Actas de las Jenui, vol. 2. 2017.
Páginas: 229-236

Objetivando el proceso de revisión de JENUI

Rosana Satorre-Cuerda, Patricia Compañ-Rosique, Rafael Molina-Carmona, Faraón Llorens-Largo
Departamento de Ciencia de la Computación e Inteligencia Artificial
Universidad de Alicante
Alicante
[rosana.satorre, patricia.company, rmolina, faraon.llorens]@ua.es

Resumen

La revisión por pares es el principal método de selección de artículos en congresos y revistas. Aunque es un proceso ampliamente aceptado diversos autores han detectado posibles problemas de validez y fiabilidad. Estos problