detección de la onda *Q* y la detección de la onda *T* por un *electrocardiograma*. Tal valor, para *personas normales*, es de 350 milisegundos.

A partir de los posibles valores de este atributo, se estableció una *relación de equivalencia* que particiona al *conjunto de datos* en tres *clases de equivalencia*:

[232, 332], [333, 433] y [434, 509]

- *Relación de equivalencia No. 2:* atributo *T interval*

  El *intervalo T* es el lapsus de tiempo que dura la onda *T*. Para *personas normales* es de 150 milisegundos.

  A partir de los posibles valores este atributo, se estableció una *relación de equivalencia* que particiona al *conjunto de datos* en tres *clases de equivalencia*:

  [108, 195], [196, 295] y [296, 381]

- *Relación de equivalencia No. 3:* atributo *vector angles in degrees on front plane of QRST*

  El *ángulo del vector en el plano frontal del conjunto de ondas QRST* es otro elemento que está presente en un *electrocardiograma*. La unidad de medida para este atributo es el *grado*.

  A partir de los posibles valores de este atributo, se estableció una *relación de equivalencia* que particiona al *conjunto de datos* en dos *clases de equivalencia*:

  [-135, 0] y [1, 166]

- *Relación de equivalencia No. 4:* atributo *number of intrinsic deflections*

  Este atributo representa la cantidad de *desvíos arteriales* inherentes a cada *persona*.

Universitat d'Alacant
Universidad de Alicante

A partir de los posibles valores de este atributo, se estableció una *relación de equivalencia* que particiona al *conjunto de datos* en tres *clases de equivalencia*:

[0, 32], [33, 66] y [67, 100]

Las personas que cumplen el *concepto* tienen las siguientes características:

- Por lo general tienen un *intervalo Q-T normal*, por lo que pertenecen a la *clase de equivalencia 2* según la *relación 1*.
- Su *intervalo T* tiende a lo *normal* por lo que, casi todas, están en la *clase de equivalencia* 1 según la *relación* 2.
- El *ángulo del vector de las ondas QRST*, casi siempre, es *positivo* por lo que están en la *clase de equivalencia 2* según la *relación* 3.
- Tienen muy pocas *desviaciones intrínsecas* por lo que, en su mayoría, están en la *clase de equivalencia* 1 según la *relación* 4.

Por tanto, se tratará que las características de los individuos que se introduzcan intencionalmente como *outliers* en el *conjunto de datos*, disten en gran medida de las que se acaban de expresar.

Universidad de Alicante

Características de la prueba

Se introdujeron 15 *outliers* en el *conjunto de datos*. La **Tabla 5-5**, refleja los *outliers* introducidos y los valores en ***negrita*** e *itálica* representan valores contradictores para los atributos a los que están asociados. El resto, son valores *normales*. En dicha tabla, los primeros 5 elementos de la misma serán contradictorios por las 4 *relaciones de equivalencia*. Los elementos del 6 al 10, serán contradictorios por 3 *relaciones* y los restantes 5 lo serán solo por 2 de ellas.

En la tabla aparecen reflejados solo los atributos que son relevantes al análisis.

Tabla 5-5 *Outliers* introducidos Prueba 5-4.2: *Arrhythmia D*

Id	*QRS duration (msec)*	*P-R interval (msec)*	*Q-T interval (msec)*	*T interval (msec)*	*vector angle of QRST (degree)*	*number of intrinsic deflections*
1	79	165	***240***	***300***	***-50***	***95***
2	68	160	***271***	***297***	***-10***	***77***
3	66	161	***280***	***299***	***-23***	***82***
4	79	164	***235***	***360***	***-34***	***90***
5	80	162	***330***	***320***	***-5***	***91***
6	77	161	***240***	***310***	5	***37***
7	80	163	***286***	***298***	22	***60***
8	65	164	***435***	***296***	***-80***	25
9	68	160	***455***	135	***-49***	***85***
10	74	162	338	***200***	***-46***	***91***
11	61	160	390	***201***	***-100***	28
12	65	163	400	***240***	***-90***	15
13	68	164	***450***	***280***	50	10
14	70	165	***480***	***291***	33	30
15	75	161	***500***	***233***	88	12

Universitat d'Alacant
Universidad de Alicante

Interpretación de la Prueba 5-4.2

La **Tabla 5-6** y la **Figura 5-15** reflejan, de distinta forma, los resultados alcanzados en la prueba.

Por su parte, la **Tabla 5-6** refleja, para cada par de valores μ y β, la cantidad de *outlier*s detectados por el algoritmo y de ellos, cuántos formaban parte del conjunto de *outlier*s introducido en el *conjunto de datos*.

Mientras tanto la **Figura 5-15** refleja la cantidad de *outliers* detectados en función de la variación de los valores de μ y β.

En esta prueba como aspecto peculiar puede comentarse lo siguiente: Cuando se trabajó con un *umbral de excepcionalidad* μ= 0,2 y se varió el valor β de 0,1 a 0,2; el número de *outliers* aumentó (de 10 a 13) en vez de disminuir que es lo que debe ser la tendencia.

Tabla 5-6 Detección de *outliers* Prueba 5-4.2: *Arrhythmia DS*

β	μ	**cantidad de *outliers* detectados**	**Del total de detectados, cuántos eran *outliers* insertados**
0	0.2	30	15
0	0.4	30	15
0	0.6	30	15
0	0.8	30	15
0	1.0	30	15
0.1	0.2	10	10
0.1	0.4	10	10
0.1	0.6	7	7
0.1	0.8	5	5
0.1	1.0	5	5
0.2	0.2	13	13
0.2	0.4	7	7
0.2	0.6	0	0
0.2	0.8	0	0
0.2	1.0	0	0
0.3	0.2	10	10
0.3	0.4	5	5

Universitat d'Alacant
Universidad de Alicante

Este hecho se debe a que —para β=0,1— al formarse las *fronteras internas* por cada *relación de equivalencia*, sucedió lo siguiente:

La *frontera interna* por una de las relaciones (sea *A*) tenía como subconjunto propio a la *frontera interna* de otra relación (sea *B*). Por tanto, según lo establecido en la

concepción teórica del método, esto conlleva a que el análisis de *outliers* se deja para cuando se esté analizando la frontera interna *B*, pues no habrá ningún elemento en *A* que pertenezca a algún *conjunto excepcional no redundante* y que no pertenezca a *B*.

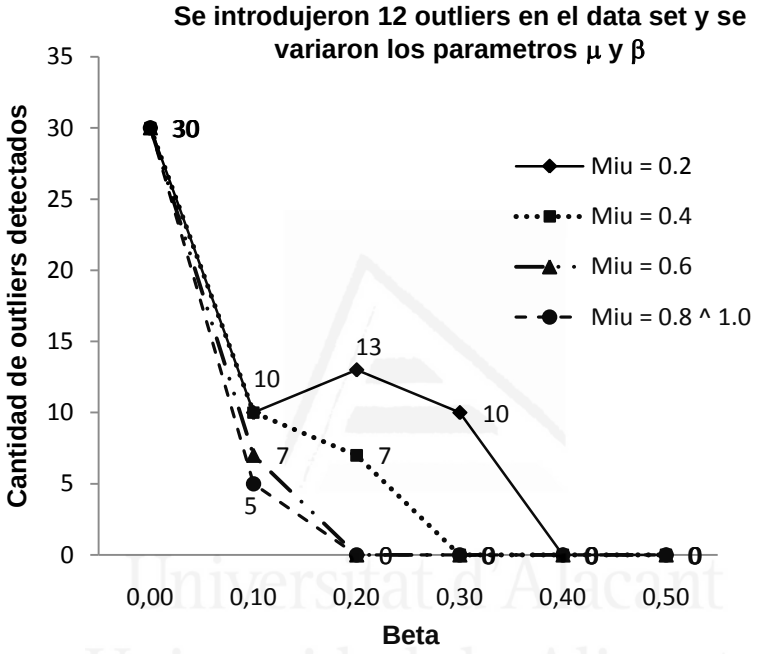


Figura 5-15 Detección de *outliers* Prueba 5-4.2*: Arrhythmia DS*

Al permitirse un *error de clasificación β*= 0,2; todos los elementos que estaban en *frontera interna B* pasaron a la *región positiva*. Por tanto, se excluyen del análisis de *outliers*. Al dejar de existir la *relación de inclusión* que había entre *A* y *B*, entonces, esto provoca que un mayor número de elementos de *A* entre en consideración como posibles *outliers* en el momento en que se esté analizando dicha *frontera interna*.

A pesar de la situación antes descrita, en el resto de los resultados que arrojó la prueba se puso de manifiesto que

al permitir un cierto *grado de desclasificación* en el cálculo de las regiones significativas —permitir mayor nivel de tolerancia en el mismo— se refina el proceso de detección.

Se puede observar cómo, en un primer momento, se detectan 30 *outliers* para β= 0. Esto sucede para todos los valores de μ. Sin embargo, permitiendo un pequeño *grado de desclasificación* (β = 0,1) esto provoca que para los distintos valores de μ, tenidos en consideración, disminuya el número de *outliers* detectados. Por ejemplo, para un umbral μ = 0,2 se reduce el número de *outliers* detectados a 10. Estos 10, coincide que fueron de los elementos insertados intencionalmente en el *conjunto de datos*.

Para μ = 0,8 y β = 0,1 se detectan solamente 5 *outliers*, que son, precisamente, los 5 más contradictorios insertados en el *conjunto de datos*.

Vale recordar que, a medida que se aumenta el valor μ, se hace más exigente la detección. Los elementos que pertenecen a un mayor número de *fronteras internas* son los que van quedando como *outliers*, pues en tal caso, la razón (*número de fronteras internas a las que pertenece el elemento*) / (*total de fronteras internas*) va tendiendo a 1.

Universidad de Alicante

Conclusiones

Como conclusión general puede señalarse que el marco teórico alcanzado permitió establecer un método de detección de *outliers* basado en *VPRSM* a partir del cual se elimina el carácter determinista —en cuanto a la *clasificación*— del método que le antecede basado en *RSBM*.

El algoritmo *VPRSM*, a partir del cual se valida la viabilidad computacional del método propuesto; además de lograr mejores resultados en la detección de *outliers* mantiene el

mismo orden de *complejidad temporal* y *espacial* del algoritmo *RSBM*.

Al permitirse un *grado controlado de desclasificación* se refina la detección de *outliers* a través de una *limpieza* paulatina de las *fronteras internas* y con ello se logra mayor precisión en la detección; haciendo que, finalmente, queden como *outliers* los elementos más contradictorios.

Las pruebas de validación realizadas al algoritmo corroboraron, además, que el mismo puede aplicarse de forma eficiente sobre *conjuntos de datos* de *gran tamaño* y *alta dimensionalidad*. En este sentido, puede señalarse que estos aspectos constituyen ventajas (al igual que del algoritmo *RSBM*) con respecto a propuestas anteriores de métodos de detección y, por tanto, pueden considerarse también aportes de la investigación: brindar soluciones computacionales eficientes, tales como la posibilidad de utilizar algoritmos *casi*-lineales, es una ventaja que cualquier analista/ingeniero de datos valorará sin duda en su justa medida, dada la elevada *complejidad* habitual de los procedimientos en el campo del *KDD–DM*.

Universitat d'Alacant

Sin embargo, pese a todo, este método sigue conservando el estilo tradicional de trabajo de la mayoría de los métodos de detección de *outliers* existentes. En él, se mantiene la necesidad de establecer las condiciones particulares —con respecto a la elección de los valores de los umbrales que intervienen en el análisis— bajo las cuales se determina un conjunto de *outliers* dentro de un *universo* de datos dado.

Atendiendo a lo anteriormente expresado, se hace necesario que el usuario realice un análisis previo del contexto a partir del cual pueda establecer el valor de dichos umbrales. Una vez que para esos valores específicos se determina un conjunto de *outliers*, es necesario un análisis posterior de dicho conjunto. Como resultado de éste, podría

ser necesario el establecimiento de nuevos valores de los umbrales bajo los cuales se puede obtener un nuevo conjunto de *outliers* que se adecúe, realmente, a los intereses particulares del análisis. Este proceso supone ejecutar el algoritmo tantas veces como pares de umbrales (μ, β) diferentes hayan sido seleccionados.

La situación descrita en el párrafo anterior, lleva a la siguiente reflexión:

Aunque esté determinada la *complejidad temporal* del método de detección, no se puede decir que la misma se corresponde con el tiempo real empleado en el proceso global de *análisis de casos excepcionales*. Para el cálculo de éste, habría que tener en cuenta el costo en tiempo que conlleva el análisis previo y posterior del contexto, ante cada ejecución del algoritmo, así como el coste total que implique la ejecución del algoritmo, tantas veces como sea necesario. En consecuencia, el coste global del proceso puede llegar a aumentar considerablemente.

No obstante, el marco teórico alcanzado hasta el momento sirve de antecedente para el establecimiento de nuevos resultados a partir de los cuales se garantice el cumplimiento del **objetivo general** propuesto en la investigación, todo lo cual se aborda en el siguiente capítulo.

Universitat de Lleida
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Capítulo 6

Algoritmo *Beta-Miu* Probabilístico

Hasta el momento se han propuesto dos métodos de detección de *outliers* basados en *RSBM* y en *VPRSM,* respectivamente, que dan solución al siguiente problema:

> *«A partir de un umbral de excepcionalidad establecido (μ) y un determinado error de clasificación permitido (β), extraer un conjunto de outliers a partir de un universo de datos dado.»*

Universitat d'Alacant
Universidad de Alicante

En este capítulo, basándonos en los resultados anteriores —que permitieron establecer y validar el marco teórico— se propone un nuevo enfoque del problema de la detección de *outliers,* que pretende resolver las limitaciones de los métodos antes mencionados. El nuevo enfoque estará en correspondencia con el **objetivo general** planteado en la investigación:

> *«Establecer un método, computacionalmente viable, que proporcione la probabilidad que tiene cada*

elemento de un universo de datos dado de ser excepcional, sin necesidad de haber establecido las condiciones previas —referidas a la determinación de los umbrales que intervienen en el análisis— en función de un contexto específico de aplicación.»

Dicho objetivo se estableció sobre la base de la siguiente **hipótesis**:

«Es posible desarrollar una nueva teoría basada en la extensión de los conceptos básicos y las herramientas formales que nos proporciona la Teoría de Conjuntos Aproximados (Pawlak, 1982), (Pawlak, 1991) y el Modelo de Conjuntos Aproximados de Precisión Variable (VPRSM) (Ziarko, 1993), aplicados al problema de la detección de outliers, que permita obtener, de forma no supervisada, para cada elemento de un universo de datos, la región de valores de los umbrales en la cual dicho elemento es outlier. A partir de dicho resultado, es posible determinar la probabilidad de cada elemento del universo de ser outlier con relación al mismo.»

Universitat d'Alacant
Universidad de Alicante

Para desarrollar el método expresado en el objetivo general expuesto —que constituye otra aportación de la investigación— se propone una ampliación del marco teórico desarrollado en los **Capítulos 4** y **5** a partir de los elementos conceptuales de la Teoría de *RS* así como de *VPRSM*, junto a la propuesta teórica de (Jiang *et al.*, 2005). Todos ellos permiten demostrar, formalmente, los elementos teóricos propuestos en la nueva concepción del método y sirven de marco de referencia para el diseño e implementación de un algoritmo, computacionalmente viable, con el que se valida la hipótesis de partida.

El método propuesto se fundamenta en dos fases bien diferenciadas:

- En la primera, se determina, para cada elemento del *universo* finito *U,* bajo qué condiciones (*umbral de excepcionalidad* μ y *error de clasificación* permitido β) dicho elemento se comporta como un elemento *excepcional* (*outlier*). Dichas condiciones (μ y β) establecen una región *R* dentro de la cual el elemento se considera *outlier.*
- En la segunda fase, teniendo en cuenta la región *R* determinada para cada elemento del *universo* finito *U*, se calcula la probabilidad de cada uno de ellos de ser *outlier* en *U*.

En función de lo expuesto, el resto del capítulo se estructura de la siguiente forma: seguidamente se abordan los aspectos referidos a la *primera fase*. Se proponen y se demuestran formalmente un conjunto de resultados teórico-matemáticos que permiten determinar *R* para cada elemento del universo en función de los umbrales μ y β. Desde el punto de vista del método científico, estos elementos teóricos se concretan en un algoritmo cuya funcionalidad permite calcular la región antes mencionada. De dicho algoritmo, además, se ofrecen detalles sobre su implementación computacional. Posteriormente, se abordan los aspectos referidos a la *segunda fase*. Teniendo en cuenta el resultado alcanzado en la *primera*, en esta fase se proponen nuevos elementos teóricos y se aplican técnicas estadísticas que permiten determinar la probabilidad de cada elemento del *universo* de ser *outlier* dentro del mismo. La viabilidad computacional del método queda validada mediante la propuesta de un algoritmo del cual se dan detalles sobre su implementación

computacional. Finalmente, se dan las conclusiones del capítulo.

Fase 1. Determinación para cada Elemento del *Universo* de su Región de Excepcionalidad

Formalmente, el problema a resolver en esta fase puede enunciarse de la siguiente forma:

$\forall x: x \in U$, determinar R. Donde R es la región que establece el conjunto de valores de los umbrales β y μ para los cuales x es *outlier* en U.

Para solucionar el mismo, se sigue la misma metodología empleada en capítulos anteriores. Se amplía el marco teórico existente, con nuevas definiciones y proposiciones, lo cual permite establecer un algoritmo, computacionalmente eficiente, que resuelve el problema.

Universitat d'Alacant
Universidad de Alicante

Marco Teórico. Algoritmo *BM*

Dividiremos el análisis en dos partes:

1. Determinación de la *región de excepcionalidad* con respecto al umbral β (*grado de desclasificación*).
2. Determinación de la *región de excepcionalidad* para el umbral μ (*umbral de excepcionalidad*).

Finalmente, se integran estas dos soluciones particulares para determinar la *región de excepcionalidad* (β, μ) para cada elemento del *universo*.

Región de Excepcionalidad con respecto a β

En este apartado se determina la *región de excepcionalidad* con respecto al conjunto de valores de β (referido a un β-*error admisible en la clasificación*).

Para una mayor claridad en la exposición, el análisis de esta problemática se subdivide en función de tres subproblemas particulares a resolver:

- **Subproblema No. 1**

  Determinar el rango de valores de β para los cuales $B_i \subseteq B_j$, $i \neq j$, $1 \leq i,j \leq m$.

- **Subproblema No. 2**

  Determinar el rango de valores de β para los cuales una *frontera interna* dada es nula.

- **Subproblema No. 3**

  Determinar el conjunto de valores de β para los cuales $B_i = B_j$, $i \neq j$, $1 \leq i,j \leq m$.

Finalmente, las tres soluciones particulares se integran para poder establecer la *región general de excepcionalidad* de un elemento cualquiera del *universo* con respecto al conjunto de valores de β.

Subproblema No. 1: Determinación del rango de valores de β para los cuales $B_i \subseteq B_j$. $i \neq j$, $1 \leq i,j \leq m$

Supongamos que se tienen $U/r_i = \{P_{1,\ldots,}P_{n1}\}$ y $U/r_j = \{Q_{1,}\ldots,Q_{n2}\}$, dos particiones inducidas en U por las *relaciones de equivalencia* r_i y r_j, respectivamente, $1 \leq i,j \leq m$.

De acuerdo con los Lemas y Proposiciones enunciados en el **Capítulo 4,** sabemos que si ninguna *frontera interna* B_i es subconjunto de otra *frontera interna* B_j, entonces todos los elementos de B_j son candidatos a ser *outliers* en U.

A partir de esto, el nuevo problema a resolver podría replantearse de la siguiente forma:

Determinar el conjunto de valores de β para los cuales una *frontera interna* B_i, $i \neq j$, es subconjunto de la *frontera interna* B_j, o sea, $B_i \subseteq B_j$. Una vez calculados éstos, $\forall i \neq j,\ 1 \leq i \leq m$, entonces el complemento de la unión de todos los intervalos de valores de β calculados, será el conjunto de valores, con respecto a dicho parámetro, para los cuales TODOS los elementos de B_j son candidatos a ser *outliers*.

Consecuentemente, supongamos entonces:

$$B_i \subseteq B_j$$

Sabemos además que:

$$P_1 \cup \ldots \cup P_{n1} = U$$

$$P_{i1} \cap P_{i2} = \phi,\ 1 \leq i_1 \neq i_2 \leq |U/r_i|$$

por ser *clases de equivalencia* definidas por r_i sobre U.

Retomemos la función (9) definida en *VPRSM* que establece la *medida del grado de inclusión relativo* del conjunto X con respecto al conjunto Y (o lo que es lo mismo, el margen de error supuesto al considerar $X \subseteq Y$). Sabemos que dicha función queda definida de la siguiente forma:

$$c(X,Y) = \begin{cases} 1 - |X \cap Y|/|X| & si\ |X| \neq 0 \\ 0 & si\ |X| = 0 \end{cases}$$

y además sabemos que a partir de ella pueden hacerse las siguientes consideraciones:

- si $c(X, Y) = 0 \Rightarrow X \subseteq Y \Rightarrow |X \cap Y| = |X|$

  $\Rightarrow$ NO hay *error en la clasificación*

- si $c(X, Y) \approx 1 \Rightarrow X, Y$ se acercan a ser disjuntos

- si $c(X, Y) = 1 \Rightarrow |X \cap Y| = 0 \Rightarrow X, Y$ son disjuntos

La expresión numérica *c(X, Y)* es un indicativo del *error relativo de clasificación*.

En el **Capítulo 5** vimos que si se toma como base la *medida de desclasificación relativa*, se puede definir la *relación de inclusión*, de la siguiente forma:

$$X \subseteq Y \Leftrightarrow c(X, Y) = 0$$

De la **Definición 17** —relación de inclusión mayoritaria— se conoce, además que:

- si $c(X, Y) \leq \beta \Rightarrow X \overset{\beta}{\subseteq} Y$, o sea, $c(X, Y) \in [0; \beta]$

- si $c(X, Y) > \beta \Rightarrow X \overset{\beta}{\not\subset} Y$, o sea, $c(X, Y) \in (\beta; 0{,}5)$

Basado en lo anterior se define la siguiente función:

Definición 21:

Dados los conjuntos *A* y *B* ($A, B \neq \phi$), $\beta \in [0; 0{,}5)$, se define la función $\varphi(A, B)$ de la siguiente forma:

$$\varphi(A,B) = \begin{cases} \phi \; si (c(A,B) \leq \beta) ó (c(A,B) \geq 1-\beta) \\ B \qquad\qquad en\_otro\_caso \end{cases}$$

Universitat d'Alacant
Universidad de Alicante

En la **Figura 6-1** se observa una interpretación gráfica del significado de esta función.

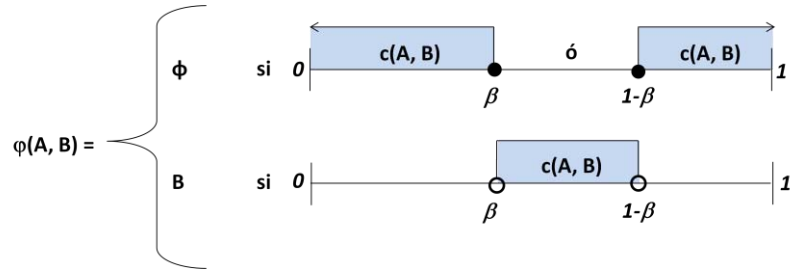


Figura 6-1 Interpretación gráfica de la función $\varphi(A, B)$

Por otra parte, desde el punto de vista teórico la interpretación de la función φ, sería la siguiente:

Supongamos que $X \subseteq U$ es el conjunto de todos los elementos del *universo* que cumplen el *concepto* y $Ce \subseteq U$, $Ce \neq \phi$, es una *clase de equivalencia* cualquiera que pertenece a la partición de U establecida por una determinada *relación de equivalencia* definida sobre el propio U. Si se evalúa $\varphi(Ce, X)$, entonces el resultado de esta evaluación puede interpretarse de la siguiente forma:

- si $\varphi(Ce, X) = \phi \Rightarrow$ la *mayoría* de los elementos de Ce cumple el *concepto* y por tanto $Ce \in X_\beta^+$ o la *mayoría* de los elementos de Ce no lo cumple y por tanto $Ce \in X_\beta^-$.
- si $\varphi(Ce, X) = X \Rightarrow$ en Ce hay un cierto número de elementos que cumple el *concepto* y otro cierto número de elementos que no lo cumplen, pero ni uno ni otro subconjunto puede decirse que representan una *mayoría* a partir del *error de clasificación* permitido, o sea, $Ce \in BN_\beta(X)$.

La **Figura 6-2** ilustra gráficamente la interpretación de la función φ dada.

Universitat d'Alacant
Universidad de Alicante

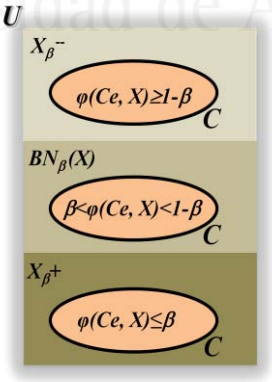


Figura 6-2 Interpretación de la función $\varphi(C, X)$ desde el punto de vista teórico. $X \subseteq U$ es el conjunto de todos los elementos del *universo* que cumplen el *concepto* y $C \subseteq U$, $C \neq \phi$

Definición 22:

Se definen los conjuntos C_l, $1 \leq l \leq |U/r_i|$, de la siguiente forma:

$C_l = P_l \cap \varphi(P_l, X)$

Esta definición puede interpretarse de la siguiente manera:

- $\varphi(P_l, X) = \phi \Rightarrow$

  La *mayoría* de los elementos de la *clase de equivalencia* P_l cumple el *concepto,* o sea $P_l \in X_\beta^+$, o la *mayoría* de los elementos de la *clase de equivalencia* P_l no cumple el *concepto,* o sea $P_l \in X_\beta^-$. (10)

- $\varphi(P_l, X) = X \Rightarrow$

  La *clase de equivalencia* $P_l \subseteq BN_\beta(X)$ (11)

entonces:

- si (10), $C_l = \phi$
- si (11), $C_l = P_l \cap X$, por tanto,

  $C_l = \{x: x \in P_l \wedge x \in X\}$ = {conjunto de los elementos de la *clase de equivalencia* P_l que cumplen el *concepto*}

y por tanto, en caso de cumplirse (10) ó (11), $C_l \subseteq B_i$ y esto se cumplirá $\forall l$, $1 \leq l \leq |U/r_i|$, o sea: Cada C_l será siempre un subconjunto de elementos de la *frontera interna* B_i .

Sin pérdida de generalidad, ilustremos lo que se acaba de expresar tomando como ejemplo un conjunto particular, C_1:

- *caso 1*: $\varphi(P_1, X) = \phi$

  $P_1 \cap \phi = \phi \Rightarrow \phi \subseteq B_i$

- *caso 2*: $\varphi(P_1, X) = X$

Universitat d'Alacant
Universidad de Alicante

$P_1 \cap X = \{x \in U: x \in P_1 \wedge x \in X\}$, por tanto, todos los elementos que pertenecen a este conjunto pertenecen también a B_i y por tanto, $C_1 = (P_1 \cap X) \subseteq B_i$.

Para el resto de los conjuntos C_l, puede hacerse un análisis similar.

Se tiene además que:

$P_i \cap P_j = \phi$

$C_i \subseteq P_i$

$C_j \subseteq P_j$

A partir de lo cual se concluye que $C_i \cap C_j = \phi$, $i \neq j$, $1 \leq i,j \leq |U/r_i|$. Esto permite establecer una nueva definición para la *frontera interna* B_i de X respecto a r_i:

Definición 23:

$$B_i = \bigcup_{l=1}^{|U/ri|} C_l$$

La validez de esta definición se justifica por el hecho de que cada C_l contiene a todos los elementos de la *clase de equivalencia* P_l que cumplen el *concepto*.

A su vez, los conjuntos C_i, $1 \leq i \leq |U/r_i|$ forman una partición de la *frontera interna* B_i, pues la intersección de todos ellos, dos a dos, es siempre el conjunto vacío.

Se puede asegurar además que,

Proposición 24:

$B_i \subseteq B_j \Leftrightarrow \forall C_l$, $1 \leq l \leq |U/r_i|$, $C_l \subseteq B_j$, $i \neq j$, $1 \leq i, j \leq m$.

Demostración:

$\Rightarrow$

Si $B_i \subseteq B_j \Rightarrow \forall C_l$, $1 \leq l \leq |U/r_i|$, $C_l \subseteq B_i \subseteq B_j$

Supongamos que existe k, $1 \leq k \leq n_1$ tal que $C_k \not\subset B_j$ (*lógicamente, bajo esta suposición* $C_k \neq \phi$, *pues, de lo contrario, con* $C_k=\phi$, *se cumpliría* $\phi \subseteq B_j$) y esto implica que existe $x \in C_k$ tal que $x \notin B_j$, pero como $C_k \subseteq B_i$ entonces $x \in B_i$ lo que implica que $B_i \not\subset B_j$ (contradicción con la hipótesis).

$\Leftarrow$

$\forall C_l$, $1 \leq l \leq |U/r_i|$, $C_l \subseteq B_j \Rightarrow B_i \subseteq B_J$:

como $B_i = C_1 \cup \ldots \cup C_{|U/ri|}$ y cada $C_l \subseteq B_j$, entonces, $B_i \subseteq B_j$.

Luego, el problema de hallar los valores de β para los cuales la *frontera interna* $B_i \subseteq B_j$ se puede solucionar determinando los valores de dicho parámetro para los cuales $C_l \subseteq B_j$, $\forall C_l$, $1 \leq l \leq |U/r_i|$.

Finalmente, interceptando TODOS los conjuntos de valores de β particulares que se obtengan para cada uno de los C_l, $1 \leq l \leq |U/r_i|$, se determina el conjunto de valores de β (o sea los valores de dicho parámetro que son comunes a todos los C_l) para los cuales se cumple $B_i \subseteq B_j$.

Por tanto, en adelante nos centraremos en resolver el siguiente problema:

Dado un l, determinar los valores de β para los cuales se cumple $C_l \subseteq B_j$ con B_i, $B_j \neq \phi$ y $B_i \neq B_j$.

Sin pérdida de generalidad, hagamos nuevamente el análisis para un conjunto particular, sea C_1 y posteriormente el razonamiento sería el mismo para el resto de los conjuntos C_l. En correspondencia:

Determinemos el intervalo de valores de β para los cuales se cumple $C_1 \subseteq B_j$.

Teniendo en cuenta que $C_1 = P_1 \cap \varphi(P_1, X)$, analicemos cuál será el resultado de dicha intersección teniendo en cuenta los posibles valores que puede tomar $\varphi(P_1, X)$.

Universitat d'Alacant
Universidad de Alicante

Recordemos que según la **Definición 21**, $\varphi(P_1, X) = \phi$ ó $\varphi(P_1, X) = X$.

En cada caso, se obtendrá un valor particular para C_1 y a partir de este valor, determinaremos el rango de valores de β para los que se satisface $C_1 \subseteq B_j$.

- *caso 1*: $\varphi(P_1, X) = \phi$

  Este valor de la función se establece cuando $\forall\beta$ se cumple una de las dos opciones que se muestran en la **Figura 6-3**:

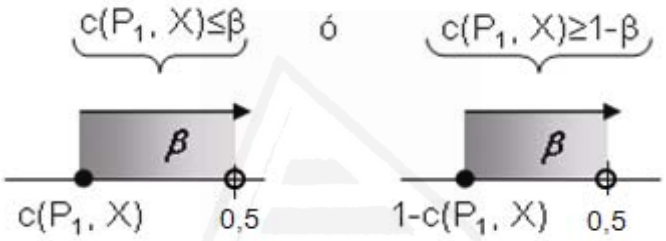


Figura 6-3 Posibles opciones para el Caso 1: $\varphi(P_1, X) = \phi$

Nótese que la propia definición de la función impone dos condiciones para β:

$[\beta \geq c(P_1, X)] \vee [\beta \geq 1 - c(P_1, X)]$,

por tanto, para garantizar que se cumpla alguna de ellas, se debe satisfacer que:

$\beta \geq mín(c(P_1, X), 1 - c(P_1, X))$

Al calcular C_1 según su definición, tendremos que:

$C_1 = P_1 \cap \phi = \phi$

Sabemos que $\phi \subseteq B_j$ y por tanto, $C_1 \subseteq B_j$.

Luego, en este *caso*, el intervalo de valores de β para el que se cumple $C_1 \subseteq B_j$ será el que se ilustra en la **Figura 6-4**.

Universitat d'Alacant
Universidad de Alicante

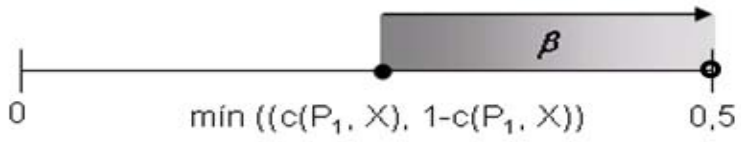


Figura 6-4 Intervalo de valores de β para el cual se cumple $C_1 \subseteq B_j$ cuando estamos en el Caso 1: $\varphi(P_1, X) = \phi$

Podemos entonces concluir el análisis para este *caso* afirmando que $\forall\beta: \beta \geq min\ (c(P_1, X), 1- c(P_1, X))$ se cumplirá $C_1=\phi$ y por tanto, $C_1 \subseteq B_j$

- *caso 2:* $\varphi(P_1, X) = X$

  Cuando estamos en este caso, entonces sucede lo contrario a lo que sucede en el *caso 1*, o sea, $\beta < c(P_1, X) < 1-\beta$.

  Al igual que en el *caso 1*, la definición de la función impone dos condiciones para β:

  $[\beta < c(P_1, X)] \wedge [\beta < 1 - c(P_1, X)]$,

  por lo cual, para garantizar que se cumplan ambas, se debe satisfacer que:

  $\beta < mín\ (c(P_1, X), 1 - c(P_1, X))$.

  Esto establece una primera restricción para los posibles valores de β cuando estamos en el *caso 2*. Dicha restricción se ilustra gráficamente en la **Figura 6-5**.

Universidad de Alicante

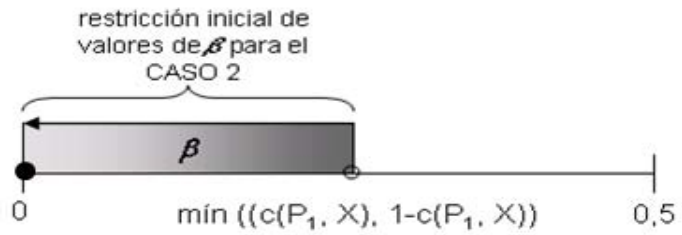


Figura 6-5 Restricción inicial de valores de β para el *caso 2*: $\varphi(P_1, X) = X$

Analicemos a continuación las siguientes situaciones posibles:

- Si $P_1 = \phi \Rightarrow$

  $c(P_1, X) = 0$ y debido a las restricciones iniciales para los valores de β que se ilustran en la **Figura 6-5**, no existen valores de β que lo satisfagan (o sea, $\beta \geq 0$ *y* $\beta < 0$ y es obvio que este sistema de inecuaciones no tiene solución).

- Si $P_1 \neq \phi \wedge P_1 \cap X = \phi \Rightarrow$

  $c(P1, X) = 1$ y debido a las restricciones iniciales para los valores de β que se ilustran en la **Figura 6-5**, no existen valores de β que lo satisfagan *(* como $mín(0; 1) = 0$, entonces las restricciones para β serían*:* $\beta \geq 0$ *y* $\beta < 0$ y es obvio que este sistema de inecuaciones no tiene solución*).*

- Si $P_1 \neq \phi \wedge P_1 \cap X \neq \phi \Rightarrow C_1 = P_1 \cap X \neq \phi \Rightarrow$

  Los elementos que pertenecen a C_1, son los que pertenecen a P_1 que cumplen el *concepto* (o sea, pertenecen a X) y en tal caso, debemos determinar las restricciones que se establecen para los valores de β. Estas restricciones se derivan de la necesidad de que

todos los elementos de C_1 pertenezcan a la *frontera interna* B_j (recordar que estamos intentando determinar valores de β para los cuales $C_1 \subseteq B_j$).

Como artificio, intentemos relacionar a los elementos de C_1 con los elementos de $U/r_j = \{Q_1,\ldots,Q_{n2}\}$, o sea, los elementos de la partición inducida en U por la *relación de equivalencia* r_j:

Como todos los elementos de C_1, son elementos de U, podemos asegurar que cada uno de ellos pertenecerá a alguna *clase de equivalencia* de la partición $Q_1, \ldots, Q_{n2}$ inducida por r_j en U.

Consideremos entonces los conjuntos C'_k , $1 \leq k \leq n_2$, asociados a la partición $Q = Q_1, \ldots, Q_{n2}$, definidos de forma similar a como se definieron los C_l, $1 \leq l \leq n_1$, para la partición $P = P_1, \ldots, P_{n1}$.

Definición 25:

Se definen los conjuntos C'_k, $1 \leq k \leq |U/r_j|$, de la siguiente forma:

$C'_k = Q_k \cap \varphi(Q_k, X)$, o sea,

$C'_1 = Q_1 \cap \varphi(Q_1, X)$, $C'_2 = Q_2 \cap \varphi(Q_2, X)$, $\ldots$, $C'_{n2} = Q_{n2} \cap \varphi(Q_{n2}, X)$

Según la **Definición 23**,

$$B_j = C'_1 \cup \ldots \cup C'_{n2} \qquad (12)$$

Para que se cumpla $C_1 \subseteq B_j$, siendo $C_1 \neq \phi$, se tendrá que cumplir que:

$\forall e: (e \in C_1 \Rightarrow e \in B_j \Rightarrow \exists k: e \in C'_k , 1 \leq k \leq n_2)$ - por (12)

Partiendo de este resultado se puede realizar la siguiente secuencia de deducciones:

Universitat d'Alacant
Universidad de Alicante

$e \in C'_k \Rightarrow e \in Q_k \Rightarrow Q_k \neq \phi \Rightarrow \varphi(Q_k, X) \neq \phi \Rightarrow \beta < c(Q_k, X) < 1-\beta$

y la restricción $\beta < c(Q_k, X) < 1-\beta$, se impondrá para cada una de las *clases de equivalencia* $Q_1, \ldots, Q_{n2}$ que contengan al menos un elemento de C_1.

Si denotamos $C'[e] = C'_k : e \in C'_k$, y suponemos que:

$C_1 = \{e_1, e_2, \ldots, e_t\}$,

entonces se tienen que cumplir todas las desigualdades siguientes:

$\beta < c(C'[e_1], X) < 1-\beta \Rightarrow \beta < min(c(C'[e_1], X), 1 - c(C'[e_1], X))$

$\beta < c(C'[e_2], X) < 1-\beta \Rightarrow \beta < min(c(C'[e_2], X), 1 - c(C'[e_2], X)$

...

$\beta < c(C'[e_t], X) < 1-\beta \Rightarrow \beta < min(c(C'[e_t], X), 1 - c(C'[e_t], X)$

Por tanto:

$\beta < min(min(c(C'[e_1], X), 1 - c(C'[e_1], X)),$

$min(c(C'[e_2], X), 1 - c(C'[e_2], X)),$

$\ldots,$

$min(c(C'[e_t], X), 1 - c(C'[e_t], X)))$

Generalizando, podemos plantear:

$\beta < min_{e \in C1} (min (c(C'[e], X), 1 - c(C'[e], X)))$

Finalmente, a estas restricciones se añaden las restricciones iniciales que se establecieron para el *caso 2* y a partir de ello, puede decirse que el rango de valores de β para los cuales se cumple dicho caso es el que se muestra en la **Figura 6-6**.

Intentando esclarecer el análisis realizado, mostremos lo expresado mediante un ejemplo.

Universitat d'Alacant
Universidad de Alicante

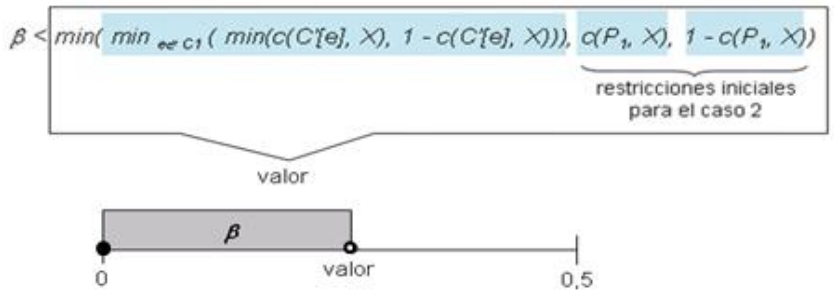


Figura 6-6 Rango de valores de β para los que se cumple el *caso 2*: $\varphi(P_1, X) = X$

Ejemplo 5-1

Supongamos que $C_1 = \{e_1, e_2, e_3, e_4\}$ y supongamos además que, $\{e_1, e_2, e_4\} \in Q_2$ y $e_3 \in Q_4$, o sea, los elementos de C_1 pertenecen sólo a Q_2 y Q_4

En tal caso, para garantizar $C_1 \subseteq B_j$, tendrían que cumplirse a la vez las siguientes condiciones:

- $\varphi(Q_2, X) \neq \phi$, lo cual sucede si $\beta < c(Q_2, X) < 1-\beta$ (a)
- $\varphi(Q_4, X) \neq \phi$, lo cual sucede si $\beta < c(Q_4, X) < 1-\beta$ (b)

Nótese que si $\varphi(Q_2, X) = \phi$ ó $\varphi(Q_4, X) = \phi$, esto significa que Q_2 ó Q_4 ,según el caso, no pertenece a la *frontera* de X, por tanto, con estas condiciones no se cumplirá que $C_1 \subseteq B_j$, que es exactamente lo que necesitamos garantizar, así que, dichos casos no son tenidos en consideración en esta parte del análisis.

- Por (a) tendría que cumplirse:

$$\beta < c(Q_2, X) < 1-\beta \Rightarrow \beta < c(Q_2, X) \quad (13)$$

$$\beta < 1 - c(Q_2, X) \quad (14)$$

Por (13) y (14):

$$\beta < min\,(c(Q_2, X),\ 1 - c(Q_2, X))$$

- Por (b) tendría que cumplirse:

$$\beta < c(Q_4, X) < 1-\beta \Rightarrow \beta < c(Q_4, X) \quad (15)$$

$$\beta < 1- c(Q_4, X) \quad (16)$$

Por (15) y (16):

$$\beta < min\ (c(Q_4, X),\ 1- c(Q_4, X))$$

Finalmente,

$$\beta < mín\ (c(Q_2, X),\ c(Q_4, X),\ 1- c(Q_2, X),\ 1- c(Q_4, X))$$

A estas restricciones, se añaden las restricciones propias del *caso 2* y a las cuales ya se había hecho referencia, o sea,

$\beta < c(P_1, X) < 1-\beta$, de lo que se deduce:

$\beta < c(P_1, X)$ y $\beta < 1- c(P_1, X)$

A partir de todo esto, se puede resumir el rango de valores de β para el *caso 2*, de la siguiente forma:

$$\beta < mín\ (c(P_1, X),\ 1- c(P_1, X),\ c(Q_2, X),\ c(Q_4, X),\ 1- c(Q_2, X),\ 1- c(Q_4, X))$$

Finalmente, podemos plantear que el conjunto de valores para los cuales $C_1 \subseteq B_j$ es la unión de las soluciones halladas para el *caso 1* y *2* respectivamente, dado que ambos casos son excluyentes a partir de que *caso 1*: $\varphi(C_1, X) = \phi$ y *caso 2*: $\varphi(C_1, X) = X$

Resumiendo todo el análisis hecho, puede plantearse lo siguiente:

$\forall\beta : \beta \in Caso\ 1 \cup Caso\ 2$, se cumple $C_1 \subseteq B_j$

Si se sigue un procedimiento similar para cada una de las C_l, $1 \leq l \leq |U/r_i|$, podremos hallar el rango general de

valores de β para el cual $B_i \subseteq B_j$. Llamaremos $\boldsymbol{I_{ij}}$ al conjunto de valores de β para los que $B_i \subseteq B_j$, $i \neq j$, $1 \leq i, j \leq m$.

Subproblema No. 2: Determinación del rango de valores de β para los cuales una *frontera interna* dada es nula

Ya sabemos determinar el rango de valores de β para los cuales cualquier *frontera interna* es subconjunto de una dada. Ahora, teniendo en cuenta que en el marco teórico en el cual se basa nuestro método de detección se supone que las *fronteras internas* tenidas en cuenta en el análisis no son nulas, debemos determinar los valores de β para los cuales una *frontera interna* dada, es nula. Para determinar esto, veamos lo que sucede para una *frontera interna* cualquiera B_i y posteriormente este resultado podrá generalizarse para cualquier otra *frontera interna* mediante un análisis similar.

Por tanto, nuestro problema en estos momentos será: Determinar el conjunto de valores de β para los cuales $B_i=\phi$, $1 \leq i \leq m$.

Si $B_i=\phi$ y además se sabe que $\forall l: C_l \subseteq B_i \Rightarrow$

$C_l=\phi \Rightarrow P_l=\phi \vee \varphi(P_l, X) = \phi$

Si $P_l=\phi \Rightarrow c(P_l, X)=0$ y como siempre $\beta \geq 0 \Rightarrow$

$c(P_l, X) = 0 \leq \beta \Rightarrow \varphi(P_l, X) = \phi$,

por tanto, siempre que algún $P_l = \phi$ podremos concluir que $\varphi(P_l, X) = \phi$

En caso de ocurrir lo expresado, o sea $\varphi(P_l, X) = \phi \Rightarrow$

$c(P_l, X) \leq \beta \vee c(P_l, X) \geq 1-\beta$ (por definición de la función φ),

por consiguiente, $\beta \geq mín(c(P_l, X), 1- c(P_l, X))$, luego

$$\forall l: \beta \geq mín(c(P_l, X), 1- c(P_l, X)) \qquad (17)$$

El conjunto de valores de β que garantiza que (17) se cumpla es el siguiente:

$$\beta \geq máx_l\,(mín(c(P_l, X),\ 1- c(P_l, X)))$$

y éste será el conjunto de valores de β para los cuales $B_i=\phi$. Llamemos $\boldsymbol{N_i}$ a dicho conjunto. En la **Figura 6-7** puede observarse una representación gráfica del mismo.

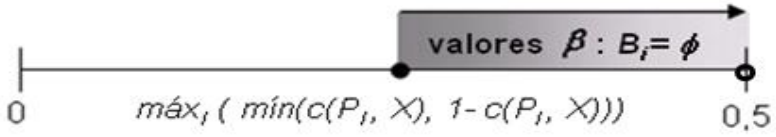


Figura 6-7 Conjunto de valores de β para los cuales $B_i=\phi$

Subproblema No. 3: Determinación del conjunto de valores de β para los cuales $B_i= B_j,\ i\neq j,\ 1\leq i,j \leq m$

Teniendo en cuenta que en el marco teórico en el cual se basa nuestro método de detección tampoco se contempla la existencia de dos *fronteras internas* iguales, es necesario también determinar el conjunto de valores de β para los cuales esto ocurre.

En esta ocasión, nuestro problema será determinar el conjunto de valores de β para los cuales $B_i= B_j,\ i\neq j,\ 1\leq i,j \leq m$ y esto se deduce fácilmente a través de la siguiente secuencia de equivalencias:

$$B_i= B_j \Leftrightarrow B_i \subseteq B_j \wedge B_j \subseteq B_i \Leftrightarrow \beta \in I_{ij} \wedge \beta \in I_{ji} \Leftrightarrow \beta \in I_{ij} \cap I_{ji}$$

A partir de lo cual podemos resumir que el conjunto de valores de β para los cuales $B_i= B_j,\ i\neq j,\ 1\leq i,j \leq m$, es el siguiente:

$$EQ_{ij} = \{\beta : \beta \in I_{ij} \cap I_{ji}\}$$

Finalmente, hemos concluido el análisis de los tres subproblemas que nos planteamos inicialmente y a partir de ello podremos dar un criterio general en relación a

cuándo una *frontera interna* es subconjunto de otra. Estableciendo la siguiente secuencia de conjuntos podemos llegar a conclusiones:

- $(I_{1j} - EQ_{1j} - N_1) \cup (I_{2j} - EQ_{2j} - N_2) \cup \ldots \cup (I_{mj} - EQ_{mj} - N_m) = A$,

  A: Conjunto de valores de β para los cuales existe una *frontera interna* no vacía que es *subconjunto propio* de la *frontera interna j*.

- A^c: Conjunto de valores de β para los cuales ninguna *frontera interna* no vacía es *subconjunto propio* de la *frontera interna j*.

- $S_j = A^c - N_j$

  S_j: Conjunto de valores de β para los cuales ninguna *frontera interna* no vacía es *subconjunto propio* de la *frontera interna j* excluidos los valores para los cuales dicha *frontera* es vacía.

Sabiendo que, para que todos los elementos de B_j sean *outliers* debe cumplirse que no exista otra *frontera interna* que sea subconjunto de ella, a partir de los resultados anteriores podemos plantear que esto sucederá solo cuando $\beta \in S_j$.

S_j representa el intervalo de valores de β para el cual, un elemento x del *universo*, $x \in B_j$, pertenece a algún *conjunto excepcional no redundante*, por lo cual x es un posible *outlier*.

A continuación haremos un análisis similar para determinar el conjunto de valores del *umbral de excepcionalidad* μ para el cual cada elemento del *universo* podrá ser considerado un *outlier*.

Determinación de la región de excepcionalidad con respecto al conjunto de valores de μ

Nuestro problema ahora es el siguiente:

Dado un elemento $a \in U$, determinar el rango de valores del umbral μ, para los cuales el *grado de excepcionalidad* de a es mayor que μ.

Los elementos teóricos que resultan necesarios para garantizar la solución de este problema se plantean siguiendo la secuencia lógica que se expresa a continuación:

- Definir el conjunto de valores de β para los cuales $\forall a$: $a \in U$ pertenece a la *frontera interna* B_i, $1 \leq i \leq m$.
- Establecer una nueva definición de *grado de excepcionalidad* $\forall a$: $a \in U$, bajo una nueva interpretación en la cual intervienen los valores de β
- Determinar $\forall a \in U$ el rango de valores de μ *para los* cuales $GrExcep(a,\beta) \geq \mu$ para un valor β dado.

Siguiendo la secuencia expresada, primeramente definiremos el conjunto de valores de β para los cuales $a \in U$ pertenece a la *frontera interna* B_i, $1 \leq i \leq m$.

Universitat d'Alacant
Universidad de Alicante

Definición 26:

Sea U un *universo* de datos dado y sea X el conjunto de valores de U que cumplen un *concepto*. $\forall a \in U$, $1 \leq i \leq m$, W es una *clase de equivalencia* de la partición inducida por la *relación de equivalencia* r_i en U, tal que, $a \in W$. El conjunto de valores de β para los cuales a pertenece a la *frontera interna* B_i se define como:

$$M_i(a) = \begin{cases} \{\beta : \beta < c(W, X) < 1 - \beta\} & si \quad a \in X \\ \phi & si \quad a \notin X \end{cases}$$

Desde el punto de vista teórico la interpretación de la **Definición 26**, sería la siguiente:

Supongamos que, $a \in U$. Si se evalúa $M_i(a)$, entonces el resultado de esta evaluación puede interpretarse de la siguiente forma:

- si $M_i(a) = \phi \Rightarrow a$ no cumple el *concepto* y por tanto no es miembro de la *frontera interna* B_i.
- si $M_i(a) = \{\beta: \beta < c(W, X) < 1-\beta\} \Rightarrow a$ cumple el *concepto* y por tanto para que a pertenezca a la *frontera interna* B_i es necesario que $\varphi(W, X) \neq \phi$.

Consecuentemente, según lo establecido por $M_i(a)$, las restricciones que deben cumplir los valores del parámetro β para garantizar que a pertenezca a la *frontera interna* B_i, son las siguientes:

$\beta < c(W, X) < 1-\beta \Rightarrow [\beta < 1-c(W, X)] \wedge [\beta < c(W, X)]$

En función de esto, a partir de $M_i(a)$ se puede establecer que el intervalo de valores de β dentro del cual se garantiza $a \in B_i$ es el siguiente:

$\forall \beta: \beta \in [0, min(c(W, X), 1-c(W, X))$

La **Figura 6-8** ilustra dicho intervalo.

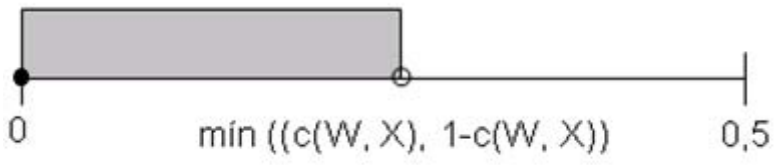


Figura 6-8 Conjunto de valores de β determinado por $M_i(a)$

A partir de lo mostrado en la **Figura 6-8**, se observa que, cuando $min(c(W, X), 1- c(W, X)) = 0$, el intervalo al que se hace referencia es *vacío*.

Universitat d'Alacant
Universidad de Alicante

Un criterio necesario para afirmar que un elemento $a \in U$ puede ser candidato a *outlier* es que pertenezca a alguna *frontera interna.* Teniendo en cuenta esto, se establece a continuación una nueva definición de *grado de excepcionalidad* de un elemento $a \in U$, bajo una nueva interpretación: dependencia de la misma con relación a los valores de β.

Previamente, se hace necesario dar una nueva definición y establecer una proposición a partir de ella.

Definición 27:

$\forall a \in U,\ 1 \leq i \leq m$

$$\lambda_i(a) = \begin{cases} Sup(M_i(a)) & si\ M_i(a) \neq \phi \\ 0 & en\ otro\ caso \end{cases}$$

donde:

$Sup(M_i(a))$ es el menor valor de β que es mayor que todos los valores del intervalo $M_i(a)$. De manera informal, podríamos decir que es el valor del extremo abierto del intervalo que se ilustra en la **Figura 6-8**.

En general, para todo $\beta < \lambda_i(a)$ el elemento a pertenece a la *frontera interna* B_i, de lo que se deriva el siguiente resultado:

Proposición 28:

$\forall a \in U,\ 1 \leq i \neq j \leq m,$

$Si\ \lambda_i(a) \leq \lambda_j(a) \Rightarrow \forall \beta: \beta < \lambda_i(a),\ a \in B_i \wedge a \in B_j$

Demostración:

Sea $a \in U,\ 1 \leq i \neq j \leq m$. Si $\lambda_i(a) \leq \lambda_j(a)$, dada la forma en que están definidos los intervalos $M_i(a)$ y $M_j(a)$ (**Figura 6-8**), entonces $M_i(a) \subseteq M_j(a)$ y por tanto, $\forall \beta: \beta < \lambda_i(a)$ se cumplirá que $\beta \in M_i(a) \wedge \beta \in M_j(a)$. Por tanto, $a \in B_i \wedge a \in B_j$, (*vea* **Definición 26**).

La **Figura 6-9** muestra la relación entre los intervalos $M_i(a)$ y $M_j(a)$ a partir de los valores $\lambda_i(a)$ y $\lambda_j(a)$, respectivamente. ($M_i(a)$ es el conjunto de valores de β para los cuales el elemento a pertenece a la *frontera interna* B_i).

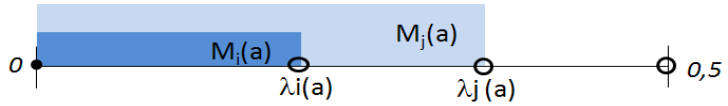


Figura 6-9 Relación entre los intervalos $M_i(a)$ y $M_j(a)$

A partir de lo que acabamos de analizar, podemos obtener para cada elemento $a \in U$ una ordenación particular de los supremos $\lambda_i(a)$, $1 \leq i \leq m$; asociados a cada una de las *fronteras internas* B_i. Sea $Z_i(a)$:

$Z_1(a),\ldots, Z_m(a)$, tal que $\lambda_{Z1(a)}(a) \leq \ldots \leq \lambda_{Zm(a)}(a)$.

Resulta evidente que $Z_i(a)$ es una permutación de índices que ordena a los $\lambda_i(a)$. Esto se muestra en la **Figura 6-10**.

Universitat d'Alacant
Universidad de Alicante

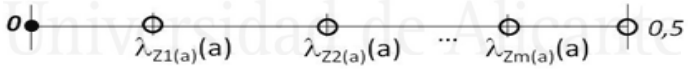


Figura 6-10 Ordenación de los $\lambda_i(a)$ a partir de la permutación $Z_i(a)$

Observemos que:

- $\forall\beta$ *tal que* $\beta \in [0; \lambda_{Z1(a)}(a))$, a pertenece a TODAS las *fronteras internas* (Por **Proposición 28**).
- $\forall\beta$ tal que $\beta \in [\lambda_{Z1(a)}(a), \lambda_{Z2(a)}(a),)$ entonces, a pertenece a TODAS *las fronteras internas, e*xcepto a la B_1. (Por **Proposición 28** y **Definición 26**).

- $\forall\beta$ *tal que* $\beta\in[\lambda_{Z2(a)}(a), \lambda_{Z3(a)}(a),)$, *a* pertenece a TODAS *las fronteras internas,* excepto a la B_1 y a la B_2. (Por **Proposición 28** y **Definición 26**).

y así sucesivamente…

La siguiente definición también ayudará a redefinir el concepto de *grado de excepcionalidad* para un elemento $a\in U$.

Definición 29:

Sean $a\in U$, $\beta\in[0; 0,5)$ y sea m la cantidad de *fronteras internas* tenidas en cuenta en el análisis. Se define el *Total* de *fronteras internas* a las que pertenece el elemento a para el valor β dado, de la siguiente forma:

$$Total(a,\beta)=\begin{cases} m & si \quad \beta < \lambda_{Z1(a)}(a) \\ 0 & si \quad \beta \geq \lambda_{Zm(a)}(a) \\ m - máx_k(\beta \geq \lambda_{Zk(a)}(a)) & en \quad otro\_caso \end{cases}$$

Las dos primeras partes de la **Definición 29** se establecen para garantizar que cuando se evalúe la función *máx* siempre se establezca un resultado definido (especialmente cuando no se satisfaga la condición establecida en el predicado $\beta \geq \lambda_{Zk(a)}(a)$).

La interpretación gráfica de la función *Total(a, β)* se ilustra en la **Figura 6-11**. En esta figura se considera $\upsilon = máx_k(\beta \geq \lambda_{Zk(a)}(a))$. Este valor es el mayor valor de k tal que $\beta \geq \lambda_{Zk(a)}(a)$, o sea, es exactamente la cantidad de *fronteras internas* a las cuales a no pertenece. Además, a partir de $k' = k+1$, se cumplirá $\beta < \lambda_{Zk'(a)}(a)$ y por tanto a pertenece a las *fronteras internas* $B_{Zk'(a)}, \ldots, B_{Zm(a)}$, por la **Proposición 28** y no pertenece a las *fronteras internas* $B_{Z1(a)}, \ldots, B_{Zk(a)}$.

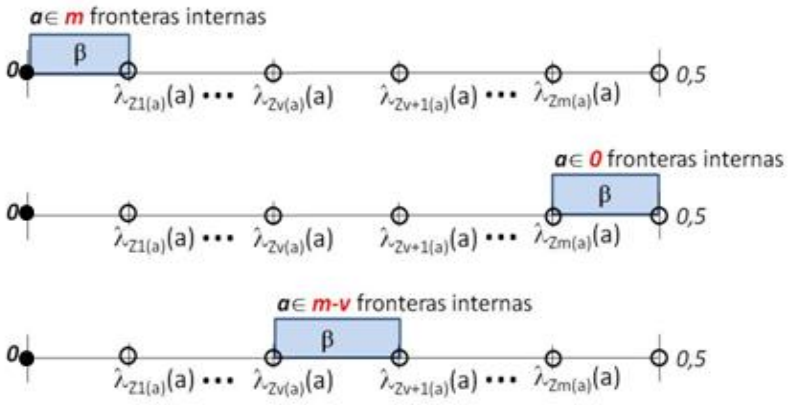


Figura 6-11 Interpretación gráfica de la función Total(a, β)

A partir de la **Definición 27**, la **Proposición 28** y la **Definición 29**, podemos redefinir el concepto de *grado de excepcionalidad* de un elemento $a \in U$ bajo una nueva interpretación, a partir de la cual se establece el *grado de excepcionalidad* de los elementos del universo en función de los valores de β.

Definición 30:

Sean $a \in U$, un valor $\beta \in [0;\ 0{,}5)$ y sea m la cantidad de *fronteras internas* tenidas en cuenta en el análisis. Se define el *grado de excepcionalidad* del elemento a para el valor β dado, de la siguiente forma:

$$GrExcep(a, \beta) = Total(a, \beta)/m$$

Es evidente que esta nueva definición de *grado de excepcionalidad* no entra en contradicción con la que hasta ahora hemos tenido en cuenta. Esta tiene como particularidad el hecho de que involucra en ella a un valor de β dado.

A partir de esto, $\forall a \in U$ se puede obtener el *grado de excepcionalidad* de dicho elemento para cualquier valor de β y por consiguiente, pueden obtenerse también los valores de μ para los cuales $GrExcep(a,\beta) \geq \mu$. La **Figura 6-12** ilustra gráficamente mediante un ejemplo lo que se acaba de expresar.

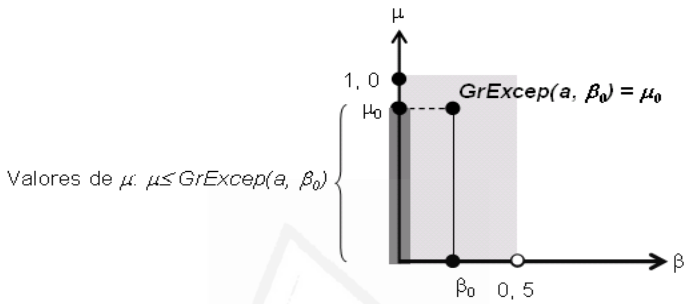


Figura 6-12 Valores de μ para los cuales $\mu \leq GrExcep(a,\beta_0)$

Una vez que se determine $\forall a \in U$ el rango de valores de μ *para los* cuales $GrExcep(a,\beta) \geq \mu$ para un valor β dado, estamos en condiciones de establecer el procedimiento general para determinar los valores de β *y* μ para los cuales el elemento $a \in U$ es *outlier* en U.

Integración de resultados

El marco teórico alcanzado permite establecer el siguiente procedimiento —método— general para determinar los valores de β *y* μ para los cuales el elemento $a \in U$ es *outlier* en U:

1. Determinar $M_i(a)$:

   Valores de β para los cuales el elemento $a \in B_i$

2. Determinar S_i:

Valores de β para los cuales no existe una *frontera interna* que sea subconjunto de la *frontera interna* B_i.

3. Determinar $D_i(a) = M_i(a) \cap S_i$:

   Valores de β para los cuales el elemento a pertenece a B_i y no existe una *frontera interna* que sea subconjunto de la *frontera interna* B_i.

   Para los valores de $\beta \in D_i(a)$, el elemento a pertenece a algún *conjunto excepcional no redundante* y es el único representante de la *frontera interna* B_i en dicho conjunto, o sea, para los valores de β en $D_i(a)$, $a \in Ei$.

4. $\forall \beta_o, \mu_o: \beta_o \in \bigcup_{k=1}^{m} D_k(a) \wedge \mu_o \leq GrExcep(a, \beta_o)$, entonces:

   a es un outlier en U

Un $\beta_o \in \bigcup_{k=1}^{m} D_k(a)$ representa un valor para el cual el elemento a pertenece a alguna *frontera interna* de la cual ninguna otra *frontera interna* es subconjunto y, en tal caso, se restringe μ_o a ser menor o igual que $GrExcep(a, \beta_o)$.

La **Figura 6-13** ilustra la región de valores β-μ para los cuales un elemento a cualquiera del *universo* es *outlier* en U. En este caso y con el objetivo de simplificar el resultado en aras de la claridad, se ha minimizado el problema suponiendo que:

$$intervalo\ (1) \cup intervalo\ (2) = \bigcup_{k=1}^{m} D_k(a)$$

La viabilidad computacional del procedimiento expuesto se valida a partir de la concepción de un algoritmo que determina, de forma no supervisada, la región de valores de los umbrales μ-β bajo la cual cada elemento del universo es

outlier. A continuación se dan detalles sobre la implementación computacional de dicho algoritmo.

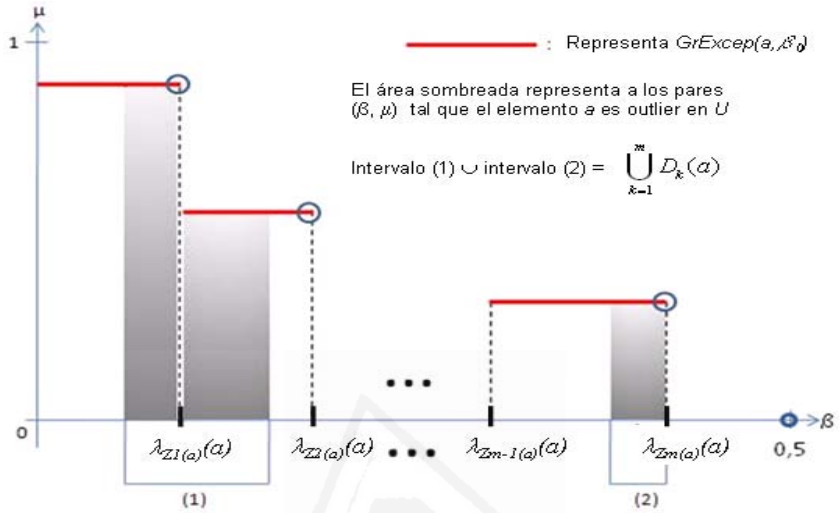


Figura 6-13 Región de valores β-μ para los cuales un elemento *a* cualquiera del *universo* es *outlier* en *U*

Ejemplos

Seguidamente se presentan varios ejemplos que ilustran la *región de excepcionalidad* determinada por el algoritmo *BM*, para elementos representativos por su nivel de *excepcionalidad* dentro de un determinado *conjunto de datos*.

Los elementos representados se han escogido del *conjunto de datos Arrhythmia Data Set* descrito en la **Prueba 5-4.2.**

Ejemplo 6-1

En la **Figura 6-14** se muestra la *región de excepcionalidad* de uno de los elementos *más contradictorios* introducidos de forma intencional en el *conjunto de datos*.

La interpretación de esta gráfica sería la siguiente:

El intervalo de valores $\beta\in$ [0; 0,36) que corresponde a la base de la región representada establece el conjunto de valores de dicho umbral para los cuales el elemento pertenece a algún *conjunto excepcional no redundante*.

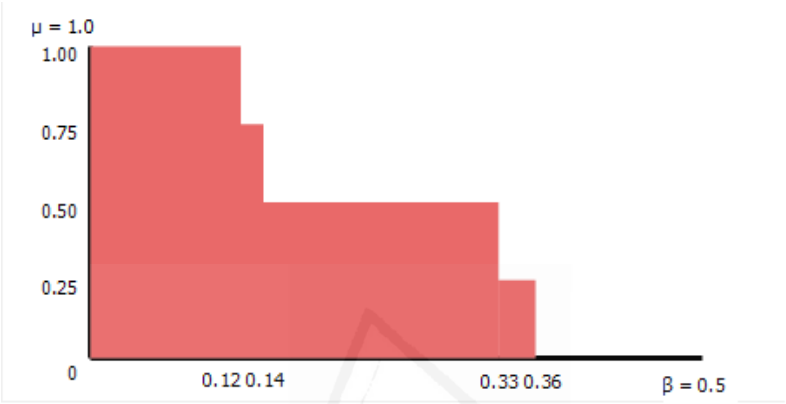


Figura 6-14 *Región de excepcionalidad* para un elemento *muy contradictorio*

Para valores de $\beta\in$[0; 0,12) el elemento en cuestión, por ser de los más contradictorios, pertenece a todas las *fronteras internas* y su *grado de excepcionalidad* en dicho intervalo es el máximo posible, o sea, 1. Por tanto, para valores de $\beta\in$[0; 0,12) y valores de$\mu\leq 1$, el elemento es *outlier* en *U.*

Al aumentar el *grado de desclasificación* permitido, β: $\beta\in$[0,12; 0,14), el elemento varía su *grado de excepcionalidad* a 0,75. Por tanto, para valores de $\beta\in$[0,12; 0,14) y valores de $\mu\leq 0{,}75$, el elemento es *outlier* en *U.*

De igual forma, para valores de$\beta\in$[0,14; 0,33), el *grado de excepcionalidad* es de 0,5. En este caso puede entonces decirse que para valores de $\beta\in$[0,14; 0,33) y valores de $\mu\leq 0{,}5$, el elemento es *outlier* en *U.* Algo similar sucede para valores de $\beta\in$[0,33; 0,36) y valores de $\mu\leq$0,25.

Para *grados de desclasificación* mayores o iguales que 0,36 el elemento ya no pertenecerá a ninguna frontera interna, por tanto, su *grado de excepcionalidad* es 0 y no existe la posibilidad de que sea *outlier*.

Ejemplo 6-2

En la **Figura 6-15** se muestra la *región de excepcionalidad* de uno de los elementos *menos contradictorios* del *conjunto de datos*.

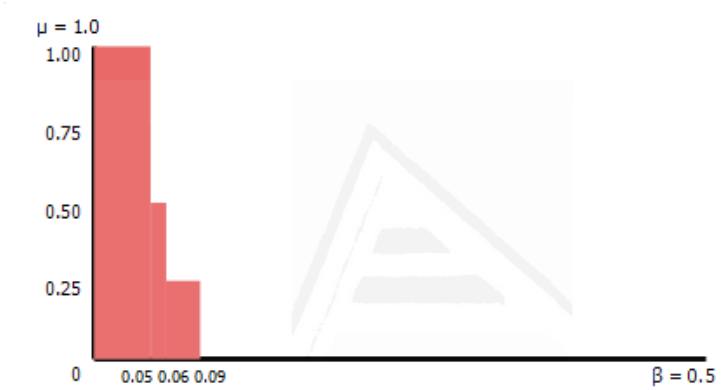


Figura 6-15 *Región de excepcionalidad* para un elemento *muy poco contradictorio*

Universitat d'Alacant
Universidad de Alicante

La interpretación de esta gráfica sería la siguiente:

El intervalo de valores $\beta \in$ [0; 0,09), que corresponde a la base de la región representada, establece el conjunto de valores de dicho umbral para los cuales el elemento pertenece a algún *conjunto excepcional no redundante*.

Como se observa, estamos en presencia de un elemento *muy contradictorio* para pocos valores de β. Un análisis similar al realizado para el ejemplo anterior permite apreciar como varía el *grado de excepcionalidad* del elemento al variar el *grado de desclasificación* permitido.

Ejemplo 6-3

En la región representada en la **Figura 6-16** se observa una variación de color en una parte de la misma. El significado de esta situación es el siguiente:

En la subregión de color más claro, para los valores de $\beta \in [0{,}09;\ 0{,}14)$, el elemento tiene un *grado de excepcionalidad* de 0,25 pero, sin embargo, no pertenece a ningún *conjunto excepcional no redundante*. Una justificación en torno a este hecho es que, en este caso, todas las fronteras internas a las que el elemento pertenece tienen como *subconjunto propio* a otra frontera interna a la cual el elemento no pertenece. Las consecuencias de una situación de este tipo fueron descritas, detalladamente, en la justificación teórica de la *Fase 2* del algoritmo *RSBM* en el **Capítulo 4**.

Aunque el algoritmo determina los valores de la subregión antes comentada, la misma no se considera como parte de la *región de excepcionalidad* del elemento.

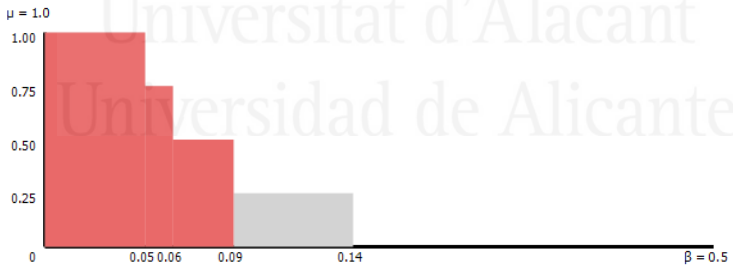


Figura 6-16 *Región de excepcionalidad* para un elemento que, para determinados valores de *β*, a pesar de su *grado de excepcionalidad*, no pertenece a ningún *conjunto excepcional no redundante*

Universitat d'Alacant
Universidad de Alicante

Implementación computacional. Algoritmo *BM*

En la concepción general del algoritmo para determinar la región de valores de los umbrales β y μ para los cuales un objeto cualquiera del *universo U* es *outlier* en *U*, se destacan dos tareas fundamentales. De cada una de ellas se hizo un análisis concreto y a cada una se le dio una solución computacional. Estas tareas son las siguientes:

Tarea No. 1

Cálculo de las dependencias entre las *fronteras internas* (relación de inclusión entre ellas).

Tarea No. 2

Búsqueda de la región β-μ para la cual cada elemento de *U* es *outlier.*

Debido a la complejidad de cada una de estas tareas, y para presentar de forma clara y modular el algoritmo, se han desglosado las mismas en diferentes subtareas (métodos). Dichas subtareas están en correspondencia con los aspectos teóricos que se han visto en el análisis precedente y cada uno de los cuales interviene en la solución general del problema.

Universidad de Alicante

El orden en que se presentan los métodos, es importante debido a que se establece según la relación de dependencia existente entre ellos. De cada uno se hace un análisis de su *complejidad temporal* y se tiene en cuenta la validez de las siguientes proposiciones:

Proposición 31:

Sea *r* una *relación de equivalencia* y *U* un *universo* de datos dados. La cantidad de *clases de equivalencia* en las que *r* particiona a *U* es, a lo sumo, la cardinalidad de *U* (caso particular en que cada *clase de equivalencia* contiene a un solo elemento del *universo*).

Proposición 32:

La solución computacional de las operaciones de intersección, unión y diferencia entre dos intervalos de números reales cualesquiera, tiene *complejidad temporal constante, O(1).* La operación unión supone como resultado un conjunto de, a lo sumo, dos intervalos ordenados, mientras que las operaciones de intersección y diferencia dejan como resultado un conjunto de, a lo sumo, un intervalo.

Proposición 33:

Sean X_1 y X_2 dos conjuntos ordenados de intervalos de números reales, donde los intervalos que pertenecen a cada conjunto particular son disjuntos dos a dos. Entonces, la solución computacional de las operaciones de intersección, unión y diferencia entre X_1 y X_2 tiene *complejidad temporal* $O(|X_1|+|X_2|)$. La cardinalidad del conjunto de intervalos resultante de cualquiera de estas operaciones es, a lo sumo, $|X_1| + |X_2|$.

Proposición 34:

Sean $X_1, X_2, \ldots, X_p$ conjuntos ordenados de intervalos de números reales, donde los intervalos que pertenecen a cada conjunto particular son disjuntos dos a dos y sea $q = máx\{|X_1|, |X_2|, \ldots, |X_p|\}$. Entonces, la solución computacional de las operaciones de intersección y unión entre todos estos conjuntos es $O(p \times q \times log(p))$. La cardinalidad del conjunto de intervalos resultante de cualquiera de estas operaciones es, a lo sumo, $p \times q$.

La demostración formal de estas proposiciones forma parte de resultados teóricos importantes de la *Matemática*, así como de disciplinas afines a la *Ciencia de la Computación*, como por ejemplo, la referida al *Diseño y Análisis de Algoritmos*. Por tal motivo, no se incluyen en este trabajo

tales demostraciones, pues no son consecuencia de resultados alcanzados en el marco de esta investigación.

Antes de presentar los métodos, se presentan las estructuras de datos que intervienen en el algoritmo, desglosadas por tareas. Existe una relación directa entre éstas y los métodos. La funcionalidad de cada uno de ellos es, en esencia, el cálculo del valor –o de los valores– asociado a cada estructura de datos.

Tarea No. 1

case1[i][ec]

Dada la *clase de equivalencia ec* inducida por la *relación de equivalencia i*, la estructura almacena la solución para el *caso 1:* $\varphi(ec, X) = \phi$, donde: $1 \leq i \leq m$, $1 \leq ec \leq |U/r_i| \leq |U|$.

case2[i][j][ec]

Dada la *clase de equivalencia ec* inducida por la *relación de equivalencia i* y ante el problema de determinar si *ec* es subconjunto de la *frontera interna* asociada a la *relación de equivalencia j*, la estructura almacena la solución para el *caso 2:* $\varphi(ec, X) = X$, donde: $1 \leq i, j \leq m$, $i \neq j$, $1 \leq ec \leq |U/r_i| \leq |U|$

Subset[i][j]

Dadas las *fronteras internas i* e *j*, la estructura almacena los valores de β para los cuales la *frontera interna i* es subconjunto de la *frontera interna j*, donde: $1 \leq i, j \leq m$, $i \neq j$.

Equal[i][j]

Dadas las *fronteras internas i* e *j,* la estructura almacena los valores de β para los cuales la *frontera interna i* es igual a la *frontera interna j,* donde: $1 \leq i, j \leq m, i \neq j$.

Null[i]

Dada la *frontera interna i*, la estructura almacena los valores de β para los cuales la *frontera interna i* es nula, donde: $1 \leq i \leq m$.

S[i]

Dada la *frontera interna i*, la estructura almacena los valores de β que conforman el conjunto S_i, donde: $1 \leq i \leq m$

Tarea No. 2

Universitat d'Alacant
Universidad de Alicante

Lambda[e][i]

Dado $e \in U$ y la *frontera interna i*, la estructura almacena el valor λ_i asociado al elemento *e*, donde: $1 \leq i \leq m$.

GrExcep[e]

Dado $e \in U$, la estructura almacena la función $F(\beta)=GrExcep(e, \beta)$. $F(\beta)$ es una proyección de la función $GrExcep(e, \beta)$ sobre el eje β.

Pertain[e][i]

Dados $e \in U$ y la *frontera interna i*, la estructura almacena los valores que conforman el conjunto

$M_i(e)$, o sea, los valores de β para los cuales el elemento e pertenece a la *frontera interna i,* donde: $1 \leq i \leq m$.

Outlier[e]

Dado $e \in U$, la estructura almacena el conjunto de valores de β para los cuales el elemento e pertenece a algún *conjunto excepcional no redundante* (o sea, almacena $D_i(e)$).

Result[e]

Dado $e \in U$, la estructura almacena la región *beta- miu* en la cual el elemento e es un *outlier* en U.

A continuación se presenta una descripción de los métodos fundamentales que fueron implementados, desglosados también por tareas.

Tarea No. 1

```
(1) for i:= 1 to m // iterar × todas las fr.int.
(2)     for ec in r_i // iterar × todas las c.eq. de r_i
(3)         case1[i][ec]=[min(c(ec,X), 1-c(ec, X)), 0.5)
```

Método 1: Cálculo de los valores de la estructura de datos case1[i][ec]

Complejidad temporal del **Método 1**:

$O(n \times m \times c)$, $\Omega(n \times m \times c)$

```
(1) for i:=1 to m //inicialización de la estructura
(2)     for j:=1 to m
(3)         if i≠j
```

```
(4)            then
(5)                 for ec in ri
(6)                 case2[i][j][ec]=[0; min(c(ec,X),1-c(ec,X))]
                     //valores iniciales del CASO2
(7)  for e in U //iterar x todos los elementos de U
(8)       if e∈X  //e cumple el concepto X
(9)       then
(10)           for i:=1 to m //itera xtodas las fr.int.
(11)               ec = ri[e] //c.eq. de e por ri
(12)               for j:=1 to m //para el resto de las fr.int.
(13)                    if i≠j
(14)                    then
(15)                         oc = rj[e] //c.eq. de e por rj
(16)                         case2[i][j][ec] =
                               case2[i][j][ec] ∩
                              [0; min(c(oc,X), 1-c(oc,X))]
// actualizar los valores para los cuales (ec ∩ X)⊆ fr. int. j
```

Método 2: Cálculo de los valores de la estructura de datos **case2[i][j][ec]**

Universitat d'Alacant
Universidad de Alicante

Complejidad temporal del **Método 2**:

$O(n \times m^2 \times c)$, $\Omega(n \times m^2 \times c)$

```
(1)  for i:=1 to m //inicialización de la estructura
(2)       for j:=1 to m
(3)            if i≠j
(4)            then
// En esencia intersecta todas las soluciones particulares
// obtenidas para las cl. equiv. inducidas por la relación de
// equiv. i ante el problema de determinar si dichas clases eran
// subconjunto de la front. int. Bj. Como resultado de esta
// intersección se determina el conjunto de valores para los cuales
```

// la frontera interna *i* es subconjunto de la frontera interna *j*

(5) $\text{Subset}[i][j] = \bigcap_{ec \in ri} (\textbf{case1}[i][ec] \cup \textbf{case2}[i][j][ec])$

// asignar el resultado de intersectar todas las soluciones
// particulares para cada clase de equivalencia

Método 3: Cálculo de los valores de la estructura de datos **Subset[i][j]**

Complejidad temporal del **Método 3**:

$O(m^2 \times n \times \log(n)), \Omega(m^2)$

Es importante resaltar que: |**Subset[i][j]**|es $O(n), \Omega(1)$

(1) **for** i := 1 **to** m // iterar por todas las fronteras internas
(2) **for** j := 1 **to** m // iterar por las restantes fronteras
(3) **if** i≠j
(4) **then**
(5) Equal[i][j]=**Subset[i][j]** ∩**Subset[j][i]**
// $B_i = B_j \Leftrightarrow (B_i \subseteq B_j) \wedge (B_j \subseteq B_i)$

Método 4: Cálculo de los valores de la estructura de datos **Equal[i][j]**

Complejidad temporal del **Método 4**:

$O(n \times m^2)$, $\Omega(m^2)$

Es importante resaltar lo siguiente: |**Equal[i][j]**|es $O(n), \Omega(1)$

(1) max_beta = 0

Universitat d'Alacant
Universidad de Alicante

```
(2) for i:= 1 to m // iterar por todas las fronteras internas
(3)       for ec in r_i  // iterar por cada clase de equivalencia
// para cada ec calculo su error de clasificación respecto a X.
// Una vez que esto ha sido calculado para cada ec, tomando β mayor
// que el mayor error de clasificación obtenido para alguna clase,
// se determina el rango de valores de β para los cuales la
// frontera interna i es nula, pues para dichos valores de β, todas
// las clases de equivalencia pasarían a región positiva o negativa
(4)             max_beta = max(max_beta, min(c(ec,X), 1-c(ec,X))
//actualizar el máximo error de clasificación de las cl. de equiv.
(5)       Null[i]=[max_beta, 0.5) // intervalo de valores de beta
                                  // en el cual la frontera
                                  // interna i es nula
```

Método 5: Cálculo de los valores de la estructura de datos **Null[i]**

Complejidad temporal del **Método 5**:

$O(n \times m)$, $\Omega(m)$

Universitat d'Alacant
Universidad de Alicante

```
(1) for i:= 1 to m // iterar por todas las fronteras internas
(2)       S[i]=([0, 0.5] - Null[i])
```

$$- \bigcup_{\substack{j=1 \\ j \neq i}}^{m} (\mathbf{Subset[j][i] - Equal[i][j] - Null[j]})$$

```
// A^c = [0, 0.5) – A
// S_i = A^c – Null[i]
//     = ([0, 0.5] – A) – Null[i]
//     = ([0, 0.5] – Null[i]) - A
```

Método 6: Cálculo de los valores de la estructura de datos **S[i]**

Complejidad temporal del **Método 6**:

$O(n \times m^2 \times log(m))$, $\Omega(m^2)$

Es importante resaltar que: |**S[i]**|es $O(n \times m)$, $\Omega(1)$

Tarea No. 2

Búsqueda de la región β-μ para la cual cada elemento de U es *outlier.*

```
(1) for e in U // iterar por todos los elementos del universo
(2)     for i := 1 to m // iterar por todas las front. int.
(3)         ec = r_i[e] // clase equiv. de e según la relación i
(4)         Lambda[e][i] = min(c(ec, X), 1-c(ec, X))
```

Método 7: Cálculo de los valores de la estructura de datos **Lambda[e][i]**

Complejidad temporal del **Método 7**:

$O(n \times m \times c)$, $\Omega(n \times m \times c)$

```
(1) for e in U // ∀e ∈U definir GrExcep[e](β)
(2)     sorted_lambdas = Sort(Lambda[e])
        // obtener los valores λ_i asociados a e de forma
        // ordenada
(3)     present = m
        // total front. Intern. a las que pertenece e
(4)     prev = 0 // prev = valor de λ en la iteración anterior
                 //(inicialmente 0)
(5)     GrExcep[e](β) = 0 // caso base
(6)     for i := 1 to m // iterar por todos los λ asociados a e
(7)         if Lambda[e][i] > prev
```

(8) **then**

(9) $$GrExcep[e](\beta) = \begin{cases} present/m & si \quad \beta \in [prev, Lambda[e][i]) \\ GrExcep[e](\beta) & otro \quad caso \end{cases}$$

//La función GrExcep[e](β) para valores de β entre *prev* y //Lambda[e][i] vale *present/m* En esencia se está definiendo una //función por partes y en cada iteración se le asigna una nueva //parte a la misma, con lo cual, su dominio se va aumentando //progresivamente

(10) prev = **Lambda[e][i]**

//Actualizar el valor previo de *Lambda* para la //próxima iteración

(11) present = present - 1

// Actualizar TOTAL *front. int.* a las que ∈e

Método 8: Cálculo de los valores de la estructura de datos **GrExcep[e]**

Complejidad temporal del **Método 8**:

$O(n \times m \times log(m))$, $\Omega(n \times m \times log(m))$

(1) **for** e in U // iterar por todos los elementos de U

(2) **for** i:= 1 **to** m // iterar × todas las front. internas

(3) Pertain[e][i] = [0; **Lambda[e][i]**) //calculo *M(e)*

Método 9 Cálculo de los valores de la estructura de datos **Pertain[e][i]**

Complejidad temporal del **Método 9**: $O(n \times m)$, $\Omega(n \times m)$

(1) **for** e in U // iterar por todos los elementos de U

(2) $Outlier[e] = \bigcup_{i=1}^{m} S[i] \cap Pertain[e][i]$

// asignar el resultado de unir las

// soluciones para el elemento por cada

Universitat d'Alacant Universidad de Alicante

```
// frontera interna (Di(e))
```

Método 10: Cálculo de los valores de la estructura de datos ***Outlier*[e]**

Complejidad temporal del **Método 10**:

$O(n^2 \times m^2 \times log(m))$**,** $\Omega(n \times m)$

Es importante resaltar que: |***Outlier*[e]**| es $O(n \times m^2)$, $\Omega(1)$

El **Método 11** que se muestra a continuación, es el método principal del algoritmo *BM*. En él se realiza el cálculo de los valores de la estructura de datos **Result[e],** que constituye la salida dicho algoritmo.

```
(1) for e in U // iterar por todos los elementos de U
(2)       Result[e] = {x, y: x∈Outlier[e] ^ 0≤y<GrExcep[e](x)}
```

Método 11. Método Principal del algoritmo BETA-MIU

Complejidad temporal del **Método 11**:

$O(n^2 \times m^2)$**,** $\Omega(n)$

Es importante resaltar que: |**Result[e]**| es $O(n \times m^2)$, $\Omega(1)$

En todos los casos, *n* es la cardinalidad del *conjunto de datos*, *m* es el número de *relaciones de equivalencia* tenidas en cuenta en el análisis y *c* es el coste de aplicar un clasificador.

Para la implementación computacional se escogió como lenguaje de desarrollo *C Sharp,* para hacer uso de las posibilidades de las herramientas que proporciona el *Visual*

Universitat d'Alacant
Universidad de Alicante

Studio 2008. Al analizar la funcionalidad de los métodos que era necesario diseñar, se observó que en los mismos había un conjunto de tareas que se repetían en varios de ellos. Por ejemplo, hacer un recorrido por todos los elementos del *universo*, recorrer todas las *fronteras internas*, etc. Ante esto, y con el objetivo de optimizar el tiempo de ejecución general del algoritmo, en el método de trabajo seguido en la implementación: se factorizaron todas las tareas que tenían cálculos comunes para disminuir el acceso a *disco* (acceso a la Base de Datos).

Estudio de la complejidad espacial y temporal. Algoritmo *BM*

Antes de hacer el análisis de la *complejidad temporal* del algoritmo, es necesario analizar la *complejidad espacial* de las principales *estructuras de datos* utilizadas en el mismo. A continuación, se presenta información detallada sobre las mismas:

Universitat d'Alacant
Universidad de Alicante

Estructura de datos: Case1[i][ec]

Complejidad espacial de un objeto de la estructura: *O(1)*

Operación necesaria para obtener el valor de UN elemento almacenado en la estructura:

`[min(c(ec,X), 1-c(ec, X)), 0.5]`

Complejidad espacial de la estructura: $O(m \times n)$

Descripción computacional de la estructura:

Array bidimensional de intervalos de números reales.

Estructura de datos: Case2[i][j][ec]

Complejidad espacial de un objeto de la estructura: *O(1)*

Operaciones necesarias para obtener el valor de UN elemento almacenado en la estructura:

Opción 1:

$[0; min(c(ec,X), 1-c(ec, X)))$

Opción 2:

$case2[i][j][ec] \cap [0; min(c(oc,X),1-c(oc,X)))$

Complejidad espacial de la estructura:

$O(m^2 \times n)$

Descripción computacional de la estructura:

Array tridimensional de conjuntos de intervalos de números reales, donde los intervalos de cada conjunto son disjuntos dos a dos.

Estructura de datos: Subset[i][j]

Complejidad espacial de un objeto de la estructura: *O(n)*

Operación necesaria para obtener el valor de UN elemento almacenado en la estructura:

$$\bigcap_{ec \in ri} (case1[i][ec] \cup case2[i][j][ec])$$

Complejidad espacial de la estructura:

$O(m^2 \times n)$

Descripción computacional de la estructura:

Array bidimensional de conjuntos de intervalos de números reales, donde los intervalos de cada conjunto son disjuntos dos a dos.

Universitat d'Alacant
Universidad de Alicante

Estructura de datos: Equal [i] [j]

Complejidad espacial de un objeto de la estructura: *O(n)*

Operación necesaria para obtener el valor de UN elemento almacenado en la estructura:

Subset [i] [j] ∩ Subset [j] [i]

Complejidad espacial de la estructura:

$O(m^2 \times n)$

Descripción computacional de la estructura:

Array bidimensional de conjuntos de intervalos de números reales, donde los intervalos de cada conjunto son disjuntos dos a dos.

Estructura de datos: Null [i]

Complejidad espacial de un objeto de la estructura: *O(1)*

Operación necesaria para obtener el valor de UN elemento almacenado en la estructura:

[max_beta, **0.5**]

Complejidad espacial de la estructura: *O(m)*

Descripción computacional de la estructura:

Array de intervalos.

Estructura de datos: S[i]

Complejidad espacial de un objeto de la estructura:

$O(m \times n)$

Operación necesaria para obtener el valor de UN elemento almacenado en la estructura:

Universitat d'Alacant
Universidad de Alicante

$$([0, 0.5] - Null[i]) - \bigcup_{\substack{j=1 \\ j \neq i}}^{m} (Subset[j][i] - Equal[i][j] - Null[j])$$

Complejidad espacial de la estructura: $O(m^2 \times n)$

Descripción computacional de la estructura:

Array de conjuntos de intervalos de números reales, donde los intervalos de cada conjunto son disjuntos dos a dos.

Estructura de datos: Lambda[e][i]

Complejidad espacial de un objeto de la estructura: $O(1)$

Operación necesaria para obtener el valor de UN elemento almacenado en la estructura:

min(c(ec, X), 1-c(ec, X))

Complejidad espacial de la estructura: $O(m^2)$

Descripción computacional de la estructura:

Array bidimensional de números reales.

Universitat d'Alacant
Universidad de Alicante

Estructura de datos: GrExcep[e]

Complejidad espacial de un objeto de la estructura: $O(m)$

Operación necesaria para obtener el valor de UN elemento almacenado en la estructura:

Opción 1: 0

Opción 2:

$$GrExcep[e](\beta) = \begin{cases} present/m & si \quad \beta \in [prev, Lambda[e][i]) \\ GrExcep[e](\beta) & otro \quad caso \end{cases}$$

Complejidad espacial de la estructura: $O(m \times n)$

Descripción computacional de la estructura:

Array bidimensional de pares *(intervalo; número).*

Estructura de datos: Pertain[e][i]

Complejidad espacial de un objeto de la estructura: $O(1)$

Operación necesaria para obtener el valor de UN elemento almacenado en la estructura:

[0; Lambda[e][i])

Complejidad espacial de la estructura:

$O(m \times n)$

Descripción computacional de la estructura:

Array bidimensional de intervalos de números reales.

Estructura de datos: *Outlier*[e]

Complejidad espacial de un objeto de la estructura: $O(m^2 \times n)$

Operación necesaria para obtener el valor de UN elemento almacenado en la estructura:

$$\bigcup_{i=1}^{m} (S[i] \cap Pertain[e][i])$$

Complejidad espacial de la estructura:

$O(m^2 \times n^2)$

Descripción computacional de la estructura:

Universitat d'Alacant
Universidad de Alicante

Array de conjuntos de intervalos de números reales, donde los intervalos de cada conjunto son disjuntos dos a dos.

Estructura de datos: Resul t [e]

Complejidad espacial de un objeto de la estructura: $O(m^2 \times n)$

Operación necesaria para obtener el valor de UN elemento almacenado en la estructura:

$\{ \beta_0, \mu_0: \ \beta_0 \in$ Out l i er [e] ^ $0 \leq \mu_0 <$ Gr Excep[e] (β_0) }

Complejidad espacial de la estructura:

$O(m^2 \times n^2)$

Descripción computacional de la estructura:

Array bidimensional de rectángulos representados por sus proyecciones sobre los ejes.

La *complejidad espacial* de las estructuras de datos influye en la *complejidad temporal* de las operaciones que se realizan durante la ejecución del algoritmo. Por tanto, influye también en la *complejidad temporal* general del mismo para el *caso peor*.

En la **Tabla 6-1** se observa la complejidad espacial de las estructuras de datos y en la **Tabla 6-2** la complejidad temporal de cada uno de los métodos, as omoc í la *complejidad temporal total* del algoritmo *BM*.

Tabla 6-1 Algoritmo *BM*. *Complejidad espacial* de las estructuras de datos

Estructura de datos:	**Complejidad Espacial (*caso peor*)**
Case1[i][ec]	$O(n \times m)$
Case2[i][j][ec]	$O(n \times m^2)$
Subset[i][j]	$O(n \times m^2)$
Equal[i][j]	$O(n \times m^2)$
Null[i]	$O(m)$
S[i]	$O(n \times m^2)$
Lambda[e][i]	$O(m^2)$
GrExcep[e]	$O(n \times m)$
Pertain[e][i]	$O(n \times m)$
***Outlier*[e]**	$O(n^2 \times m^2)$
Result[e]	$O(n^2 \times m^2)$
Complejidad Espacial TOTAL del Algoritmo	*O(máx de las complejidades de las estructuras particulares)=* $O(n^2 \times m^2)$

Tabla 6-2 Algoritmo *BM*. *Complejidad temporal* de los métodos y del algoritmo

Método que calcula:	**Complejidad Temporal (*caso peor*)**
Case1[i][ec]	$O(n \times m \times c)$
Case2[i][j][ec]	$O(n \times m^2 \times c)$
Subset[i][j]	$O(n \times m \times \log(n))$
Equal[i][j]	$O(n \times m^2)$
Null[i]	$O(n \times m)$
S[i]	$O(n \times m^2 \times \log(m))$
Lambda[e][i]	$O(n \times m \times c)$
GrExcep[e]	$O(n \times m \times \log(m))$
Pertain[e][i]	$O(n \times m)$
***Outlier*[e]**	$O(n^2 \times m^2 \times \log(m))$
Result[e]	$O(n^2 \times m^2)$
Complejidad TOTAL del Algoritmo	*O(máx de las complejidades de los métodos particulares)=* $O(n^2 \times m^2 \times \log(m))$

Universitat d'Alacant
Universidad de Alicante

Conclusiones

El aspecto más original del algoritmo propuesto es que permite, de forma no supervisada; establecer la región de valores de los umbrales (parámetros β y μ) que intervienen en el análisis, en la cual cada elemento del *universo* será considerado *outlier*. No obstante, después de analizada la *complejidad temporal* y *espacial* del algoritmo para el *caso peor*, se comprueba que son de orden mayor que la de las anteriores propuestas. Esto no nos sorprende, pues el algoritmo *BM* es más general que los algoritmos vistos anteriormente (*RSBM* y *VPRSM*).

Podemos afirmar que, al ejecutar una vez el algoritmo *BM* para un *universo* de datos dado; se pueden obtener las salidas particulares de los algoritmos previos para cualquier valor de (β, μ). Al determinar para cada elemento del *universo* la región total de valores de los umbrales para los cuales tal elemento es *outlier*, estamos garantizando que posteriormente se pueda hacer un recorrido por todo el *universo* buscando si pares particulares de valores de los umbrales (β, μ) pertenecen a la *región de excepcionalidad* de cualquier elemento del mismo. Por tanto, el uso de este algoritmo es recomendable cuando se necesite obtener un resultado acerca de la condición de *outlier* de los elementos del *universo* para un conjunto de valores de los umbrales.

El análisis hecho en el párrafo anterior, nos permite afirmar también que el resultado que se obtiene tras la ejecución del algoritmo *BM* contiene cualquier resultado particular que pudiese obtenerse a partir de la ejecución de los algoritmos anteriores. Esto constituye la principal ventaja de este algoritmo frente a la desventaja que supone el aumento en la *complejidad temporal* y *espacial* del mismo.

Resulta importante señalar que, a pesar del orden de *complejidad temporal* señalado para el *caso peor*; el algoritmo puede llegar a alcanzar un orden de *complejidad temporal* similar al de los anteriores (*RSBM* y *VPRSM*), *casi lineal* para el *caso mejor* ($\Omega(n \times m^2 \times c)$).

La región general de valores obtenida durante la **Fase 1**, permite establecer una aproximación estocástica a la solución del problema de determinar si un elemento dado es *outlier* dentro de un determinado *universo* de datos. Esto significa que, a partir de ella; se facilita el establecimiento de un criterio probabilístico sobre dicha condición. Por tanto, estamos en disposición de abordar el problema siguiente, que es la propuesta definitiva de un algoritmo que sea capaz de proporcionar la probabilidad que tiene cada elemento del *universo* de ser *excepcional* (*outlier*) en dicho universo.

Fase 2. Estimación de la Probabilidad de cada Elemento del *Universo* de ser *Outlier*

Marco Teórico. Algoritmo *BM*/Probabilístico

En el diseño del nuevo algoritmo, se utilizan algunos conceptos sencillos de Teoría de Probabilidades, tales como: *variable aleatoria*, *vector aleatorio*, *función de densidad de probabilidad*, *función de distribución*, *independencia de variables aleatorias* y *distribución uniforme*, cuyas definiciones pueden encontrarse en cualquier texto básico de esa disciplina. En particular (Fristedt & Gray, 1996) constituye una referencia significativa a tales efectos.

Como ya sabemos, tras la primera fase de este proceso y tras la ejecución del algoritmo *BM* diseñado en ella; se

obtiene como resultado, para cada elemento $x \in U$, la región de valores para los parámetros β y μ, en la cual dicho elemento es un *outlier*. Denotemos por R_x a la región encontrada para un elemento dado, $x \in U$.

Considerando β y μ dos *variables aleatorias*, denotemos por $\varphi(\beta,\mu)$ a la *función de densidad de probabilidad* del *vector aleatorio* (β,μ). Entonces, la *función de distribución* de (β,μ) quedaría:

$$P(\beta \leq i, \mu \leq j) = \int_{-\infty}^{i} \int_{-\infty}^{j} \varphi(\beta,\mu) d\beta d\mu \qquad (18)$$

Luego, la probabilidad P_x que estamos interesados en calcular, o sea la probabilidad de que $x \in U$ sea *outlier* conociendo R_x, puede calcularse a partir de (18) de la siguiente forma:

$$P_x = P((\beta,\mu) \in R_x) = \int_{R_x} \varphi(\beta,\mu) d\beta d\mu \qquad (19)$$

teniendo en cuenta que x es *outlier* para los valores de β y μ que pertenecen a R_x.

Como β y μ son dos *variables aleatorias* independientes**,** entonces:

$$\varphi(\beta,\mu) = f(\beta) * g(\mu)$$

donde, $f(\beta)$ y $g(\mu)$ son las *funciones de densidad de probabilidad* de β y μ respectivamente.

Por tanto:

$$P_x = \int_{R_x} f(\beta) g(\mu) d\beta d\mu \qquad (20)$$

Universitat d'Alacant
Universidad de Alicante

Para calcular P_x sólo tenemos que sustituir en (20) las *funciones de densidad de probabilidad* de los parámetros β y μ y luego calcular la integral resultante.

Sin embargo, en la práctica lo más común es que no se cuente con ninguna información con respecto a la *distribución* de los parámetros β y μ; por ello en este trabajo asumiremos *distribuciones uniformes* para obtener una estimación de P_x. Debe destacarse que, si en algún contexto específico se conociera la *distribución* de los parámetros β y μ y esta fuera diferente a la supuesta; el esfuerzo necesario para adecuar los resultados que se exponen a tales circunstancias sería mínimo, pues solo sería necesario modificar la implementación del cálculo de P_x con las nuevas *funciones de densidad de probabilidad* utilizando algún *método numérico* para el cálculo de la integral, si fuera necesario.

Partiendo del hecho de que hemos supuesto que los parámetros β y μ *distribuyen uniformemente*, no es necesario utilizar *método numérico* alguno, pues la integral resultante resulta muy fácil de calcular. Veamos entonces lo que sucede cuando β y μ *distribuyen de manera uniforme*.

Universidad de Alicante

La *función de densidad de probabilidad* para una variable t que *distribuye uniformemente* en el intervalo [a, b], es la siguiente:

$$f(t) = \begin{cases} \dfrac{1}{b-a} & si\ a < t < b \\ 0 & en\ otro\ caso \end{cases} \qquad (21)$$

Como sabemos que el dominio de valores posibles para β y μ es el siguiente:

$$0 \leq \beta < 0.5 \quad \text{y} \quad 0 \leq \mu \leq 1$$

entonces, bajo la *hipótesis de uniformidad* para β y μ:

$$f(\beta) = \frac{1}{0{,}5 - 0} = 2 \tag{22}$$

$$g(\mu) = \frac{1}{1 - 0} = 1 \tag{23}$$

Sustituyendo estos valores en (20), tenemos:

$$P_x = \int_{R_x} 2 * 1 d\beta d\mu$$

$$P_x = 2 \int_{R_x} d\beta d\mu \tag{24}$$

Y como $\int_{R_x} d\beta d\mu$ es el área de la región R_x, entonces:

$$P_x = 2 * Area(R_x) \tag{25}$$

Este resultado puede interpretarse como:

$$P_x = \frac{Area(R_x)}{0{,}5} \tag{26}$$

que no es más que el cociente entre el área de la *región favorable* (región de valores (β, μ) para los cuales x es *outlier*) y el *área total* (rectángulo que define el dominio de los valores (β, μ) en el plano).

Optimización

El resultado que se acaba de obtener es la base teórica a partir de la cual se obtiene la probabilidad de que cada que $x \in U$ sea un *outlier*. Como puede verse, su implementación computacional es trivial, una vez que se cuenta con la región R_x calculada por el algoritmo *BM*. Por tanto, teniendo en cuenta que este algoritmo constituye el núcleo de la solución del problema planteado; hemos mantenido el

nombre y consideramos el nuevo resultado como una versión que generaliza aún más el resultado alcanzado por el algoritmo *BM*.

Al hacer las primeras validaciones del nuevo algoritmo (*BM/probabilístico*), nos dimos cuenta de que su funcionamiento podría mejorarse estableciendo algún tipo de heurística de manera que el cálculo de la probabilidad, en función de la región R_x, sea más objetivo.

Ilustremos esto con un ejemplo y el mismo permitirá introducir con más claridad nuestra propuesta.

Ejemplo 6-4

Supongamos que el *conjunto de datos* que consideramos como *universo* es el representando en la **Tabla 6-3**. En ella se almacenan datos referidos a países y cada elemento de la misma tiene tres atributos categóricos:

- *País*
- *Idioma oficial*
- *Región geográfica donde se encuentra el país*

Universitat d'Alacant
Universidad de Alicante

Tabla 6-3 Conjunto de datos sobre *países*

País	Idioma oficial	Región geográfica
Angola	Portugués	África
Brasil	Portugués	América del Sur
Cabo Verde	Portugués	África
Timor del Este	Portugués	África
Guinea Ecuatorial	Portugués	África
Guinea Bissau	Portugués	África
Macau	Portugués	Asia
Mozambique	Portugués	África
Portugal	Portugués	Europa
Sao Tomé y Príncipe	Portugués	África

De dicho *conjunto de datos,* supongamos que los *países suramericanos* son los que interesan en el análisis. Por tanto, esto definirá nuestro *concepto.* Consideraremos una *relación de equivalencia* asociada al atributo *idioma oficial* (cada posible idioma define una *clase de equivalencia* particular) y veremos que, al aplicar la misma sobre los elementos del *universo*; la partición que genera posee sólo una *clase de equivalencia* a la que pertenecen todos los elementos de la tabla. Esto es debido a que todos ellos tienen como idioma oficial el portugués. Además, según los datos de la tabla, Brasil es el único país que cumple el *concepto*.

La región de valores *beta-miu* hallada para el elemento Brasil es la que se ilustra en la **Figura 6-17**.

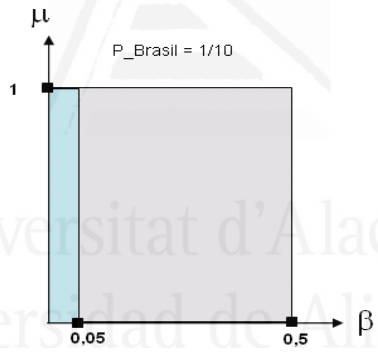


Figura 6-17 Región de valores beta-miu asociada al país Brasil

Si determinamos la probabilidad asociada a este elemento, a partir de (26) la misma sería P_{Brasil} = 0,05/0,5 =1/10 = 0,1. Si analizamos este resultado, nos damos cuenta que no expresa con exactitud la realidad en relación a lo que representa Brasil en dicho contexto de datos; pues es el único país suramericano cuyo idioma oficial es el portugués y, por tal motivo, la lógica indica que debería ser uno de los más fuertes candidatos a ser *outlier*.

Esto puede explicarse a partir del hecho de que tal situación puede provocarse al considerar márgenes de *error de clasificación* relativamente altos (valores de β próximos a 0.5). Reflexionando en torno a este fenómeno, consideramos que tales situaciones se pueden evitar permitiendo que en el algoritmo se establezca una cota al margen de *error de clasificación* permitido. Sea ésta, *t.* Dicha cota supone una forma de flexibilizar el máximo valor establecido para el valor β, según la interpretación de *VPRSM:* es decir, *t* no tiene necesariamente que ser 0,5. Tal flexibilidad supone que, $0 \leq \beta \leq t$, con $0 < t \leq 0{,}5$. Al establecimiento de un valor específico para *t*, le llamaremos *corte*, pues nemotécnicamente expresa lo que en realidad sucede, o sea: Reducir a partir de dicho valor el área de las regiones —*favorable* y *total*— que intervienen en el cálculo de la probabilidad.

Lo que acabamos de describir se ilustra en la **Figura 6-18,** donde se establecido un valor *t=0,1* para la situación planteada en el **Ejemplo 6-4**.

Universitat d'Alacant
Universidad de Alicante

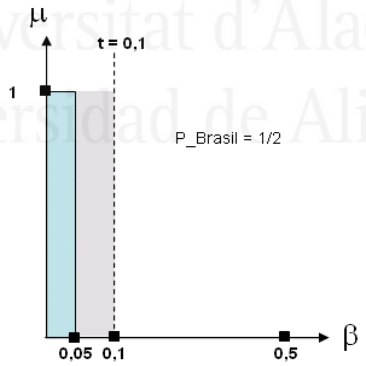


Figura 6-18 Aplicación de un *corte*, a partir del valor t=0,1, al área total de la región beta-miu

En este caso, el corte implica reducir el área total, con lo cual, aumenta el valor de la probabilidad. Esto hace que la misma se corresponde más objetivamente con lo que en

realidad representa Brasil dentro del *conjunto de datos*: es el único país suramericano cuyo idioma oficial es el portugués y por tanto debe ser uno de los más fuertes candidatos a ser *outlier*. En este caso, P_{Brasil} = 0,05/0,1 = 5/10 = 0,5.

La problemática vista en el **Ejemplo 6-4** puede generalizarse a partir de las dos situaciones siguientes:

Situación No. 1

Supongamos que tenemos un elemento x que es poco contradictorio, pero lo es para todos los valores de β.

En tal caso, el establecimiento del corte hace que:

a) disminuya el área para la cual dicho elemento se considera *outlier*.
b) reduce el área total tenida en cuenta en el cálculo de la probabilidad.

Por tanto, esto hace que el valor de la probabilidad calculada se mantenga. La **Figura 6-19** ilustra esta situación.

Universitat d'Alacant
Universidad de Alicante

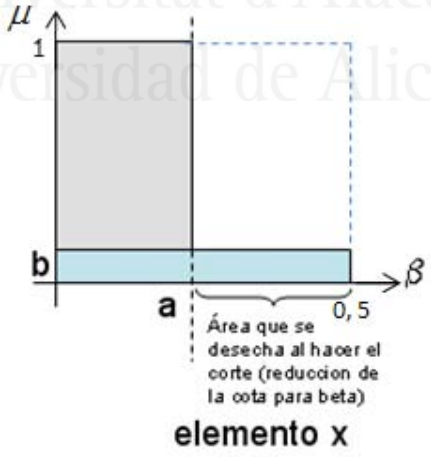


Figura 6-19 Efectos del *corte* al ser aplicado a la región que caracteriza a un elemento poco contradictorio para muchos valores de β

Situación No. 2

Supongamos que tenemos un elemento *y*, tal que el área para la cual dicho elemento es *outlier* es igual —en superficie— a la del elemento *x* al cual se hizo referencia en la descripción de la situación anterior. Para el elemento *y,* ocurre lo contrario de lo que ocurría para el elemento *x*, o sea, *y* es *muy excepcional* pero para pocos valores de β (hecho que es muy frecuente debido a la naturaleza de los *outliers*). Entonces, si mantenemos un corte similar al definido para la situación anterior, el establecimiento del mismo provocaría que el área total disminuyera; pero, como se mantiene el área para la cual el elemento es *outlier*, aumentará la probabilidad que se calcula para el elemento *y* (o sea, una situación similar a la descrita en el **Ejemplo 6-4**). La **Figura 6-20** ilustra esta situación.

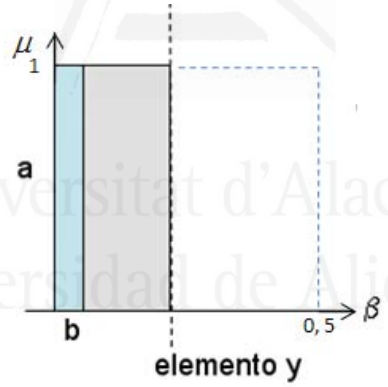


Figura 6-20 Efectos del *corte* al ser aplicado a la región que caracteriza a un elemento muy contradictorio para pocos valores de β

Para garantizar la equivalencia en la forma en que se calcula la probabilidad, se establece el corte en todos los casos; pero realmente, es verdaderamente útil en la situación 2, pues en tal caso permite ajustar con más objetividad el cálculo de la probabilidad en correspondencia

con lo que ocurre en la vida real. No obstante, en los casos como el descrito en la situación 1, el establecimiento del corte no implica un cambio significativo en el cálculo de la probabilidad y por tanto no puede decirse que tenga algún efecto perjudicial.

La situaciones mostradas, tanto en la **Figura 6-19** como en la **Figura 6-20,** puede que sean la más sencillas y evidentes donde la problemática que pretendemos explicar se cumpla. La esencia de este planteamiento viene dada por la estructura que se ha asumido para el área de la *región de excepcionalidad* de los elementos *x* e *y*. En este caso, hemos supuesto que la misma coincide con la de un solo rectángulo. Esto se ha hecho de manera intencional para no empañar la claridad del fenómeno que se desea explicar, pero debe tenerse en cuenta que, en la vida real, la estructura de dicha región puede ser mucho más compleja y estar constituida por más de un rectángulo. No obstante, la esencia del fenómeno es la misma. Considerando que esto es algo que puede verse de forma bastante intuitiva, no creemos que sea relevante una demostración rigurosa de ello para todos los posibles casos.

Universitat d'Alacant
Universidad de Alicante

Veamos las modificaciones que provoca el establecimiento del corte en el marco teórico del problema que estamos abordando.

Si suponemos que el parámetro *t* es la cota superior para el valor de β en lugar de 0,5 y se sustituye dicho valor en (22), entonces la nueva fórmula para determinar la probabilidad de que el elemento *x* sea *outlier* en el *universo* dado sería la siguiente:

$$P_x = \frac{Area\ (R'_x)}{t} \tag{27}$$

donde, $R'_x = \{puntos\,(\beta,\mu)\, tal\ que\ (\beta,\mu) \in R_x \wedge \beta \le t\}$

A continuación se exponen los aspectos esenciales acerca de la implementación computacional del algoritmo.

Implementación Computacional. Algoritmo *BM*/Probabilístico

El algoritmo a partir del cual se resuelve el problema de establecer la probabilidad que tiene cada elemento de un *universo* de datos dado de ser *outlier*, al que hemos llamado algoritmo *BM*/Probabilístico; presenta las siguientes características:

Los siguientes elementos constituyen las entradas del mismo:

- El *universo U* (*conjunto de datos*).
- El *concepto C.*
- $\Re = \{r_1, r_2, ..., r_m\}$. Los criterios que distinguen a las *relaciones de equivalencia* tenidas en cuenta en el análisis.
- El valor de *t.*

Universitat d'Alacant
Universidad de Alicante

Debemos destacar que el parámetro *t* es un parámetro opcional que el usuario especifica si desea o no. Si no lo especifica, el algoritmo asume por defecto el valor *t*=0,2. Esta decisión ha sido tomada de forma empírica a partir de lo observado en las pruebas que progresivamente se fueron haciendo para verificar la efectividad del algoritmo *BM* y tomando en consideración la estructura de las regiones que el mismo proporcionaba.

Atendiendo a lo anterior, podemos considerar como un problema abierto la determinación automática del valor de *t* adecuado, de manera tal que se garantice la mejorara de la precisión de los cálculos.

Con respecto a la salida del algoritmo *BM*/Probabilístico, ya hemos comentado que la misma es una estimación de la probabilidad para cada elemento de *U* en cuanto a su condición de *outlier* en dicho *universo*.

Es importante resaltar que el peso teórico, en este caso, recae fundamentalmente sobre el cálculo de la *región de excepcionalidad* de cada elemento del *universo,* de lo cual se encarga el algoritmo *BM*. Una vez que *BM*/Probabilístico recibe como entrada dicha región, sólo agrega a su funcionalidad un método principal donde se hace el cálculo de la probabilidad en la forma expresada en (27). Una descripción en *pseudo*-código de dicho método, se ilustra a continuación:

```
(1) result = Result[e]  // Salida del algoritmo BM para: U-Universo,
                        // C-Concepto, R-Relaciones de equivalencia
(2) for e in U
(3)     result = ptos (β, μ) de la región Result[e] tal que β<t
(4)     p[e] = (Área de r) / t
```

Método 12 Método principal del algoritmo *BM*/Probabilístico

Analicemos a continuación la *complejidad temporal* del algoritmo. Ésta se ve afectada por la *complejidad temporal* de la **Fase 1**, donde se ejecuta el algoritmo *BM*, pues, como hemos señalado, la salida que se obtiene tras la ejecución de este algoritmo, es precisamente la entrada del algoritmo *BM*/Probabilístico que se ejecuta en la **Fase 2**. Es decir, antes de ejecutar el algoritmo *BM*/Probabilístico es imprescindible haber ejecutado el algoritmo *BM*.

Por tanto, para establecer la *complejidad temporal* del algoritmo que nos ocupa, podemos estructurar el análisis de la siguiente forma:

Coste de la **Fase 1**:

Complejidad temporal del algoritmo *BM: O($n^2 \times m^2 \times log(m)$)*

Coste de la **Fase 2**:

(cardinalidad del *conjunto de datos*) × (cota para la cantidad de rectángulos que pueden pertenecer a una región *beta-miu*) = *(n) × (n × m^2) = O($n^2 \times m^2$)*

Complejidad temporal TOTAL del algoritmo *BM*/Probabilístico para el *caso peor*:

O(máx (Costo de la **Fase 1**, Costo de la **Fase 2***))*

= O($n^2 \times m^2 \times log(m)$)

Validación de los resultados

Las pruebas realizadas, al igual que en el caso de los algoritmos *RSBM* y *VPRSM*, tuvieron como objetivo fundamental, validar aspectos relacionados con el *tiempo de ejecución* del algoritmo y la *calidad de la detección* del mismo.

Universitat d'Alacant
Universidad de Alicante

Prueba 6-1

Objetivo de la prueba: Validar *tiempo de ejecución* del algoritmo *BM/Probabilístico* —comparando el mismo con el algoritmo *VPRSM*— al trabajar con *conjuntos de datos* de *gran tamaño* y *alta dimensionalidad.*

Conjunto de datos utilizado: *Conjunto de datos sintético*

Descripción del *conjunto de datos*: *conjunto de datos aleatorio generado automáticamente* a partir del uso de técnicas estadísticas.

Tipo del *conjunto de datos*: *multivariado.*

Tipo de los atributos: *categóricos* y *continuos.*

Cantidad de filas: 500.000

Cantidad de atributos (columnas): 100

Dispositivo de cálculo utilizado en la prueba

Procesador: *Intel(R) Core(TM)2 Quad CPU Q6600 @ 2.40Ghz 2.39Ghz*

Memoria: *3.25GB.* Sistema Operativo: *Windows 7 Ultimate*

Descripción de la prueba

Teniendo en cuenta que la forma en que se generó el *conjunto de datos sintético* fue abordada de manera explícita al describir la **Prueba 5-2** del acápite **Validación de los resultados** del **Capítulo 5**, creemos que no resulta necesario retomar nuevamente dicha explicación. En tal sentido, remitimos al lector al citado apartado del referido capítulo.

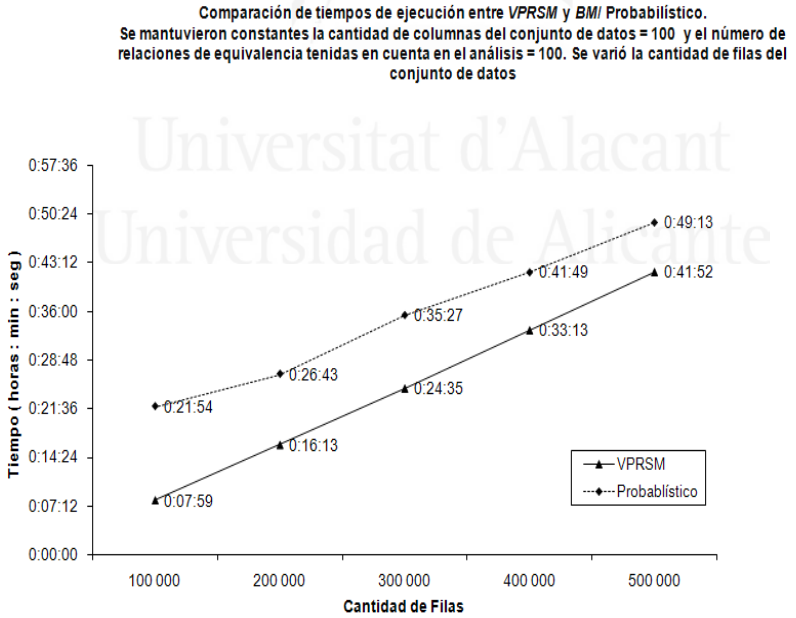


Figura 6-21 Comparación de tiempos de ejecución entre los algoritmos *VPRSM* y *BM*/Probabilístico

La **Figura 6-21** muestra los resultados obtenidos —en cuanto a *tiempo de ejecución*— tanto por el algoritmo *BM*/Probabilístico como por el algoritmo *VPRSM*.

Interpretación de los resultados

Las curvas muestran que ambos algoritmos se comportan de manera similar —en cuanto a *tiempo de ejecución*— y que son computacionalmente eficientes al ejecutarse sobre un *conjunto de datos* que puede considerarse de *gran tamaño* y *alta dimensionalidad*. Se aprecia la linealidad en los *tiempos de ejecución* alcanzados por los mismos para dicho *conjunto de datos*.

Consideramos importante destacar que el resultado alcanzado por *BM*/Probabilístico, al ser ejecutado sobre este *conjunto de datos*, pone de manifiesto que, a pesar de que el *orden de complejidad temporal* del mismo es *cuadrático* para el caso peor, este puede llegar a alcanzar un orden de *complejidad temporal casi lineal* (o sea similar al del *VPRSM*) al ejecutarse sobre *conjuntos de datos* con las características antes mencionadas.

Universitat d'Alacant
Universidad de Alicante

Prueba 6-2

Objetivo de la prueba: Validar *calidad de la detección* del algoritmo *BM/Probabilístico*

Conjunto de datos utilizado: *Adult Data Set*

Descripción del *conjunto de datos*: Contiene datos extraídos del *Census Bureau Database of USA* (*CENSUS*, 2009)

Fuente: *UCI Machine Learning Repository*. *Center for Machine Learning and Intelligent Systems*. Universidad de California, Irvine (*UCI*, 2009)

Tipo del *conjunto de datos*: *multivariado.*

Tipo de los atributos: *categóricos* y *enteros.*

Cantidad de filas: 48.842

Cantidad de atributos (columnas): 14

Dispositivo de cálculo utilizado en la prueba

INTEL Pentium 4, CPU 1.5 Ghz, 256 MB de RAM. Plataforma: Windows XP SP3

Descripción de la prueba

La descripción de esta prueba es similar a la dada en la **Prueba 5-3** del acápite **Validación de los resultados** del **Capítulo 5**. En esta ocasión se mide calidad de la detección del algoritmo *BM*/Probabilístico. Se utilizó el mismo *conjunto de datos*, el mismo *concepto* y las mismas *relaciones de equivalencia* que en la prueba antes mencionada. En este caso se introdujeron en el *conjunto de datos* 14 *outliers* que representaban *niños menores de 10 años* con valores inapropiados en los atributos que intervenían en el análisis.

La **Figura 6-22** muestra los resultados alcanzados. La estrategia seguida en la realización de la prueba fue la siguiente: para diferentes valores de *k*, se analizaron los *k* elementos del *conjunto de datos* que el algoritmo detectó con mayor probabilidad de ser *outliers* y se determinó, de ellos, cuántos pertenecían al conjunto de *outliers* que fueron introducidos intencionalmente en el *conjunto de datos*.

Los valores de *k* tenidos en consideración fueron: 5, 10, 15, 20, 25, 30 y 35. Se escogieron, además, dos valores representativos para el corte, *t = 0,2 y 0,4*, con el objetivo de mostrar una comparación de los resultados alcanzados para dos cortes diferentes.

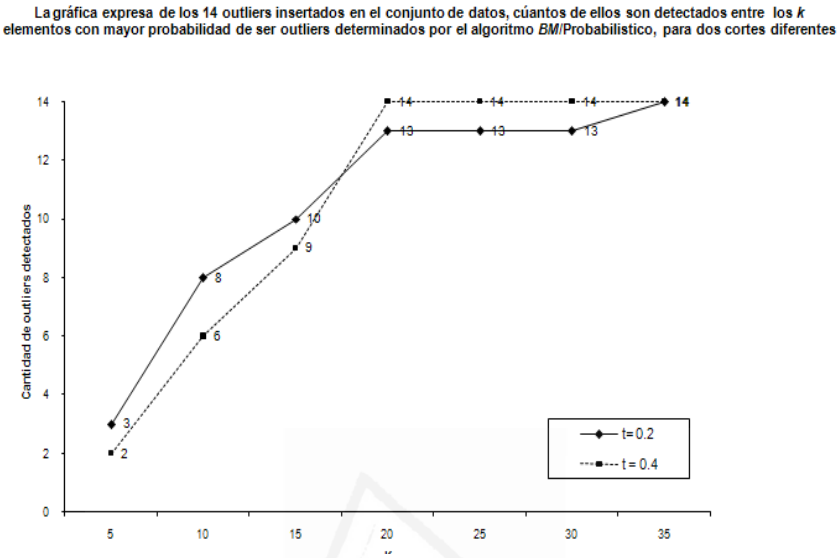


Figura 6-22 Resultados de la Prueba 6-2 realizada al algoritmo *BM*/Probabilístico

Interpretación de los resultados

Los resultados alcanzados demuestran la efectividad del algoritmo en relación al estimado probabilístico que calcula para los elementos del *conjunto de datos*. En tal sentido, debe señalarse que entre los elementos de mayor probabilidad, para cualquier valor de *k* tenido en cuenta, siempre se encontraron los *outliers* que fueron introducidos en el *conjunto de datos*.

A modo de ejemplo, se comentan algunos casos que pueden ser representativos:

- Para un *corte* establecido por *t=0,2:*
  - Al ser analizados los 5 elementos con mayor probabilidad, entre ellos aparecen los 3 más contradictorios que se introdujeron en el *conjunto de datos*.

- Al ser analizados los 20 primeros, entre ellos se encontraban 13 de los 14 introducidos.
- Al aumentar la muestra a los 35 primeros, en este grupo se detectaron todos los introducidos.

- Para un corte establecido por t=0.4:
  - Al ser analizados los 5 elementos con mayor probabilidad, entre ellos aparecen 2 de los más contradictorios que se introdujeron en el *conjunto de datos*.
  - Al analizar los 20 primeros, ya en este grupo se detectaron todos los *outliers* introducidos.

Otro hecho a destacar es el siguiente: al analizar los elementos que fueron detectados entre los de mayor probabilidad y que no eran exactamente alguno de los 14 introducidos, pudo observarse que el alto valor de su probabilidad se debía a que eran contradictorios con respecto a la información brindada en el *conjunto de datos*, pero el análisis permitió comprobar que semánticamente no eran *outliers*. No obstante, cabe señalar que los elementos con estas características eran sólo unos pocos.

Puede verse que los resultados para ambos valores de t, son bastante similares. Sin embargo, si no se tiene un amplio dominio del conocimiento brindado por el *conjunto de datos*, se recomienda utilizar el *corte* haciendo t = 0,2 pues empíricamente se ha podido comprobar que es un valor que permite un grado aceptable de *desclasificación*. Con esto se evita el riesgo de que, al permitir una *desclasificación* muy alta; los elementos *poco contradictorios para muchos valores de* β, se clasifiquen de *outliers*.

Prueba 6-3

Objetivo de la prueba: Validar *calidad de la detección* del algoritmo *BM/Probabilístico*

Conjunto de datos utilizado: *Arrhythmia Data Set*

Descripción del *conjunto de datos*: Datos de *pacientes* con problemas cardiovasculares

Fuente: *UCI Machine Learning Data Repository*

Tipo del *conjunto de datos*: *multivariado*

Tipo de los atributos: *reales, enteros* y *categóricos.*

Cantidad de filas: 452

Cantidad de atributos (columnas): 279

Dispositivo de cálculo utilizado en la prueba

INTEL(R) Core(TM) 2 Duo, CPU T5450 @ 1.66 Ghz (2 CPUs), 2046 MB de RAM. Plataforma: Windows Vista

Sobre este *conjunto de datos se* hicieron dos pruebas diferentes, por tanto, la descripción de cada una de ellas se identificará de manera diferente también.

Universidad de Alicante

Descripción de la Prueba 6-3.1

La descripción de esta prueba es similar a la dada en la **Prueba 5-4.1** del acápite **Validación de los resultados** del **Capítulo 5**. En esta ocasión se mide calidad de la detección del algoritmo *BM*/Probabilístico. Se utilizó el mismo *conjunto de datos*, el mismo *concepto* y las mismas *relaciones de equivalencia* que en la prueba antes mencionada.

Vale recordar que las *personas* que clasificaban dentro del *concepto* se consideraban de *bajo peso*. Teniendo en cuenta los valores habituales de los atributos implicados en las

relaciones de equivalencia descritas para dichas personas, se introdujeron, intencionalmente, en el *conjunto de datos* 12 *outliers* que representaban *personas de bajo peso* con valores contradictorios para varios de dichos atributos.

La **Figura 6-23** muestra los resultados alcanzados en esta ocasión. La estrategia seguida en la realización de la prueba fue la misma que la seguida en la **Prueba 6-2**, es decir, se consideraron diferentes valores de *k*, se analizaron los *k* elementos del *conjunto de datos* que el algoritmo detectó con mayor probabilidad de ser *outliers* y se determinó, de ellos, cuántos pertenecían al conjunto de *outliers* que fueron introducidos en el *conjunto de datos*. Los valores de *k* tenidos en consideración, en esta ocasión, fueron: 5, 10, 15 y 20. En este caso, se asumió el *corte por defecto*, o sea, *t = 0,2.*

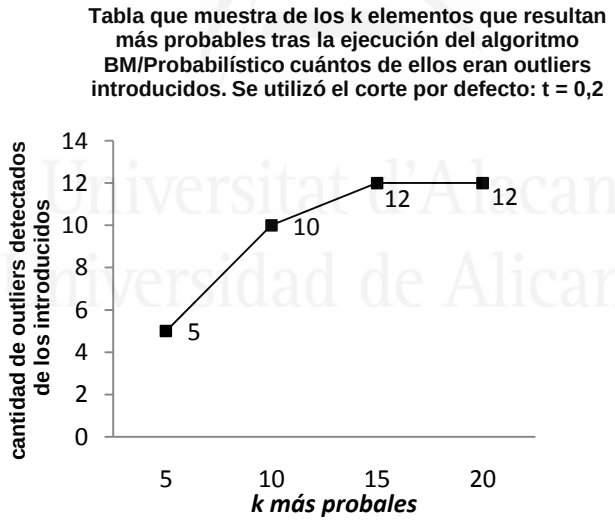


Figura 6-23 Resultados de la Prueba 6-3.1 realizada al algoritmo *BM*/Probabilístico

La **Tabla 6-4** muestra el valor de la probabilidad determinado por el algoritmo para los *outliers* que fueron introducidos en el *conjunto de datos*. Se resaltan los

elementos más contradictorios y se observa que son los de mayor probabilidad.

Tabla 6-4 Probabilidad para los *outliers* introducidos. Prueba 6-3.1: *Arrhythmia DS*

Id	*weight (Kg)*	*heart rate*	*number of intrinsic deflections*	*height (cm)*	*probabilidad de ser outlier*
1	15	***60***	17	***180***	0.61884
2	31	93	***68***	***178***	0.7557252
3	39	***50***	***82***	130	0.6151009
4	10	***53***	16	***188***	0.61884
5	19	***45***	***90***	***190***	**0.8779342**
6	20	***48***	***86***	***183***	**0.8779342**
7	25	***50***	***71***	***180***	**0.8779342**
8	29	***55***	***75***	***179***	**0.8779342**
9	33	90	***60***	***176***	0.7557252
10	40	***61***	20	***186***	0.61884
11	26	***50***	***99***	***180***	**0.8779342**
12	38	92	***100***	***178***	0.7557252

Universitat d'Alacant
Universidad de Alicante

Interpretación de los resultados

A continuación se comentan los resultados alcanzados para cada uno de los valores de *k* tenidos en consideración:

k=5

Los 5 elementos con mayor probabilidad de ser *outliers,* resultaron ser los 5 elementos más contradictorios introducidos en el conjunto de datos. Lo eran por todos los atributos tenidos en cuenta en el análisis.

k=10

Los 10 con mayor probabilidad eran *outliers* introducidos en el *conjunto de datos.*

k=15, 20

Entre los 15 con mayor probabilidad, ya estaban los 12 que fueron introducidos en el *conjunto de datos.*

Esta prueba es un caso concreto de aplicación del algoritmo a problemas médicos del mundo real.

Descripción de la Prueba 6-3.2

La descripción de esta prueba es similar a la dada en la **Prueba 5-4.2** del acápite **Validación de los resultados** del **Capítulo 5**. En esta ocasión se mide calidad de la detección del algoritmo *BM*/Probabilístico. Se utilizó el mismo *conjunto de datos*, el mismo *concepto* y las mismas *relaciones de equivalencia* que en la prueba antes mencionada. Debe recordarse que en esta ocasión se insertaron intencionalmente 15 *outliers* en el conjunto de datos bajo los mismos criterios descritos en la **Prueba 5-4.2**.

La estrategia seguida para la realización de la prueba fue similar a la descrita en la **Prueba 6-3.1**. La **Figura 6-24** muestra los resultados alcanzados en esta ocasión.

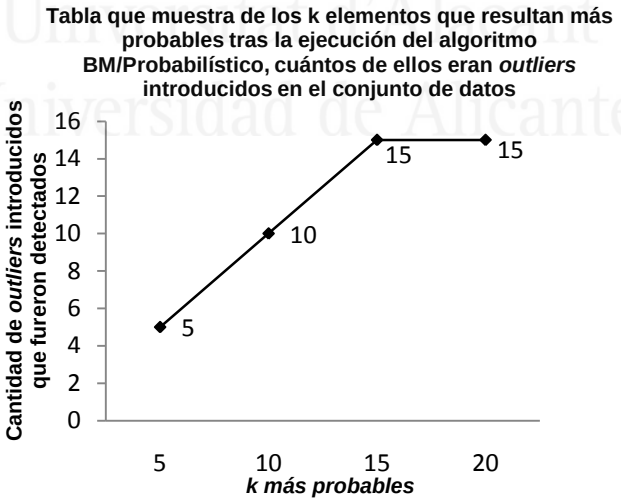


Figura 6-24 Resultados de la Prueba 6-3.2 realizada al algoritmo *BM*/Probabilístico

La **Tabla 6-5** muestra el valor de la probabilidad determinado por el algoritmo para los *outliers* que fueron introducidos en el *conjunto de datos*. Se resaltan los elementos más contradictorios y se observa que son los de mayor probabilidad.

Tabla 6-5 Probabilidad para los *outliers* introducidos. Prueba 6-3.2: *Arrhythmia DS*

Id	*Q-T interval*	*T interval*	*vector angle of*	*number of intrinsic*	*probabilidad de ser outlier*
1	***240***	***300***	***-50***	***95***	**0.8299794**
2	***271***	***297***	***-10***	***77***	**0.8299794**
3	***280***	***299***	***-23***	***82***	**0.8299794**
4	***235***	***360***	***-34***	***90***	**0.8299794**
5	***330***	***320***	***-5***	***91***	**0.8299794**
6	***240***	***310***	5	***37***	0.4635234
7	***286***	***298***	22	***60***	0.4635234
8	***435***	***296***	***-80***	25	0.7497731
9	***455***	135	***-49***	***85***	0.7266043
10	338	***200***	***-46***	***91***	0.5918933
11	390	***201***	***-100***	28	0.3467498
12	400	***240***	***-90***	15	0.3467498
13	***450***	***280***	50	10	0.4447436
14	***480***	***291***	33	30	0.4447436
15	***500***	***233***	88	12	0.4447436

Universitat d'Alacant
Universidad de Alicante

Interpretación de los resultados

A continuación se comentan los resultados alcanzados para cada uno de los valores de *k* tenidos en consideración:

k=5

Los 5 elementos con mayor probabilidad de ser *outliers,* resultaron ser los 5 elementos más contradictorios

introducidos en el *conjunto de datos*. Lo eran por todos los atributos tenidos en cuenta en el análisis.

k=10

Los 10 con mayor probabilidad, resultaron ser 10 *outliers* introducidos en el *conjunto de datos.* Los 5 más contradictorios y otros *5* que lo eran por dos *relaciones de equivalencia*

k=15, 20

Los 15 con mayor probabilidad, coincidían con los 15 *outliers* que fueron introducidos en el *conjunto de datos.*

Universitat d'Alacant
Universidad de Alicante

Capítulo 7

Comparación de Métodos

En los capítulos 4, 5 y 6 se han expuesto paulatinamente los principales resultados de la investigación, junto con su validación formal a partir de los algoritmos particulares que se diseñaron para comprobar la viabilidad computacional de cada uno de los métodos propuestos. En especial, la validación realizada en el capítulo 6 demuestra la viabilidad de la solución propuesta, así como la idoneidad de la hipótesis de partida. En este capítulo se contextualiza dicha propuesta mediante su comparación con otros métodos existentes.

De acuerdo con lo expuesto, el resto del capítulo se estructura de la siguiente forma: seguidamente se exponen algunas consideraciones generales sobre las comparaciones entre métodos de detección de *outliers; p*osteriormente se establece una comparación entre los distintos algoritmos propuestos en esta investigación como paso previo a una comparación de los mismos con algunos métodos representativos de detección de *outliers*. En cada caso, los

aspectos esenciales que han sido expuestos se resumen en cuadros comparativos.

Consideraciones Generales

Entre los aspectos esenciales en los que se basan las comparaciones de métodos de detección de *outliers* pueden señalarse los siguientes (Otey *et al.*, 2005c):

- Parámetros que recibe el método.
- Tipo de los valores que pueden tomar los atributos asociados a los elementos del *conjunto de datos*.
- En qué medida el tamaño o la dimensionalidad del *conjunto de datos* puede afectar a la efectividad del método.
- Calidad de la detección.
- *Complejidad espacial* y *temporal* de los algoritmos que validan la viabilidad computacional del método.
- Naturaleza y estructura de los datos sobre los cuales se aplican dichos algoritmos.

Universitat d'Alacant
Universidad de Alicante

A partir de la información recopilada en el estudio del *estado del arte*, pudo comprobarse que la cantidad de técnicas (bases teóricas) que sirven de soporte a los métodos de detección de *outliers* es considerablemente grande. Por tanto, resulta impracticable una comparación general de dichos métodos (Ben-Gal, 2005). Una buena parte de ellos se basan en un conjunto disjunto de aspectos y se aplican a tipos de datos diferentes dentro de entornos diferentes. En otros casos, los nuevos métodos insertan en sus propuestas otras técnicas ya conocidas (Ren *et al.*, 2004) que constituyen, en sí mismos métodos de detección de *outliers* específicos. Como ejemplo de tales técnicas, pueden señalarse las de *clustering* y las basadas en *distancias*. En tal caso, pueden citarse varios enfoques

basados en *densidades* (Reif *et al.*, 2008), (Tao & Pi, 2009) que incorporan al mecanismo de detección de *outliers* criterios de *distancia* y técnicas de *clustering*.

En la bibliografía revisada en el estudio del *estado del arte* sólo se encontraron comparaciones parciales de métodos (Penny & Jolliffe, 2001), (Simon *et al.*, 2002), (Shekhar *et al.*, 2003), (Hodge & Austin, 2004), (Otey *et al.*, 2005c), en los que se comparan grupos pequeños de técnicas o métodos atendiendo a varios de los aspectos antes mencionados.

Algoritmos de Detección de *Outliers* basados en la Teoría de *Rough Sets*

La base teórica de los métodos propuestos en este trabajo es la Teoría de *RS*. Por tanto, el primer aspecto que los distingue del resto de los métodos recogidos en la literatura es, precisamente, su base teórica. No existen antecedentes de otros con un planteamiento similar. No caen dentro de ninguna de las categorías que la bibliografía recoge para clasificar los métodos de detección según el principio en el cual se basan. Es el primer resultado práctico de la aplicación de la Teoría de *RS* al problema de la detección de *outliers*. Esto los hace originales en cuanto a su concepción.

Los métodos expuestos tienen solidez y simplicidad en su base matemática. La misma queda establecida por la teoría de las *relaciones de equivalencia* que es la esencia de la Teoría de *RS*. La aplicación de estos métodos resulta especialmente adecuada en contextos donde sea necesario resolver problemas de clasificación.

Los métodos propuestos son aplicables a datos en forma tabular y sus atributos deben ser monovaluados. Si los

atributos no cumplen esta condición, entrarían en contradicción con la esencia del método, pues no existiría la posibilidad de establecer, a partir de ellos, *relaciones de equivalencia*. Esto delimita el ámbito de aplicación del problema.

A partir de la concepción de cada uno de los métodos propuestos, se han concebido cuatro algoritmos fundamentales que validan la viabilidad computacional de los mismos. Vale aclarar que, en dos de ellos —*BM* y *BM*/Probabilístico—, uno necesita de la salida del otro para garantizar su ejecución. No obstante, cada uno, individualmente, juega un papel importante en la solución de problemas concretos.

Cada algoritmo tiene sus particularidades en función del marco teórico sobre el cual se diseñó. Por tanto, consideramos necesario establecer una comparación entre ellos antes de compararlos con otros enfoques existentes.

Algoritmo *RSBM*

El algoritmo *RSBM* es el algoritmo basado en el modelo básico de *RS*. Sus principales ventajas y limitaciones se sintetizan en los siguientes apartados.

Principales Ventajas

- Valida la viabilidad computacional del método de detección de *outliers* basado en el modelo básico de *RS*.
- El algoritmo es computacionalmente factible para *conjuntos de datos* grandes y de alta dimensionalidad.
- Es un algoritmo fácil de usar para resolver el problema de la detección de *outliers* haciendo uso de la Teoría de *RS*. Por tanto, se podría usar como primera aproximación a una solución de dicho problema donde

se requiere poco esfuerzo para alcanzar la misma; debido a la simplicidad del método y a la *complejidad temporal* del algoritmo: $O(n \times m^2)$, donde: *n es* $|U|$ y *m es* $|\ \Re\ |$, para el *caso peor.*

- Con respecto a este orden de *complejidad temporal* se debe puntualizar que, teniendo en cuenta que el número de *relaciones de equivalencia* que intervienen en el análisis, en la inmensa mayoría de los casos, no es muy grande en relación al número de filas de la tabla; puede decirse que la dependencia cuadrática del tiempo de ejecución con respecto a la cantidad de *relaciones de equivalencia* tenidas en cuenta en el análisis, no afecta en gran medida al tiempo de ejecución del algoritmo. Esta dependencia cuadrática es *casi lineal* para valores pequeños de m ($m \leq 40$) que son los usuales.
- Su *complejidad espacial* es $O(n \times m)$

Principales Limitaciones

- Determinista en lo referido a *clasificación*. La concepción del método en el cual se basa el algoritmo se fundamenta en el modelo básico de *RS,* del cual se critica su incapacidad para modelar información *incierta*. La *clasificación* con un *grado controlado de incertidumbre* o un posible *error de clasificación*, está fuera del alcance de este modelo. Sin embargo, como se ha señalado; en la práctica, poder admitir algún nivel de *incertidumbre* en el proceso de *clasificación* puede llevar a una comprensión más profunda y a una mejor utilización de las propiedades de los datos analizados. La definición estándar de inclusión de conjuntos tenida en cuenta en el modelo básico es demasiado rigurosa para modelar una inclusión de conjuntos *casi completa*.

Universidad de Alicante

- Se requiere del establecimiento del umbral μ, por parte del usuario, para la ejecución del algoritmo.

Algoritmo *VPRSM*

El algoritmo *VPRSM* se basa en el modelo de *RS* de Precisión Variable. Sus principales ventajas y limitaciones se sintetizan en los siguientes apartados.

Principales Ventajas

- Valida la viabilidad computacional del método de detección de *outliers* basado en *VPRSM*.
- El algoritmo subsana el carácter determinista de *RSBM* mediante la relajación del *concepto* estándar de *inclusión de conjuntos*, usando ciertos umbrales definidos por el usuario.
- El algoritmo es computacionalmente eficiente, i.e. mantiene la *complejidad espacial* y *temporal* del algoritmo *RSBM,* al tiempo que permite ampliar la aplicación del método original a contextos en los que sea necesaria una *clasificación* con un cierto *grado de incertidumbre*.

Universitat d'Alacant
Universidad de Alicante

Principales Limitaciones

- El usuario debe definir, además del *umbral de excepcionalidad*, el *error de clasificación* permitido. Una selección no adecuada de este *error* puede conllevar a resultados no satisfactorios, por tanto, es necesario tener conocimiento sobre aspectos específicos del *conjunto de datos* a partir de los cuales se haga una selección adecuada del mismo.

Consideraciones comunes para *RSBM* y *VPRSM*

En ambos algoritmos —*RSBM* y *VPRSM*— se señala como limitación el hecho de que en ellos se requiere del establecimiento de los umbrales que intervienen en la ejecución de los mismos: *umbral de excepcionalidad* en *RSBM* y en *VPRSM*, además, se debe proporcionar el *error de clasificación* admitido. Esta situación conduce a la siguiente reflexión: aunque en el análisis teórico de la *complejidad temporal* de ambos algoritmos se determina que la misma es *casi lineal* con respecto a la cardinalidad del *universo*; no se puede decir lo mismo del proceso global de detección de *outliers* a partir de los mismos. Tal proceso requiere, para cada caso, de un análisis del contexto, previo y posterior a la ejecución de los algoritmos. En el análisis previo, se establecen las condiciones bajo las cuales se desea ejecutar el algoritmo, mientras que en el análisis posterior, se analiza el conjunto de *outliers* detectado y, dependiendo del resultado de dicho análisis; puede que sea necesario establecer nuevas condiciones bajo las cuales se pueda obtener un nuevo conjunto de *outliers* que se adecúe con más objetividad a los intereses del análisis. Este proceso es necesario repetirlo, cada vez que se ejecuten los algoritmos para valores concretos de los umbrales antes mencionados y esto atenta, por tanto, contra el tiempo de ejecución total requerido por el proceso de detección.

A partir de lo anteriormente dicho, se puede concluir que el uso de los dos algoritmos resulta especialmente factible cuando el proceso que se acaba de describir se repite pocas veces como consecuencia de una buena elección de los umbrales.

Algoritmo *BM*

El algoritmo BM es el responsable de establecer la región de valores de los umbrales β y μ en la cual cada elemento del *universo* es *outlier*. Sus principales ventajas y limitaciones se detallan a continuación.

Principales Ventajas

- Valida la viabilidad computacional del método a partir del cual se obtiene la *región de excepcionalidad* para cada elemento del universo.
- Mantiene el carácter no determinista del algoritmo *VPRSM*.
- La funcionalidad de este algoritmo es más general que la de *RSBM* y *VPRSM*. La concepción de cómo abordar el problema de la detección de *outliers* es diferente a la de los algoritmos anteriores:
- Al obtenerse para cada elemento del *universo* la región de valores de los umbrales en la cual dicho elemento es *outlier* —según *VPRSM*—, tenemos la posibilidad de hacer un recorrido por todos los elementos del mismo buscando si pares particulares de valores (β, μ) pertenecen a su *región de excepcionalidad*. Por tanto, el resultado que se obtiene ejecutando este algoritmo, contiene cualquier resultado particular que pudiese obtenerse a partir de la ejecución de los algoritmos anteriores. Esto constituye su principal ventaja.
- La región obtenida tras la ejecución de este algoritmo permite establecer una aproximación estocástica a la solución del problema de determinar la probabilidad de cada elemento del universo de ser *outlier* dentro del mismo.

- Su uso resulta especialmente factible cuando se necesita obtener un resultado acerca de la condición de *outlier* de los elementos del *conjunto de datos* para un conjunto determinado de valores de los umbrales, lo cual resulta imposible realizar a partir de las versiones anteriores del algoritmo.

Principal Limitación

- Las principales limitaciones de este algoritmo son su *complejidad temporal*: $O(n^2 \times m^2 \times log(m))$ y su complejidad espacial: $O(n^2 \times m^2)$ para el *caso peor.*

  El costo, en cuanto a tiempo y espacio, de este algoritmo es mayor que el de las anteriores propuestas. Esto no es para sorprenderse, ya que es más general que sus versiones anteriores. Independientemente del orden de *complejidad temporal* señalado para el *caso peor*, el algoritmo puede llegar a alcanzar un orden de *complejidad temporal* similar al de los algoritmos *RSBM* y *VPRSM*: *Casi lineal* con respecto a la cardinalidad del *conjunto de datos*, para el *caso mejor:* $\Omega(n \times m^2 \times c)$, donde *c:* costo de aplicar un *clasificador* a un elemento del *universo*.

Universitat d'Alacant
Universidad de Alicante

Algoritmo *BM/Probabilístico*

El algoritmo *BM/Probabilístico* proporciona la probabilidad de cada elemento del *universo* de ser *outlier*: Sus principales ventajas y limitaciones se analizan a continuación.

Principales Ventajas

- El resultado que se obtiene tras su ejecución es más general también que el que se obtiene tras la ejecución

de *RSBM* y *VPRSM*. La concepción de cómo abordar el problema de la detección de *outliers,* al igual que en *BM,* es diferente a la de dichos algoritmos y a la de la mayoría de los algoritmos de detección de *outliers* existentes:

- El principio básico de funcionamiento que caracteriza a el algoritmo *BM/Probabilístico* es original con relación al del resto de los métodos de detección de *outliers* existentes. En todos ellos, independientemente de la técnica en la que estén basados y con mejores o peores costes computacionales, se establecen unas condiciones particulares, según el contexto en el que serán aplicados los algoritmos bajo las cuales son capaces de proporcionar *un conjunto de outliers* dentro de un *universo* o *conjunto de datos* dado. No obstante, aunque en nuestro caso y en todas las variantes de los algoritmos diseñados se mantiene la necesidad de establecer como condiciones particulares el *concepto* y las *relaciones de equivalencia*, el resultado que se obtiene tras la ejecución del *BM/Probabilístico* ya no es un *conjunto de outliers*. Es un resultado más general que permite, de forma no supervisada —no es necesario que el usuario establezca el valor de los umbrales—, determinar la probabilidad de cada elemento del *universo* de ser *outlier* en dicho *universo*.
- El algoritmo valida la viabilidad computacional del método a partir del cual se determina el resultado que se acaba de expresar.

Principal Desventaja

- La principales limitaciones de este algoritmo, al igual que *BM*, son su *complejidad temporal*: $O(n^2 \times m^2 \times log(m))$ para el *caso peor* y su *complejidad espacial*: $O(n^2 \times m^2)$

para el *caso peor*. No obstante, son válidas también las consideraciones que fueron hechas para *BM* con respecto al orden de *complejidad temporal* para el *caso mejor*.

Resumen

En la **Tabla 7-1** se muestra un cuadro comparativo donde se resumen, de manera abreviada, las principales ventajas y desventajas de los algoritmos *RSBM*, *VPRSM*, *BM* y *BM*/Probabilístico.

Tabla 7-1 Cuadro comparativo, atendiendo a VENTAJAS y DESVENTAJAS, de los algoritmos *RSBM, VPRSM, BM* y *BM/Probabilístico*

	Ventajas	Desventajas
RSBM	• Valida la viabilidad computacional del método de detección basado en *RSBM* • Complejidad temporal y espacial lineal con respecto a la cardinalidad del conjunto de datos	• DETERMINISTA en lo que respecta a la clasificación • Necesidad de establecer, por parte del usuario, el *umbral de excepcionalidad*
VPRSM	• Valida la viabilidad computacional del método de detección basado en *VPRSM* • Complejidad temporal y espacial lineal con respecto a la cardinalidad del conjunto de datos • NO DETERMINISTA en lo que respecta a la clasificación	• El usuario debe definir, además del *umbral de excepcionalidad*, el *error de clasificación* permitido • Una selección no adecuada de este *error* puede conllevar a resultados no satisfactorios. Por tanto, es necesario tener suficiente conocimiento de aspectos específicos del *conjunto de datos* a partir del cual se pueda hacer una selección adecuada del mismo

Tabla 7-1 (continuación)

	Ventajas	Desventajas
BM	• Valida la viabilidad computacional del método a partir del cual se obtiene la *región de excepcionalidad* para cada elemento del universo • La condición de *outlier* se establece a partir de *VPRSM*, por tanto mantiene el no determinismo de este modelo • A partir del resultado que se obtiene tras la ejecución de este algoritmo, se puede determinar cualquier resultado particular que pudiese obtenerse a partir de la ejecución de los algoritmos *RSBM* y *VPRSM* • La región obtenida tras la ejecución de este algoritmo permite establecer una aproximación estocástica a la solución del problema de determinar la probabilidad de que un elemento dado es *outlier* dentro de un determinado conjunto de datos • Su uso resulta especialmente factible cuando se necesita obtener un resultado a cerca de la condición de *outlier* de los elementos del *conjunto de datos* para un conjunto determinado de valores de los umbrales	• Complejidad temporal: $O(n^2 \times m^2 \times \log(m))$ para el caso peor • Complejidad espacial: $O(n^2 \times m^2)$ para el caso peor
BM/P	• Valida la viabilidad computacional del método a partir del cual se obtiene la probabilidad para cada elemento del universo de ser *outlier* dentro del mismo • La condición de *outlier* se establece a partir de *VPRSM*, por tanto mantiene el no determinismo de este modelo • El resultado que se obtiene tras su ejecución es más general que el que se obtiene tras la ejecución de *RSBM* y *VPRSM*. La concepción de cómo abordar el problema de la detección de *outliers* es diferente a la de dichos algoritmos y a la de la mayoría de los algoritmos de detección de *outliers* existentes: enfoque original	• Complejidad temporal: $O(n^2 \times m^2 \times \log(m))$ para el caso peor • Complejidad espacial: $O(n^2 \times m^2)$ para el caso peor

Universitat d'Alacant Universidad de Alicante

Comparación con otros Métodos de Detección de *Outliers*

Métodos Estadísticos

Entre los aspectos fundamentales que se le critican a los modelos estadísticos y que limitan su aplicación en los contextos de *KDD-DM* se señalan los siguientes:

- Estos métodos son apropiados, generalmente, para el procesamiento de *conjuntos de datos* con valores *reales continuos* o, al menos, datos *cualitativos* con valores *ordinales*. Contrasta con ello el hecho de que en la actualidad se necesita, cada vez más, procesar datos expresados de manera *categórica* (*no ordinales*) y por tanto, dichos métodos no son aplicables, de forma directa, a este tipo de datos. En algunos casos, para lograr su aplicación, se hace necesario realizar complejas transformaciones de los datos que empeoran la *complejidad temporal* de los algoritmos (Walfish, 2006). Esta es una de las desventajas que se señala, especialmente, a los métodos basados en *distribuciones* (Han & Kamber, 2000). Ellos, por lo general, están orientados a *conjuntos de datos* donde los atributos tienen valores *reales* (*continuos*).
- Los enfoques basados en *distribuciones,* además, suponen la necesidad de conocer a priori la *distribución* de los datos como algo necesario para poder ejecutar el método. Sin embargo, en la mayoría de los problemas del mundo real, la *distribución* de los valores de los atributos se desconoce. Para adaptar las observaciones a una *distribución estándar,* y seleccionar el *test* adecuado; se requieren esfuerzos computacionales no triviales cuando se trabaja con un *conjunto de datos* de

gran tamaño (Hodge & Austin, 2004). Los métodos *paramétricos*, por ejemplo, asumen que los datos deben seguir una *distribución paramétrica*. Como caso típico, una *distribución univariada*. En tales casos, los métodos no funcionan correctamente en contextos *multivariados* (Otey *et al.*, 2005c).

- El problema de la alta dimensionalidad de muchos de los *conjuntos de datos* de hoy en día es otro de los factores que implica que determinados métodos de detección de *outliers* no funcionen correctamente (Aggarwal, 2007).

Ninguno de los aspectos antes mencionados constituye una limitación para la aplicación de los algoritmos (*RSBM, VPRSM, BM y BM/Probabilístico*) basados en la Teoría de *RS*.

Los cuatro algoritmos mencionados pueden ser aplicados sobre *conjuntos de datos* donde en los valores de sus atributos se mezclen datos *continuos* y datos *discretos* (*categóricos*). Resulta interesante destacar que a partir de la base matemática que caracteriza a la Teoría de *RS* —la teoría de las *relaciones de equivalencia*— puede comprobarse que éstas constituyen un mecanismo natural a partir del cual se pueden discretizar datos continuos. El siguiente ejemplo ilustra lo antes expuesto: En las primeras validaciones que se hicieron del algoritmo *RSBM,* se trabajó sobre un *conjunto de datos* que contenía información de *pacientes* con problemas cardiovasculares. Los datos sobre ellos se representaban por los atributos: *edad*, *peso*, *talla*, *cintura*, *tensión sistólica-TAS* y *tensión diastólica-TAD*. Algunas de las *relaciones de equivalencia* definidas sirvieron para discretizar los valores continuos de algunos atributos, como por ejemplo, el *peso*. En tal caso, una de ellas se definió de la siguiente forma:

El elemento x pertenece a la clase de equivalencia k ⇔ k-0,5 ≤ x[PESO] < k + 0,5.

Según esta relación, por ejemplo, a la *clase de equivalencia* 44 de la partición, pertenecerán los pacientes cuyo peso esté entre *43,5 ≤ x[PESO] < 44,5*

Métodos basados en *Distancias*

Como ya se ha señalado, entre las aproximaciones no paramétricas más representativas están los métodos de detección de *outliers* basados en *distancias*. En (Otey *et al.*, 2005c) se señala como una limitación de los mismos, el hecho de que una buena parte de ellos tiene *complejidad temporal* cuadrática para el *caso peor*, lo cual resulta una limitación cuando se trabaja con *conjuntos de datos* muy grandes o dinámicos. Los algoritmos *RSBM y VPRSM* puede decirse que resuelven los problemas asociados al orden cuadrático de la mayoría de los métodos basados en *distancias* pues su *complejidad temporal* es *casi* lineal —con respecto a la cardinalidad del universo— para el *caso peor*. Esa misma *complejidad temporal* la presenta *BM y BM/Probabilístico* para el *caso mejor,* aunque ya sabemos que la concepción de estos algoritmos no se corresponde con la tradicional.

Universitat d'Alacant
Universidad de Alicante

No obstante, existen variantes optimizadas de métodos de detección de *outliers* basados en *distancias* que alcanzan *complejidad temporal lineal* con respecto a la cardinalidad del *conjunto de datos*, pero exponencial con respecto a la cantidad de atributos asociados a cada dato. Este hecho limita la aplicación de dichas variantes solo a contextos donde la cantidad de atributos es pequeña.

Es obvio que la necesidad de definir *funciones* o *criterios de distancia* adecuados es algo que caracteriza a estos

métodos (Angiulli *et al.*, 2006), (Angiulli *et al.*, 2007), sin embargo, este aspecto es otra de las desventajas que se le señala a los mismos. A continuación se dan algunos ejemplos de las dificultades que se pueden presentar ante esta necesidad:

- Cuando en los cálculos de la función de *distancia* intervienen dos atributos donde los posibles valores de uno varían mucho más que los del otro, entonces aplicar un criterio de *distancia* donde estén implicados, a la vez, el valor de ambos, puede conllevar a que los cálculos no sean representativos (Prayote, 2007). Un modo de proceder podría ser, en estos casos, dividir cada atributo por su *desviación estándar* para obtener atributos estandarizados y luego aplicar la función estándar de *distancia euclidiana*. Esto implica tener que hacer cálculos adicionales. No obstante, existen situaciones donde no es razonable el uso de una *distancia métrica*.
- El problema de la alta dimensionalidad de los actuales *conjuntos de datos* es otro de los factores que puede traer dificultades a la hora de aplicar métodos de detección basados en criterios de *distancia*. Aumentar la dimensión del *conjunto de datos* puede afectar la aplicación de un determinado método. Por ejemplo, el concepto de *distancia* en un espacio de dimensión *k*, no es el mismo que en un espacio de dimensión *k+1*.
- En general, investigaciones recientes (Angiulli *et al.*, 2007) han demostrado que el concepto de *proximidad* en un *conjunto de datos* de gran dimensionalidad puede no ser cualitativamente significativo. Por tal motivo, los enfoques basados en *distancias* no se consideran apropiados para la detección de *outliers* en *conjuntos de datos* con tales características.

Universitat d'Alacant
Universidad de Alicante

Con respecto a estas consideraciones, podemos decir que ninguno de los algoritmos propuestos trabaja con criterios de *distancia*. Sin embargo, en el estudio del *estado del arte* realizado, pudo comprobarse que una buena parte de los métodos de detección existentes en la actualidad incorporan de alguna forma a sus análisis el uso de alguna técnica de *distancia*. De igual forma, para los algoritmos de detección basados en *RS,* el tamaño y la dimensionalidad del *conjunto de datos* no representan obstáculos para que puedan ejecutarse eficazmente.

Enfoques basados en *Densidades*

Una buena parte de las definiciones de *distancia* consideradas en los métodos de detección que se basan en esta técnica, sólo capturan la esencia de un cierto tipo de *outliers*; que podría denominarse global, pues es el resultado de analizar el *conjunto de datos* como un todo, sin tener en cuenta características propias de algunas de sus áreas. Sin embargo, ante algunos problemas de la vida real con una estructura más compleja; el concepto de *outliers* se concreta aún más. Se trata de ciertos objetos que con respecto a la *densidad* de datos que existe en su *vecindad local* parecen seguir un patrón diferente. Esta es la esencia de la concepción de los métodos de detección basados en *densidades*. No obstante, estos métodos suelen incorporar también al análisis el uso de técnicas de *clustering* y criterios de *distancia*. Por lo general los métodos que siguen este enfoque asignan un *factor de ruido* a cada objeto y mientras más *ruidoso* sea un objeto, mayor es la probabilidad de que sea un *outlier.* Cada enfoque que sigue este paradigma define el *factor de ruido* según criterios particulares. Por ejemplo, Breunig (Breunig *et al.*, 1999), (Breunig *et al.*, 2000), define la noción de *outliers locales*. A partir de ella, a cada objeto se le asigna un valor

LOF (*Local Outlier Factor*) y este es su *grado de ruido* tomando en cuenta la estructura de un *cluster* en una *vecindad* acotada del objeto. En este aspecto, existe una semejanza de estos métodos con la concepción del método basado en *RS* a partir del cual, para cada objeto del *universo,* se define el *grado de excepcionalidad* y a partir de cuyo valor se establece la condición de *outlier* de los mismos.

DBSCAN es uno de los algoritmos de *clustering* basados en *densidades* más citados. A partir del establecimiento de los *clusters*, que es el objetivo principal del algoritmo, el mismo se usa colateralmente para detectar *outliers* en un *universo* dado. El algoritmo fue expuesto por primera vez en (Martin *et al.*, 1996) y fue concebido para establecer *clusters* que permiten clasificar *puntos* en *bases de datos espaciales —bases de datos optimizadas que almacenan datos relacionados con objetos en el espacio como puntos, rectas y polígonos*— de gran tamaño. Es una técnica que estima la *densidad* a partir de un punto considerado *centro* y teniendo en cuenta la cantidad total de otros *puntos* que estén en un *radio* ε dado de dicho *centro*. En tal caso la *densidad* de cada *punto* depende del *radio* seleccionado. En general, el método clasifica los *puntos* en tres categorías:

Puntos *núcleos* o *centros*: estos *puntos* se encuentran en el interior del *cluster*. Un *punto* es un *núcleo* si el número de *puntos* en su *vecindad* excede un cierto umbral *MinPts*.

Puntos *fronteras*: un *punto frontera* no es un *núcleo*, aunque deberá estar en la *vecindad* de alguno. Es posible que un *punto frontera* pertenezca a la *vecindad* de varios *puntos núcleos,* aun estando estos en diferentes *clusters*.

Puntos de *ruido* o *outliers*: es cualquier *punto* que no es un *núcleo* ni *frontera*.

En la funcionalidad de este algoritmo, el *radio* ε y el umbral *MinPts* juegan un papel importante. Ambos parámetros deben ser seleccionados por el usuario y, en la mayoría de los casos, una adecuada selección de los mismos no resulta trivial y el no acertar en ella puede llevar aparejadas determinadas dificultades. Por ejemplo, cuando se selecciona un ε muy grande, el *conjunto de datos* completo constituye un *cluster* de gran tamaño. Sin embargo, cuando se selecciona un ε muy pequeño, cada *punto* pertenecerá a un *cluster* en el cual él es su único elemento. La necesidad de que el usuario seleccione estos parámetros es una de las desventajas que se señala a este método y en tal caso, puede señalarse como una semejanza con los algoritmos *RSBM* y *VPRSM,* donde los umbrales deben ser suministrados por el usuario y lo cual se señala también como una de las desventajas de los mismos. En el algoritmo *BM* y en *BM*/Probabilístico se libera al usuario de tal responsabilidad. No obstante, a criterio del usuario, en *BM/Probabilístico* quizás se necesite establecer el valor donde se fija el *corte* que no es más que un valor de β.

Como otra desventaja de *DBSCAN* se señala lo siguiente: Cuando los *conjuntos de datos* contienen *clusters* de diferentes *densidades*, *DBSCAN* puede tener dificultades para identificar los mismos, lo cual incide directamente en la detección de *outliers*. Cabe destacar que aunque *DBSCAN* es usado para detectar *outliers;* su principal objetivo no es ese, sino establecer *clusters* en un *conjunto de datos* bajo determinados criterios de *densidad*. En este algoritmo la definición de *outliers* se establece indirectamente a través de la noción de *clusters* y en esencia, su objetivo principal es optimizar el *clustering* y no la detección de *outliers*. Este aspecto también establece una diferencia con los algoritmos basados en *RS* en los que la detección es el principal objetivo.

Universitat d'Alacant
Universidad de Alicante

Como ventaja de este algoritmo y a su vez similitud con los algoritmos antes mencionados, cabe señalar que la mezcla de tipos de atributos (*discretos* y *continuos*) no es una limitación para su ejecución.

Con respecto a la *complejidad espacial* y *temporal* puede decirse que este algoritmo tiene buenos órdenes. Con respecto al primero, es O(*n*). Apenas una pequeña cantidad de datos es almacenada para un *punto*: el número del *cluster* y quizás la *clasificación* como *punto núcleo*, *frontera* o *ruido*. Con respecto a la *complejidad temporal* para el *caso peor*, este algoritmo es O*(n log n*). En tal sentido puede decirse que, comparando éste con los algoritmos *RSBM y VPRSM,* se comportan aproximadamente de forma similar en cuanto a *complejidad temporal*.

Enfoques basados en *Profundidades* (*depth-based*)

Los enfoques basados en *profundidades* (Ruts & Rousseeuw, 1996), (Jhonson *et al.*, 1998), han sido propuestos con el objetivo de eliminar las limitantes de las técnicas basadas en *distribuciones*. Los datos se representan como *puntos* en un *espacio k*-dimensional y son organizados en *capas* (*layers*), con la expectativa de que las capas superficiales sean las más propensas a contener *outliers*. Estos métodos pueden superar el problema de la adecuación de la *distribución*, y conceptualmente pueden procesar datos en un espacio multidimensional.

Sin embargo, en la práctica, hay un problema computacional en el enfoque. Para computar las capas *k*-dimensionales, la técnica depende del cómputo de la *envoltura convexa* (*convex hull*) *k*-dimensional, que tiene una *complejidad computacional* ($n^{k/2}$). Por lo tanto, los métodos basados en la *profundidad* de los datos no son

prácticos para *conjuntos de datos* de *dimensión* mayor que 4 que posean grandes volúmenes de datos (Knorr & Ng, 1998). De hecho, suele señalarse que los algoritmos existentes basados en esta técnica alcanzan su mayor eficiencia cuando la *dimensión del conjunto de datos* ≤2 (Ruts & Rousseeuw, 1996).

Las principales ventajas de los algoritmos basados en *RS* con respecto a este enfoque, están centradas en dos aspectos fundamentales: La *complejidad temporal* de los mismos y el hecho de que la dimensión del *conjunto de datos* no es una restricción para que funcionen correctamente.

Enfoques basados en *Particiones* (*Partition-based*)

Los métodos basados en *particiones* son uno de los métodos clásicos que usan técnicas de *clustering*. A partir de ellas se crean varias particiones en el *conjunto de datos* y en ellos, igualmente, intervienen criterios de *distancia* en el análisis. El usuario proporciona el número de *clusters M* que desea crear y un número *K* de variables que intervienen en la creación de esas particiones. Entre los algoritmos más citados que se basan en este enfoque, pueden señalarse: *k-means* (Hautamäki *et al.*, 2005), *PAM* y *CLARA* (Kaufman & Rousseeuw, 1990), (Han & Kamber, 2000).

A modo de ejemplo, veamos algunos aspectos del *k-means* que es uno de los métodos de *clustering* más conocidos y estudiados por su aplicación a objetos en *espacios euclidianos*. Seguidamente se da una explicación general del método de detección: *Clustering Outliers Removal* (Hautamäki *et al.*, 2005) que está basado, precisamente en *k-means clustering*.

Las entradas del método son: el *conjunto de datos X*, el número de *clusters* a formar *M*, un umbral *Th* y el número de iteraciones que se desea realizar, *R*.

Primero se seleccionan *M* objetos que serán los *centroides* de cada *cluster*. Luego, se itera sobre los elementos de *X* y cada elemento del *conjunto de datos* se ubica dentro del *cluster* que le corresponde a partir de la búsqueda del *centroide* más cercano a él. Para ello, utiliza una determinada función de *distancia*. Tras esta primera iteración, vuelve a calcular los *centroides* de acuerdo a la partición formada. Luego recorre cada *cluster* calculando la *distancia* del elemento más lejano al *centroide* (d_{max}) y del más cercano (d_{min}) y luego calcula la *distorsión* ($d = d_{min} / d_{max}$). Si $d<th$, entonces el elemento más lejano del *centroide* es considerado *outlier* y es eliminado del *conjunto de datos*. Los elementos que estén solos en un *cluster* también se consideran *outliers*. Las condiciones de parada son dos: Cuando se alcanza el número máximo de iteraciones previstas *R* o cuando tras una iteración no se eliminan *outliers* del *conjunto de datos*.

A continuación se muestra un ejemplo sencillo que sirve para esclarecer la esencia del método.

Ejemplo 7-1

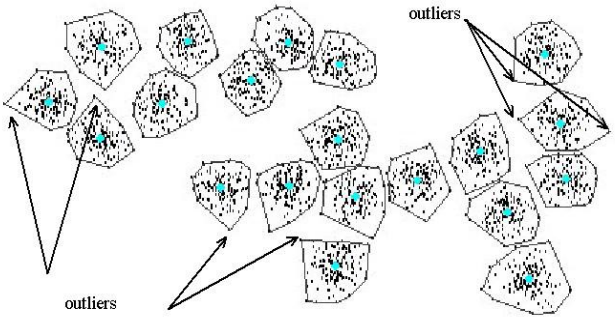


Figura 7-1 *Clustering Outliers Removal. Clusters* y *outliers*

La **Figura 7-1** muestra el estado de los *clusters* después de ejecutar el algoritmo para 5 iteraciones y un umbral de 0,009. En ella se puede ver que los objetos que se han señalado como *outliers* se encuentran bastante lejos de su *centroide*. Por tanto, al ser los más alejados, el algoritmo los elimina. La situación resultante tras la eliminación de los mismos puede apreciarse en la **Figura 7-2**.

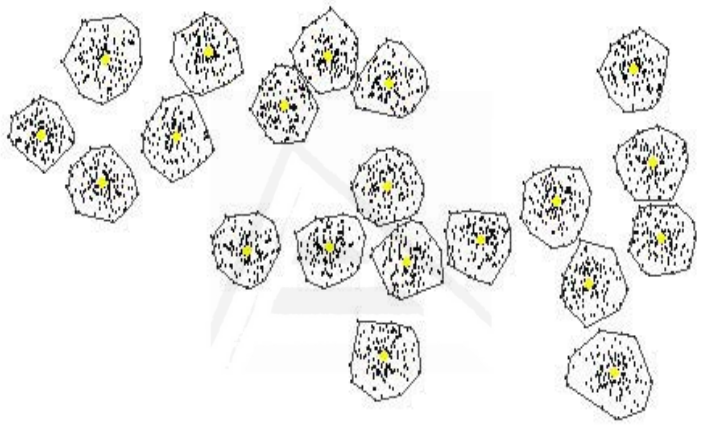


Figura 7-2 *Clustering Outliers Removal. Clusters* sin *outliers*

Complejidad temporal: $O(R \times k \times n)$ donde R es el número de iteraciones que se realizan, k el número de variables a analizar y n la cardinalidad del *conjunto de datos*. En este caso, el orden de *complejidad temporal* de este algoritmo, en comparación con el algoritmo *RSBM* y el *VPRSM,* puede decirse que es similar, es decir, *casi* lineal con respecto a la cardinalidad del *conjunto de datos*.

Un problema que tiene este método es que el usuario debe seleccionar parámetros tales como el umbral, el número de iteraciones y el número de *cluster*s que generalmente son difíciles de definir. Este aspecto puede señalarse como una similitud con los algoritmos *RSBM* y *VPRSM*, donde, como

ya hemos apuntado, la selección de los umbrales es responsabilidad del usuario.

Otra semejanza de este algoritmo con los algoritmos basados en *RS,* y a su vez una ventaja del mismo, es el hecho de que trabaja sin dificultad tanto con variables *discretas* como *continuas*.

La selección del número de *clusters* es un importante subproblema de los algoritmos de *clustering* y su solución afecta directamente la calidad de los resultados. Si se selecciona un número pequeño, se corre el peligro de que objetos diferentes no queden separados. Del mismo modo, la selección de un número muy grande puede llevar a que regiones relacionadas estén en diferentes *clusters*. Además, la necesidad de establecer una función de *distancia* hace que las limitaciones que esto supone sean heredadas por este tipo de métodos. Estas desventajas no son inherentes a los algoritmos basados en *RS*, lo que es una ventaja de los mismos con respecto a los que estamos analizando. En ellos, para el análisis, no se aplican ni técnicas de *clustering* ni de *distancia*. No obstante, cabe señalar que un riesgo similar al que representa una mala selección del número de *clusters* a tener en cuenta, puede establecerse al no hacer una adecuada selección de las *relaciones de equivalencia* más adecuadas. El mismo riesgo se corre si no se hace una adecuada selección del *error de clasificación* en el algoritmo *VPRSM.*

Enfoques basados en *Redes Neuronales*

En (Hodge & Austin, 2004) se hace un análisis detallado de la aplicación de modelos de *redes neuronales* al problema de la detección de *outliers*. Los criterios expuestos en dicho trabajo han servido como punto de partida para el análisis que se hace a continuación.

Los aspectos esenciales que constituyen desventajas de los métodos de detección basados en *redes neuronales* con respecto a los métodos basados en *RS* presentados en esta investigación son los siguientes:

Métodos Supervisados

- En varios métodos de detección basados en algunos tipos de *redes neuronales*, la dimensionalidad del *conjunto de datos* influye negativamente en la efectividad del método.
- En varios métodos, por ejemplo (Bishop, 1994), es necesario establecer criterios sobre la *densidad* de los datos como requerimientos esenciales para la ejecución de los mismos.

Métodos No Supervisados

- Algunos métodos de detección basados en modelos de *redes neuronales no supervisados*, como por ejemplo *SOM* (Kohonen, 1995), requieren modelar la *distribución* de los datos para establecer la funcionalidad del método. De igual forma, en otros casos, se hace necesario establecer criterios de *distancia* (Saunders & Gero, 2001), (Vesanto *et al.*, 1998) entre los datos. Como ya hemos señalado con anterioridad, esto limita considerablemente la aplicación del método en determinados tipos de *conjuntos de datos*.

Mientras tanto, a modo de semejanza entre los métodos de detección basados en *redes neuronales* y los basados en *RS*, concretamente los algoritmos *RSBM* y *VPRSM,* puede señalarse lo siguiente*:* algunos enfoques basados en *redes supervisadas* establecen el uso de umbrales con diversos fines en el proceso de detección de *outliers*. Por ejemplo, en el caso del método basado en *MLP-RNN* (Graham *et al.*, 2002)

Universitat d'Alacant
Universidad de Alicante

Tabla 7-2 Principales limitaciones de otros métodos de detección que resuelven los algoritmos de detección basados en *RS*

Los algoritmos de detección basados en *RS*, resuelven las siguientes limitaciones de los:
MÉTODOS ESTADISTICOS Y MÉTODOS BASADOS EN *DISTANCIAS*
• Son aplicables a conjuntos de datos donde haya mezcla de atributos, continuos y discretos. Las relaciones de equivalencia suponen una forma natural de discretizar datos continuos. • Para poder ejecutar los algoritmos, no es necesario conocer la distribución de los datos, ni es necesario establecer criterios de *distancia* sobre los mismos. • En particular, *RSBM* y *VPRSM*, resuelven el problema de la complejidad temporal, orden cuadrático, que presenta la mayoría de los métodos basados en *distancias*. • La dimensionalidad y el tamaño del conjunto de datos no es una limitación para la ejecución de los algoritmos.
MÉTODOS BASADOS EN DENSIDADES Y MÉTODOS BASADOS EN PROFUNDIDADES
• Para poder ejecutar los algoritmos, no es necesario establecer criterios sobre la *densidad* de los datos en el conjunto de datos. • La dimensionalidad del conjunto de datos no es una limitación para la ejecución de los algoritmos. • Para poder ejecutar los algoritmos no es necesario establecer cálculos previos que consumen gran cantidad de tiempo, como por ejemplo, el cálculo de la *envoltura convexa*, necesario en la mayoría de los métodos basados en *profundidades*. • *BM* y *BM/P* permiten obtener los resultados de forma no supervisada, sin necesidad de que el usuario establezca, como paso previo a su ejecución, el valor de ciertos parámetros que intervienen en el análisis, lo cual es necesario en métodos basados en *densidades* como es el caso de *DBSCAN*. • *RSBM* y *VPRSM* suponen mejoras en cuanto a complejidad temporal con respecto a métodos basados en *profundidad*.
MÉTODOS BASADOS EN REDES NEURONALES
• Previo a la ejecución de los algoritmos, no es necesario establecer procesos que consumen gran cantidad de tiempo, como por ejemplo el entrenamiento de la red, necesario en algunos modelos de redes neuronales para garantizar el aprendizaje de la misma. • La dimensionalidad del conjunto de datos no es una limitación para la ejecución de los algoritmos. • La funcionalidad de los algoritmos no depende de criterios sobre la *densidad* de los datos, como se requiere en algunos modelos supervisados. • Para poder ejecutar los algoritmos, no es necesario modelar la *distribución* de los datos. como se requiere en algunos modelos supervisados. • Algunos enfoques basados en redes supervisadas establecen el uso de umbrales con diversos fines en el proceso de detección de *outlier*s. Esto queda resuelto a partir de la concepción de los algoritmos *BM* y *BM/P*.
MÉTODOS DE DETECCIÓN DE OUTLIERS EN GENERAL
A diferencia de la mayoría de los métodos de detección, que requieren de sucesivas ejecuciones de los algoritmos hasta obtener el conjunto de *outliers* que se adecue realmente a los intereses del análisis, *BM/P* determina, de forma no supervisada y con una sola ejecución del mismo, la probabilidad de cada elemento del universo de ser *outlier* en dicho universo. Su concepción, de cómo abordar el tema de la detección de *outlier,* es diferente a la tradicional que asume la mayoría de los métodos de detección existentes.

Universitat d'Alacant
Universidad de Alicante

se establece un *factor de excepcionalidad* para clasificar los datos como *outliers*. De igual forma, en algunos enfoques basados en *auto-associative neural networks* como por ejemplo (Japkowicz *et al.*, 1995), se establecen umbrales que posteriormente son usados en el proceso de *clasificación* de las entradas de datos.

Como hemos señalado, existen métodos de detección basados en ciertos tipos de redes neuronales que requieren que el conjunto de datos sea recorrido múltiples veces para garantizar el entrenamiento previo de la red. En tales casos, esto conlleva a un costo adicional de tiempo que afecta el orden de complejidad temporal global del proceso de detección de outliers. En cierta medida, esto se puede ver como una semejanza con los algoritmos RSBM y VPRSM, pues, como ya se ha explicado, independientemente de la linealidad de su complejidad temporal con respecto a la cardinalidad del conjunto de datos, no se puede decir lo mismo del costo del proceso global del análisis de casos excepcionales. Sin embargo, el resultado que se obtiene tras una ejecución del algoritmo BM —aun cuando la complejidad temporal del mismo sea un poco mayor que la de RSBM y VPRSM— contiene cualquier resultado particular que pudiese obtenerse a partir de la ejecución de los algoritmos anteriores. Esto conduce a una optimización del coste, en cuanto a tiempo, del proceso general de análisis de casos excepcionales.

En la **Tabla 7-2** se resumen las principales limitaciones de otros métodos de detección de *outliers* que resuelven los algoritmos basados en *RS*.

Universitat d'Alacant
Universidad de Alicante

Capítulo 8

Conclusiones

En esta investigación se ha creado un novedoso método basado en la *Teoría de Conjuntos Aproximados* que determina, de forma no supervisada, la probabilidad de cada elemento del universo de ser *excepcional* en dicho universo. A partir del método se ha concebido un algoritmo computacionalmente viable que proporciona una herramienta realista con la que aplicar el método en entornos reales y que ha permitido demostrar la validez de la propuesta.

Teniendo en cuenta la incidencia que en los últimos años ha tenido el problema de la detección de *casos excepcionales* u *outliers* en diferentes contextos y en especial en los *procesos de búsqueda de información en grandes volúmenes de datos* o *KDD-DM*, así como lo efectiva que ha resultado la aplicación de la Teoría de *Conjuntos Aproximados* o *Rough Sets* en la solución de disímiles problemas dentro de dichos procesos, el aspecto esencial que motivó la realización del presente trabajo de investigación fue el interés por explorar cuan efectiva

podría ser la aplicación de dicha Teoría al ámbito de la detección de *outliers*.

Estudiando los métodos tradicionales de detección de *casos excepcionales* u *outliers* —basados en diferentes técnicas y con mejores o peores órdenes de *complejidad temporal*— se pudo constatar que en su gran mayoría necesitan el preestablecimiento de unas condiciones particulares que dependen directamente del contexto en el que serán aplicados y bajo las cuáles los algoritmos son capaces de proporcionar *un conjunto de outliers* dentro de un *universo* de datos dado. Por lo tanto, los algoritmos tradicionales requieren un análisis previo del contexto y, posteriormente a su ejecución, un análisis del conjunto de *outliers* que han sido identificados, de forma que puede ser necesario establecer nuevamente las condiciones particulares bajo las cuales se pueda obtener un nuevo conjunto de *outliers.* Este proceso puede requerir múltiples iteraciones hasta que el conjunto de *outliers* que se obtenga se adecue realmente a los intereses iniciales del análisis. Por tanto, aunque esté determinada la *complejidad temporal* de los algoritmos para el *caso peor*, no se puede decir lo mismo del proceso global de análisis de *casos excepcionales*.

La principal hipótesis propuesta para solucionar este problema se ha basado en la búsqueda de enfoques diferentes para la concepción de nuevos algoritmos para la detección de casos excepcionales, en concreto, que es posible desarrollar una nueva teoría basada en la extensión de los conceptos básicos y las herramientas formales que nos proporciona la Teoría de Conjuntos Aproximados (*RSBM*) y el *Modelo de Conjuntos Aproximados de Precisión Variable* (*VPRSM*), aplicados al problema de la detección de *outliers,* que permita obtener, de forma no supervisada, para cada elemento de un *universo* de datos, la región de valores de los umbrales bajo la cual dicho elemento es

outlier. A partir de dicho resultado, es posible determinar la probabilidad de cada elemento del *universo* de ser *outlier* dentro del mismo.

En función de esta hipótesis, se plantea como objetivo general de la investigación: establecer un método, basado en la Teoría de *RS* que determine, de forma no supervisada —respecto a la elección de los umbrales que intervienen en el análisis—, la probabilidad de cada elemento del *universo* de ser *outlier* dentro del mismo. La concepción de este método supone abordar el problema de la detección de *outliers* bajo un nuevo enfoque, original y diferente al tradicional.

En el estudio del *estado del arte* realizado resultó atractiva la propuesta de un método de detección de *outliers* basado en *RS* cuyo marco formal era muy cercano a nuestro interés y, hasta ese momento, constituía el primer antecedente de su aplicación al problema de la detección de *outliers.* Esto motivó el análisis crítico de la propuesta, revelando que, aunque el enfoque aporta un marco teórico muy sólido y coherente, la instrumentación computacional del método que se porpone implica caer en un problema computacionalmente no tratable.

Universidad de Alicante

En este trabajo, teniendo en cuenta el marco teórico general de la Teoría de *RS* y los elementos conceptuales que sirven como base a la propuesta antes mencionada, se hace una extensión del marco teórico existente, a partir de lo cual se establece, como primer resultado de la investigación, un método de detección de *outliers*, computacionalmente viable, basado en el modelo básico de la Teoría de *RS* o *RSBM*.

El método anterior hereda el carácter determinista de *RSBM* en lo relativo a la *clasificación.* Para subsanar tal limitación se incorporan al nuevo marco teórico propuesto

elementos conceptuales del *Modelo de Conjuntos Aproximados de Precisión Variable* o *VPRSM,* a partir de lo cual se establece —como segundo resultado de la investigación— un método de detección de *outliers* basado en *RS* computacionalmente viable y, además, no determinista.

El marco formal alcanzado sirvió de antecedente a la concepción general de un procedimiento que permite determinar, para cada elemento del *universo*, la *región de excepcionalidad* del mismo con respecto a los umbrales que intervienen en el análisis. Este tercer resultado establece las bases teórico-matemáticas necesarias para conceptualizar un método que determina, de forma no supervisada —con respecto a la elección de los umbrales antes mencionados— y mediante la aplicación de técnicas estocásticas, la probabilidad de cada elemento del *universo* de ser *outlier* en dicho universo. Con este resultado se da cumplimiento al objetivo general de la investigación y, a su vez, se da solución al problema general planteado dentro de la misma.

La viabilidad computacional de todos los métodos propuestos fue validada mediante la concepción de un algoritmo concreto. En todos los casos, la funcionalidad de los algoritmos propuestos así como la *complejidad temporal* determinada en el análisis teórico, fueron validadas mediante *conjuntos de datos* del mundo real y con *conjuntos de datos* generados de forma *sintética* por métodos estadísticos.

Ámbito de Aplicación de la Propuesta

Como paso previo a la aplicación de cualquiera de los algoritmos propuestos, los especialistas que intervienen en

el análisis de los datos, en dependencia de los objetivos propuestos; deben definir determinados aspectos que resultan esenciales para la ejecución de los mismos a partir del enfoque de detección propuesto. Estos aspectos son: el establecimiento del *concepto*, es decir, el subconjunto de elementos del *universo* en los que se desea centrar el análisis, así como el establecimiento de un conjunto de *relaciones de equivalencia* que constituyen las *aristas del conocimiento*, a partir de las cuales deseamos aproximarnos a dicho *concepto* y que aportan *conocimiento* acerca del mismo. Esto permite, en gran medida, decantar el ámbito de aplicación de los métodos propuestos y, por tanto, de los algoritmos presentados:

El método es aplicable a datos en forma tabular. Los atributos de la *Tabla* deben ser monovaluados pues, de lo contrario, entrarían en contradicción con la esencia del método ya que, de no ser así, no existe la posibilidad de establecer *relaciones de equivalencia* a partir de ellos.

Principales Aportaciones

El principal aporte de esta investigación es el siguiente:

Un marco teórico general basado en la extensión de los conceptos básicos y las herramientas formales que proporciona la Teoría de Conjuntos Aproximados (*RSBM*) y el *Modelo de Conjuntos Aproximados de Precisión Variable* (*VPRSM*), aplicados al problema de la detección de *outliers,* a partir del cual se establecen cuatro métodos que dan solución al problema antes mencionado desde diferentes concepciones:

A. Un método de detección de *outliers*, computacionalmente viable, basado en el modelo básico de la Teoría de Conjuntos Aproximados (*RSBM*).

B. Un método de detección de *outliers*, computacionalmente viable y no determinista, basado en el *Modelo de Conjuntos Aproximados de Precisión Variable* (*VPRSM*).

C. Un método que, de forma no supervisada —con respecto a la elección de los umbrales que intervienen en el análisis—, determina la *región de excepcionalidad* —a partir de los valores de dichos umbrales— de cada elemento de un *universo* de datos dado.

D. Un método que, de forma no supervisada —con respecto a la elección de los umbrales que intervienen en el análisis—, determina la probabilidad de cada elemento de un *universo* de datos dado de ser *outlier,* según *VPRSM,* dentro del mismo.

Como aspectos trascendentes de los resultados alcanzados pueden señalarse los siguientes:

- Permitieron explorar nuevas aplicaciones de la Teoría de Conjuntos Aproximados *(Rough Sets).*
- Constituyen los primeros reportes de soluciones computacionalmente viables al problema de la detección de *outliers* basadas en la Teoría de *RS.*
- La definición de *outliers* propuesta es intuitiva, novedosa y a la vez, consistente con las existentes:
- *Objetos que, según las particiones descritas por las relaciones de equivalencia que intervienen en el análisis, pertenecen a clases de equivalencia que agrupan a objetos con atributos similares y sin embargo, tienen comportamiento contradictorio con respecto a otros objetos de su misma clase.*
- Aunque el principio de la detección de *outliers* inherente a los métodos referidos en A. y B. es el tradicional, ellos resultan simples en cuanto a su planteamiento teórico y

Universitat d'Alacant
Universidad de Alicante

constituyen una primera aproximación que, de forma sencilla, da solución al problema de la detección de *outliers* a partir de la Teoría de *RS*.

- La manera en que se aborda el problema de la detección de *outliers* en los métodos referidos en C. y D. es novedosa y original. Rompe con el esquema tradicional de abordar el mismo por la mayoría de los métodos de detección existentes. A partir del establecimiento de ciertas condiciones iniciales —*concepto* y *relaciones de equivalencia*—, proporcionan, de forma no supervisada, resultados generales con respectos a todos los elementos el *universo* de datos.

- En especial, D. proporciona la probabilidad de cada elemento del *universo* de ser *outlier* en dicho *universo* sin la necesidad de haber establecido —excepto las señaladas en el párrafo anterior— las condiciones previas para ello en función del contexto de aplicación. Este hecho da trascendencia y originalidad a este resultado pues a partir de él, se allana el camino para el análisis y la solución de otros problemas particulares y además, permite tener una visión general sobre los datos que son objeto de estudio en el sentido de poder poner a prueba su representatividad.

- Teniendo en cuenta que la aplicación del modelo de *RS* en el contexto del *KDD-DM* ha demostrado su efectividad y su capacidad para modelar un gran número de situaciones reales y dar solución a disímiles problemas dentro del mismo, la aplicación de los métodos propuestos debe resultar factible en dicho contexto. La validación de los resultados demuestra, en cierta medida, esta afirmación.

- Los algoritmos presentados permitieron validar la viabilidad computacional de los métodos propuestos.

Universitat d'Alacant
Universidad de Alicante

Constituyen además, soluciones computacionales eficientes —en cuanto a *complejidad temporal* y *espacial*— para la solución de los problemas para los cuales fueron concebidos. Esto es una ventaja que cualquier analista/ingeniero de datos valora considerablemente.

- Los métodos propuestos resuelven, además, otras limitaciones de varios métodos de detección:
  - Pueden ser aplicados a *conjuntos de datos* donde exista mezcla de tipos de atributos (continuos y discretos).
  - Para su aplicación no se requiere conocimiento a priori sobre la distribución de los datos.
  - Dentro del ámbito de aplicación de los mismos, el tamaño y la dimensionalidad del *conjunto de datos* no es una limitación para su correcto funcionamiento.
  - Ninguno requiere para su aplicación el establecimiento de criterios de *distancia* o de *densidad* con relación a los datos del conjunto.
  - Los métodos referidos en C. y D. permiten obtener de forma no supervisada —en lo que respecta al establecimiento del valor de los umbrales que intervienen en el análisis— resultados generales para cada elemento del *universo*. Sin embargo, el establecimiento de *umbrales de excepcionalidad* por parte del usuario son requisitos indispensables para garantizar el funcionamiento correcto de varios métodos de detección de *outliers*.

Universitat d'Alacant
Universidad de Alicante

Problemas Abiertos y Líneas Futuras de Investigación

Los resultados expuestos en el presente documento, a pesar de su gran calado, no son más que el comienzo de una investigación más profunda en el contexto del problema general de la detección de *outliers* basada en el modelo de *RS*. Por tanto, se pueden identificar varios problemas que aún no han sido solucionados y que pueden constituir objetivos inmediatos para dar continuidad a la investigación. En tal sentido se han identificado los siguientes:

Problema No. 1

En la versión actual del algoritmo *BM*/Probabilístico la opción de establecer un *corte* a partir del cual se fija una nueva cota para el valor de β, es una opción que asume o no el usuario. En caso que no sea así, el algoritmo establece un valor por defecto. Dicho valor se decidió de forma empírica y a partir de los resultados de las pruebas de validación realizadas.

Universidad de Alicante

Hipótesis

Incorporar a la funcionalidad del algoritmo un análisis teórico particular para cada situación concreta, a la hora de establecer el *corte*, haría más objetiva la selección de la cota para el valor de β que define al mismo y por tanto sería más preciso el cálculo de la probabilidad.

Propuesta de Solución

El algoritmo *BM*/Probabilístico debe establecer de forma no supervisada el valor adecuado para el valor del parámetro β

a partir del cual se determina el *corte*, de manera tal que se garantice optimizar los resultados. En términos gráficos, dicha optimización consiste en minimizar las áreas de las regiones que intervienen en el cálculo de la probabilidad de forma tal que el cálculo realizado sea más preciso.

Problema No. 2

En la versión actual de los algoritmos, el *concepto* está entre las condiciones particulares esenciales que deben quedar establecidas como paso previo a la ejecución de los mismos y en función del contexto donde serán aplicados.

Hipótesis

Se podría aprovechar el marco teórico establecido para poder establecer la detección de *outliers* en casos en que el *concepto* no esté definido explícitamente, ya que parece ser posible conformar el mismo a partir del establecimiento de operaciones lógicas entre varios *subconceptos* ya conocidos.

Propuesta de Solución

En una nueva propuesta podríamos pensar en términos de reusabilidad del código y concebir una base de *conceptos*, donde se expresaran determinados campos del conocimiento y a partir de ellos, poder establecer un algoritmo que de forma automática permitiera generar operaciones lógicas entre los *conceptos base*, a partir de las cuales se puedan conformar nuevos y más complejos *conceptos*. Con ello se minimizan los esfuerzos de programación a partir del principio de reusabilidad.

Restricción

La generación de nuevos *conceptos* debe ser acotada en base a determinados criterios que impidan una generación

exponencial de los mismos.

Conclusiones sobre Proceso de Investigación

- Se ha se ha realizado una investigación siguiendo el método científico que avala las etapas del proceso realizado a partir del cual se han podido identificar un conjunto de aportaciones.
- Se han obtenido resultados originales dentro del campo de la detección de *outliers* a partir de los cuales se ha dado cumplimiento a los objetivos parciales y al objetivo general propuesto en la investigación.
- Los resultados alcanzados validan tanto la investigación como el proceso investigativo que se ha seguido durante el desarrollo de la Tesis. Han permitido, además, establecer líneas futuras concretas de trabajo que permiten dar continuidad a la investigación a partir de los problemas abiertos que han sido identificados.
- El desarrollo del presente trabajo ha permitido la integración con equipos de investigación multidisciplinarios donde participan especialistas del Departamento de Tecnología Informática y Computación (DTIC) de la Universidad de Alicante —GrupoM. Redes y Middleware—, el grupo de Bases de Datos del Departamento de Lenguajes y Sistemas Informáticos e Ingeniería del Software (DLSIIS) de la Universidad Politécnica de Madrid y el Departamento de Ciencia de la Computación de la Universidad de La Habana.
- En el marco de esta colaboración, he asumido el rol de coordinador del programa de doctorado conjunto: "Tecnologías de la sociedad de la Información" por la Universidad de La Habana (UH). Gracias a esta

responsabilidad he tenido la oportunidad de ayudar a fomentar el desarrollo científico-técnico de los profesores jóvenes del departamento de Ciencia de la Computación de la UH mediante la posibilidad real que supone el referido programa de doctorado para la obtención, por parte de ellos, del grado científico de Doctor. Lo acabado de expresar, además, tributa a la elevación del nivel profesional del claustro de profesores de la Facultad de Matemática y Computación de la mencionada universidad.

Universitat d'Alacant
Universidad de Alicante

Referencias Bibliográficas

(Aggarwal & Yu, 2001) Aggarwal, C. C. & Yu, P. S. (2001). *Outlier* Detection for High Dimensional Data. Proceedings of the 2001 ACM SIGMOD Conference on Management of Data, pp. 37-46.

(Aggarwal & Yu, 2005) Aggarwal, C. C. & Yu, P. S. (2005). An efficient and effective algorithm for high dimensional *outlier* detection. VLDB Journal, 14(2), pp. 211–221.

(Aggarwal, 2007) Aggarwal, C. C. (2007). Towards Exploratory Test Instance Centered Diagnosis in High Dimensional Classification. IEEE Transactions on Knowledge and Data Engineering, 19(8), pp. 1001-1015.

(Aggarwal & Yu, 2008) Aggarwal, C. C. & Yu, P. S. (2008). *Outlier* Detection with Uncertain Data. Proceedings of the 2008 SIAM Conference on Data Mining, pp. 483-493.

(Angiulli & Pizzuti, 2002) Angiulli, F. & Pizzuti, C. (2002). Fast *outlier* detection in high dimensional spaces. Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery PKDD'02, pp.15-26.

(Angiulli & Pizzuti, 2005) Angiulli, F. & Pizzuti, C. (2005). *Outlier* Mining in Large High-Dimensional Data Sets. IEEE Transactions on Knowledge and

Universitat d'Alacant
Universidad de Alicante

Data Engineering, Vol. 17, Art. No. 2, pp. 203-215.

(Angiulli *et al.* 2006) Angiulli, F., Basta, S. & Pizzuti, C. (2006). Distance-Based Detection and Prediction of *Outlier*s. IEEE Transactions on Knowledge and Data Engineering, Vol. 18, Art. No. 2, pp. 145-160.

(Angiulli *et al.*, 2007) Angiulli, F., Greco, G. & Palopoli, L. (2007). *Outlier* Detection by Logic Programming. ACM Transaction on Computational Logic (TOCL), Vol. 9, Art. No. 7, ISSN: 1529-3785.

(Barnett & Lewis, 1994) Barnett, V. & Lewis, T. (1994). *Outlier*s in Statistical Data, 3rd edn., John Wiley & Sons. Chichester, ISBN 0-471-93094-6.

(Bay & Schwabacher, 2003) Bay, S. D. & Schwabacher, M. (2003). Mining distance-based *outlier*s in near lineal time with randomization and a simple pruning rule. Proceedings of the 9th annual ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

(Ben-Gal, 2005) Ben-Gal, I. (2005). *Outlier*s Detection. The Data Mining and Knowledge Discovery Handbook, Chap.7, Springer 2005, ISBN 0-387-24435-2.

(Beynon & Driffield, 2005) Beynon, M. J., & Driffield, N. (2005). An illustration of variable precision rough sets model: an analysis of the findings of the UK Monopolies and Mergers Commission. Computers and Operations Research, Vol. 32, Issue 7, pp. 1739-1759.

(Beynon, 2006) Beynon, M. J. (2006). An introduction of the condition class space with continuous value discretization and rough set theory. Wiley InterScience; International Journal of Intelligent Systems , Vol. 21, Issue 2, pp. 173-191.

(Bing-Zhen *et al.*, 2004) Bing-Zhen, S., Ya-Bin, S., Zeng-Tai, G., De-Gang, C., & Qiang, H. (2004). Variable Precision Rough Set model based on general relation. Proceedings of the Third International Conference on Machine Learning and Cybernetics. Shanghai, 2004.

(Bishop, 1994) Bishop, C. M. (1994). Novelty detection and Neural Network validation. Proceedings of the IEEE Conference on Vision, Image and Signal Processing, pp. 217–222.

Universidad de Alicante

(Bouyer *et al.*, 2009) Bouyer, A., Abdullah, A. H., Ebrahimpour, H., & Nasrollahi, F. (2009). Fault-Tolerance Scheduling by Using Rough Set Based Multi-checkpointing on Economic Grids. 2009 International Conference on Computational Science and Engineering, Vol. 1, pp.103-109.

(Breunig *et al.*, 1999) Breunig, M. M., Kriegel, H-P., Ng, R. T., & Sander, J. (1999). OPTICS-OF: Identifying Local *Outlier*s. Principles of Data Mining and Knowledge Discovery, pp. 262–270.

(Breunig *et al.*, 2000) Breunig, M. M., Kriegel, H-P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local *outlier*s. SIGMOD, Rec. 29, No. 2, ISSN 0163-5808, pp. 93–104.

(Cao *et al.*, 2003) Cao, L. J., Lee, H.P., & Chong, W. K. (2003). Modified support vector novelty detector using training data with *outlier*s. Pattern Recognition Letters, (24), No.14, pp. 2479-2487.

(Chawla & Sun, 2006) Chawla, S., & Sun, P. (2006). *Outlier* Detection: Principles, Techniques and Application. Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore, 2006.

(Cherednichenko, 2005) Cherednichenko, S. (2005). *Outlier* Detection in Clustering. University of Joensuu, Departament of Computer Science, Master's Thesis.

(Chiu & Fu, 2003) Chiu, A. L., & Fu, A. W. (2003). Enhancements on local *outlier* detection. Proceedings of the IDEAS´03.

(Cramer *et al.*, 2004) Cramer J, A., Shah S, S., Battaglia T, M., Banerji S, N., Obando L, A., & Booksh K, S. (2004). *Outlier* detection in chemical data by fractal analysis. Journal of Chemometrics; Vol. 18, Issue 7- 8, pp. 317-326.

(Fayyad *et al.*, 1996) Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AAAI Press/The MIT Press.

(Fix & Hodges, 1951) Fix, E., & Hodges, J. (1951). An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation.

Universitat d'Alacant
Universidad de Alicante

lnternational Statistics Review, 57, pp. 233 247.

(Fristedt & Gray, 1996) Fristedt, B. E., Gray, L. F. (1996). A Modern Approach to Probability Theory, Birkhäuser, 1996, Chapter 2.

(Gang, 2009) Gang, F. (2009). Network Teaching Resource Evaluation Method Based on Rough Set Theory. Proceedings of the 2009 International Conference on Management of e-Commerce and e-Government (icmecg'09), pp.188-192.

(Ghoting *et al.*, 2004) Ghoting, A., Otey, M. E., Parthasarathy, S. (2004). LOADED linked-based *outlier* and anomaly detection in evolving data sets. Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04), pp. 387-390.

(Gong *et al.*, 2004) Gong, Z. T., Sun, B. Z., Shao, Y. B., Chen, D. G., & He, Q. (2004). Variable precision rough set model based on general relations. Proceedings of the Third Conference on Machine Learning and Cybernetics, Shanghai'04.

(Graham *et al.*, 2002) Graham, W., Baxter, R. A., He, H. X., Hawkins, S., & Gu, L. (2002). A comparative study of RNN for *Outlier* Detection in Data Mining. IEEE International Conference on Data-Mining (ICDM'02). CSIRO Technical Report CMIS-02/102. Maebashi City, Japan.

(Grubbs, 1969) Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. Technometrics 11, pp. 1–21.

(Hair *et al.*, 1999) Hair, J. F., Anderson, R. E., Tatham, R.L. & Black, W.C. (1999). Análisis multivariante. Prentice Hall. Madrid, 5ª edición.

(Han & Kamber, 2000) Han, J., Kamber, M. (2000). Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems, Jim Gray Series, Morgan Kaufmann Publishers, 550 pp.

(Hautamäki *et al.*, 2005) Hautamäki, V., Cherednichenko, S., Kärkkäinen, I., Kinnunen, T. & Fränti, P. (2005). Improving k-means by *outlier* removal. Proceedings of the 14th Scandinavian Conference on Image Analysis (SCIA'05), Joensuu, Finland, June 2005, pp. 978–987.

Universitat d'Alacant
Universidad de Alicante

(Hawkins, 1980) Hawkins, D. (1980). Identification of *outlier*s. Chapman and Hall, Reading.

(Hawkins *et al.* 2002) Hawkins, S., He, H., Williams, G. J., & Baxter, R. A. (2002). *Outlier* Detection Using Replicator Nueral Networks. Proceedings DaWaK'02, pp. 170-180.

(He *et al.*, 2002) He, Z., Deng, S., & Xu, X. (2002). *Outlier* detection integrating semantic knowledge. Proceedings of the International Conference for Web Information Age WAIM'02. Lecture Notes in Computer Science, Springer.

(He *et al.*, 2003) He, Z., Xu, X., & Deng, S. (2003). Discovering Cluster Based Local *Outlier*s. Pattern Recognition Letters, 24(9-10), pp. 1651-1660.

(He *et al.*, 2004) He, Z., Xu, X., Huang, J. Z., & Deng, S. (2004). Mining class *outlier*: concepts, algorithms and applications in CRM. Expert System with Applications, 27 (2004), pp. 681–697.

(Hodge & Austin, 2004) Hodge, V. J., & Austin, J. (2004). Survey of *Outlier* Detection Methodologies. Artificial Intelligence Review, 22, pp. 85-126.

(Hu & Sung, 2003) Hu, T., & Sung, S. Y. (2003). Detecting pattern-based *outlier*s. Pattern Recognition Letters, 24 (16), pp. 3059-3068.

(Japkowicz *et al.*, 1995) Japkowicz, N., Myers, C., & Gluck M. A. (1995). A Novelty Detection Approach to Classification. Proceedings of the 14th International Conference on Artificial Intelligence (IJCAI-95), pp. 518–523.

(Jhonson *et al.*, 1998) Johnson, T., Kwok, I. & Ng, R. T. (1998). Fast Computation of 2d depth contours. Proceedings of the ACM SIG KDD98, pp. 224-228.

(Jiang *et al.* 2001) Jiang, F., Tseng, S. S., & Su, C. M. (2001). Two-phase Clustering Process for *Outlier*s Detection. Pattern Recognition Letters, 2001, pp. 691-700.

(Jiang *et al.*, 2005) Jiang, F., Sui, Y., & Cao, C. (2005). *Outlier* detection using rough sets theory. Proceedings of the Rough Sets, Fuzzy Sets, Data

Universitat d'Alacant
Universidad de Alicante

Mining, and Granular Computing (RSFDGrC'05). Springer.

(Jiang *et al.* 2006) Jiang, F., Sui, Y., & Cao, C. (2006). *Outlier* detection based on rough membership function. Rough Sets and Current Trends in Computing, 5th International Conference, RSCTC'06. Kobe, Japan: Springer.

(Jin *et al.*, 2001) Jin, W., Tung, A. K., & Han, J. (2001). Mining top-n local *outlier*s in large databases. Proceedings of the KDD'01, pp. 293-298.

(Junding & Suxia, 2009) Junding,S., Suxia, Ch. (2009). ROI Extraction Based on Rough Set. Proceedings of the International Conference on Environmental Science and Information Application Technology, ESIAT'09, Vol. 3, pp.207-209.

(Last & Kandel, 2001) Kandel, A., & Last, M. (2001). Automated detection of *outlier*s in real-world data. Proceedings of the Second International Conference on Intelligent Technologies. Bangkok, Thailand.

(Kaufman & Rousseeuw, 1990) Kaufman, L., & Rousseeuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley.

(Knorr & Ng, 1997) Knorr, E., & Ng, R. (1997). A unified notion of *outlier*s: Properties and computation. Proceedings of the KDD'97, pp. 219-222.

(Knorr & Ng, 1998) Knorr, E., & Ng, R. (1998). Algorithms for mining distance-based *outlier*s in large datasets. Proceedings of the 24th Int. Conf. on Very Large Database VLDB'98, New York, pp. 392-403.

(Knorr & Ng, 1999) Knorr, E., & Ng, R. (1999). Finding intentional knowledge of distance-based *outlier*s. In Proceedings of the VLDB'99, pp. 211-222.

(Knorr & Ng, 2000) Knorr, E., Ng, R., & Tucakov, T. (2000). Distance-based *outlier*s: Algorithms and Applications. VLDB Journal, 8 (3 and 4), pp. 237-253.

(Kohonen, 1995) Kohonen, T. (1995). Self-Organizing Maps. Springer Series in Information Sciences, Vol. 30, Springer, Heidelberg.

(Kohonen, 1996) Kohonen, T. (1996). Self-organizing maps of symbol strings.

Universitat d'Alacant
Universidad de Alicante

Report A42, Helsinki University of Tech-nology, Laboratory of Computer and Information Science, Espoo, Finland.

(Kollios *et al.*, 2003) Kollios, G., Gunopulus, D., Koudas, N., & Berchtold, S. (2003). Efficient biased sampling for approximate clustering and *outlier* detection in large datasets. IEEE transactions on Knowledge and Data Engineering, Sep. 2003, pp 1170-1187.

(Last & Kandel, 2001) Last, M., & Kandel, A. (2001). Automated Detection of *Outlier*s in Real-World Data. Proceedings of the Second International Conference on Intelligent Technologies, Bangkok, Thailand, pp. 292-301.

(Lazarevic *et al.*, 2003) Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A., & Srivasta., J. (2003). A comparative study of anomaly detection schemes in network intrusion detection. Proceedings of the Third SIAM Conference on Data Mining.

(Li *et al.*, 2006) Li, S., Lee, R., & Lang, S.-D. (2006). Detecting *outlier*s in interval data. Proceedings of the ACM Southeast Regional Conference, pp. 290-295.

(Lin & Brown, 2001) Lin, S., & Brown, D. E. (2001). *Outlier*-based Data Association:Combining OLAP and Data Mining. Technical Report. Department of Systems Engineering University of Virginia.

(Liu *et al.*, 2009) Liu, H., Kong, W., Qiu, T.-S., Li, G.-L. (2009). A Neural Network Based on Rough Set (RSNN) for Prediction of Solitary Pulmonary Nodules. Proceedings of the International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, ijcbs'09, pp.135-138.

(Lu *et al.*, 2003) Lu, C., Chen, D., & Kou, Y. (2003). Algorithms for spatial *outlier* detection. in Proceedings of the 3rd IEEE International Conference on Data-mining (ICDM' 03). Melbourne, FL.

(Maheswari *et al.*, 2001) Maheswari, V. U., Siromoney, A., Mehata, K.M. (2001). The Variable Precision Rough Set Model for Web Usage Mining. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Vol. 2198/2001. Web Intelligence Research and Development, pp. 520-524.

Universitat d'Alacant

(Markus *et al.*, 2000) Markus, M. B., Hans-Peter, K., Raymond, T. N., & Jorg, S. (2000). LOF: Identifying density-based local *outlier*s. Proceedings of the International Conference on Management of Data ACM SIGMOD Record, Vol. 29, No. 2, pp.93-104.

(Martin *et al.*, 1996) Martin, E., Kriegel, H-P., Sander, J., Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, Portland, Oregon, pp. 226–231.

(Nairac *et al.*, 1999) Nairac, A., Townsend, N., Carr, R., King, S., Cowley, P., & Tarassenko, L. (1999). A system for the analysis of jet system vibration data. Integrated ComputerAided Engineering 6(1), pp. 53 – 65.

(Nguyen *et al.*, 2006) Nguyen, D. T., Memik, G., Choudhary, A. (2006). A reconfigurable architecture for network intrusion detection using principal component analysis. Proceedings of the 2006 ACM/SIGDA 14th international symposium on field programmable gate arrays. Monterey, California, USA. ISBN: 1-59593-292-5, pp. 235 - 235.

(Otey *et al.*, 2005a) Otey, M. E., Ghoting, A. & Parthasarathy, S. (2005). Fast Distributed *outlier* Detection in mixed-atribute data sets. Technical report OSU-CISRC-6/05-TR42, Department of Computer Science and Engineering. The Ohio State University.

(Otey *et al.*, 2005b) Otey, M. E., Ghoting, A. & Parthasarathy, S. (2005). Fast Ligthweight *outlier* Detection in mixed-atribute data sets. Technical report OSU-CISRC-6/05-TR43, Department of Computer Science and Engineering, The Ohio State University.

(Otey *et al.*, 2005c) Otey, M. E., Parthasarathy, S. & Ghoting, A. (2005). An Empirical Comparison of *Outlier* Detection Algorithms. Proceedings of the International Workshop on Data Mining Methods for Anomaly Detection - KDD 2005, Chicago, Illinois, USA, pp. 45 – 52.

(Papadimitriou *et al.*, 2003) Papadimitriou, S., Kitagawa, H., Gibbons, P. B. & Faloutsos, C. (2003). LOCI: Fast *Outlier* Detection Using the Local Correlation Integral. Proceedings of the ICDE'03, Bangalore, India, March 2003.

Universidad de Alicante

(Pawlak, 1982) Pawlak, Z. (1982). Rough sets. International Journal of Computer and Information Sciences, 11(5), pp. 341–356.

(Pawlak, 1991) Pawlak, Z. (1991). Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht, The Netherlands.

(Penny & Jolliffe, 2001) Penny, K. I. & Jolliffe, I. T. (2001). A comparison of multivariate *outlier* detection methods for clinical laboratory safety data. The Statistician 50 (3), pp. 295-308.

(Petrovsky, 2003) Petrovsky, M. (2003). A Hybrid Method for Patterns Mining and *Outlier*s Detection in the Web Usage Log. Proceedingas of the Atlantic Web Intelligence Conference - AWIC'03, Madrid, Spain, pp. 318-328.

(Qingkui & Junhu, 2009) Qingkui, C. & Junhu, R. (2009). Study on the Supplier Evaluation in Green Supply Chain Based on Rough Sets Theory and Analytic Hierarchy Process. Proceedings of the International Conference on Electronic Commerce and Business Intelligence, ecbi'09, pp.280-283.

(Qinglin *et al.* 2007) Qinglin, G., Kehe, W. & Wei, L. (2007). Fault Forecast and Diagnosis of Steam Turbine Based on Fuzzy Rough Set Theory. Proceedings of the Second International Conference on Innovative Computing, Information and Control - ICICIC'07, Kumamoto, Japan: IEEE Computer Society.

(Raghavan & Sever, 1995) Raghavan, V. V. & Sever, H. (1995). The state of rough sets for database mining applications. In T. Y. Lin, editor, Proceedings of the 23rd Computer Science Conference Workshop on Rough Sets and Database Mining, San Jose State University, San Jose, CA, pp. 1-11.

(Ramaswamy *et al.*, 2000) Ramaswamy, S., Rastogi, R. & Kyuseok, S. (2000). Efficient Algorithms for mining *outlier*s from large data sets. Proceedings of the SIGMOD00, pp. 93 – 104.

(Reif *et al.*, 2008) Reif, M., Goldstein, M., Stahl, A. & Breuel, T. M. (2008). Anomaly Detection by Combining Decision Trees and Parametric Densities Pattern Recognition, 2008. ICPR 2008. 19th International

Universitat d'Alacant
Universidad de Alicante

Conference on BibTeX.

(Ren *et al.*, 2004) Ren, D., Wang, B. & Perrizo, W. (2004). RDF: A density-based *Outlier* Detection Method using Vertical Data Representation. Proceedings of the Fourth IEEE International Conference on Data Mining - ICDM'04, Brighton, UK.

(Rousseeuw & Leroy, 1987) Rousseeuw, P. & Leroy, A. (1987). Robust Regression and *Outlier* Detection. John Wiley & Sons, Inc. New York, NY, USA. 329 pp.

(Ruts & Rousseeuw, 1996) Ruts, I. & Rousseeuw, P. (1996). Computing depth contours of bivariate point clouds. Computational Statistics and Data Analysis, 23, No. 1/ 1996. ISSN 0167-9473, pp. 153-168.

(Saunders & Gero, 2001) Saunders, R. & Gero, J. S. (2001). A Curious Design Agent: A Computational Model of Novelty-Seeking Behaviour in Design. Proceedings of the Sixth Conference on Computer Aided Architectural Design Research in Asia (CAADRIA 2001). Sydney.

(Scholkopf *et al.*, 2001) Scholkopf, B., Platt, J., Shawe-Taylo, J., Smola, A. J. & Williamson, R. (2001). Estimating the support of a high dimensional distribution. Neural Computation, 13(7), pp. 1443-1472.

(Schwabacher & Bay, 2003) Schwabacher, M. & Bay, S. D. (2003). Mining Distance – Based *outlier*s in near linear time with randomization and a simple pruning rule. Proceedings of the 9th annual ACMSIGKDD.

(Shawne-Taylor & Cristianini, 2004) Shawne-Taylor, J. & Cristianini, N. (2004). Kernel Methods for Pattern Analysis. Cambridge University Press, 462 pp.

(Shekhar *et al.*, 2003) Shekhar, S., Lu, C. & Zhang, P. (2003). A unified Approach to Spatial *Outlier*s Detection. GeoInformática, An International Journal on Advances of Computer Science for Gegraphic Information System, No. 7, pp. 139-166.

(Simon *et al.*, 2002) Simon, H., Hongxing, H., Rohan, B. & Graham, W. (2002). *Outlier* Detection Using Replicator Neural Networks. Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery, pp. 170-180.

(Śle¸zak & Ziarko, Śle¸zak D. & Ziarko W. (2002). Bayesian Rough Set Model.

Universidad de Alicante

2002) Proceedings of FDM'2002. December 9, Maebashi, Japan (2002), pp. 131–135.

(Stomoimenova *et al.*, 2005) Stomoimenova, E., Mateev, P. & Dobreva, M. (2005). *Outlier* detection as a method for Knowledge extraction from digital resources. Proceedings of the South-Eastern European Digitalization Initiative (SEEDI) Conference, Ohrid, Macedonia, pp. 503-506.

(Su & Hsu, 2006) Su, C.-T. & Hsu, J.-H. (2006). Precision parameter in the variable precision rough sets model: an application. Int. J. Manage. Sci. Vol. 34, Issue 2, pp. 149-157.

(Sudipto *et al.*, 2000) Sudipto, G., Rajeev, R. & Kyuscok, S. (2000). ROCK: A robust clustering algorithm for categorical attributes. Information Systems, 25(5), pp. 345-366.

(Takeshi & Einoshin, 2002) Takeshi, W. & Einoshin, S. (2002). *Outlier* Detection Based on Decision Tree and Boosting. Proceedings of the Annual Conference of JSAI. Vol. 16th, pp. 1A3.04.1-1A3.04.4.

(Tang *et al.*, 2002) Tang, J., Chen, Z., Fu, A. & Cheung, D. (2002). A Robust *Outlier* Detection Scheme in Large Data Sets. Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Taipei, Taiwan.

(Tao & Pi, 2009) Tao, Y. & Pi, D. (2009). Unifying Density-Based Clustering and *Outlier* Detection. Proceedings of the 2009 Second International Workshop on Knowledge Discovery and Data Mining, Moscow, Russia, ISBN: 978-0-7695-3543-2, pp. 644-647.

(Tax & Duin, 1999) Tax, D. & Duin, R. (1999). Support Vector data description. . Pattern recognition Letters, 20 (11-13), pp. 1191-1199.

(Toth & Gosztolya, 2004) Toth, L. & Gosztolya, G. (2004). Replicator Neural Networks for *Outlier* Modeling in Segmental Speech Recognition. LNCS, Springer Berlin/ Heidelberg, ISSN 0302-9743, Vol. 3173/2004, pp. 996-1001.

(Tukey, 1977) Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley, ISBN 0-201-07616-0.

Universitat d'Alacant

(UCI, 2009) University of California Irvine. Center for Machine Learning and Intelligent Systems: http://cml.ics.uci.edu. Último acceso: 11/09.

(CENSUS, 2009) USA Census Bureau DB: http://www.census.gov. Último acceso: 11/ 09.

(Vesanto *et al.*, 1998) Vesanto, J., Himberg, J., Siponen, M. & Simula, O. (1998). Enhancing SOM Based Data Visualization. Proceedings of the 5th International Conference on Soft Computing and Information/Intelligent Systems. Methodologies for the Conception, Design and Application of Soft Computing, Vol. 1. Singapore: WorldScientific, pp. 64–67.

(Wei *et al.*, 2003) Wei, L., Qian, W., Zhou, A., Jin, W. & Yu, J. X. (2003). Hypergraph-based *outlier* test for categorical data. Proceedings of the PAKDD03, pp. 399-410.

(Walfish, 2006) Walfish, S. 2006). A Review of Statistical *Outlier* Methods. Pharmaceutical Technologies, 30(11), pp. 82-88.

(Xiaowen *et al.*, 2009) Xiaowen, X., Wei, X. & Beirong, Z. (2009). Reconfigurability analysis of manufacturing system based on rough sets. Proceedings of the 2nd IEEE International Conference on Computer Science and Information Technology iccsit09, pp.513-517.

(Xu et. al., 2005) Xu, X., He, Z. & Deng, S. (2005). An optimization model for *outlier* detection in Categorical Data. Proceedings of the International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I, Springer Berlin / Heidelberg, pp. 400-409.

(Yamanishi *et al.*, 2000) Yamanishi, K., Takeuchi, J., & Williams, G. (2000). On-line Unsupervised *Outlier* detection using finite mixtures with discounting Learning Algorithms. Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'00, pp. 320-325.

(Yamanishi & Takeuchi, 2001) Yamanishi, K. & Takeuchi, J. (2001). Discovering *outlier* filtering rules from unlabeled data- combining a superviser learner with am unsuperviser learner. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data

Universitat d'Alacant

Mining - KDD´01, pp. 389-394.

(Yao, 2009) Yao, P. (2009). Hybrid Classifier Using Neighborhood Rough Set and SVM for Credit Scoring. Proceedings of the International Conference on Business Intelligence and Financial Engineering, bife, pp.138-142.

(Yin *et al.*, 2009) Yin, L., Liu, X., Jiang W. & Xie, J. (2009). Evaluation of Enterprise Innovation Ability Based on Rough Set Theory. Proceedings of the Asia-Pacific Conference on Information Processing, apcip, Vol. 1, pp. 471-475.

(Yu *et al.*, 2002) Yu, D., Sheikholeslami, G. & Zhang, A. (2002). FindOut: Finding Out *Outlier*s in Large Datasets. Knowledge and Information Systems, Springer London, Vol. 4, No. 4, pp. 387-412.

(Wu *et al.*, 2009) Wu, Ch., Lei, H., Ma, M. & Yan, X. (2009). Severity Analyses of Single-Vehicle Crashes Based on Rough Set Theory. Proceedings of the International Conference on Computational Intelligence and Natural Computing, cinc, Vol. 2, pp. 59-62.

(Zeng *et al.*, 2009) Zeng, F., Yin, K., Chen, M. & Wang, X. (2009). A New Anomaly Detection Method Based on Rough Set Reduction and HMM. Proceedings of the Eigth IEEE/ACIS International Conference on Computer and Information Science, icis 2009, pp. 285-289.

(Zengyou *et al.*, 2005) Zengyou, H., Xiaofei, X., Joshua Zhexue, H., & Shengchun, D. (2005). FP-*Outlier*: Frequent Pattern Based *Outlier* Detection. Computer Science and Information Systems, Publisher ComSIS Consortium, Vol. 2 , Issue 01, pp. 103-118.

(Zhao *et al.*, 2009) Zhao, M., Liu, H., Abraham, A. & Corchado,E. (2009). A Swarm-Based Rough Set Approach for Group Decision Support Systems. Proceedings of the Ninth International Conference on Hybrid Intelligent Systems, his, Vol. 3, pp. 365-369.

(Zhong *et al.*, 2009) Zhong, C., Lin, X., Zhang, M. (2009). A Local *Outlier* Detection Approach Based on Graph-Cut. Proceedings of the International Joint Conference on Computational Sciences and Optimization, Sanya, Hainan, China. ISBN: 978-0-7695-3605-7.

(Zhu *et al.*, 2005) Zhu, C., Kitagawa, H., Papadimitriou, S., & Faloutsos, C. (2005).

Example-based *Outlier* Detection with Relevance feedback. The Database Society of Japan (DBSJ) Letters, Vol. 3, No. 2.

(Ziarko, 1993) Ziarko, W. (1993). Variable precision rough set model. Journal of Computer and System Sciences, 46(1), pp. 39–59.

(Ziarko, 1994) Ziarko, W. (1994). Rough Sets, Fuzzy Sets and Knowledge Discovery, Springer Verlag, pp. 326-334.

(Ziarko, 2001) Ziarko, W. (2001). Probabilistic Decision tables in the Variable Precision Rough Set Model. Computational Intelligence: an International Journal, Vol. 17 (2001), No. 3, pp. 593–603.

(Ziarko, 2002) Ziarko, W. (2002). Set approximation quality measures in the variable precision rough set model. Proceedings of the 2nd Int. Conf. on Hybrid Intelligent Systems (HIS'02), Soft Computing Systems 87 (2002), pp. 442–452.

Universitat d'Alacant
Universidad de Alicante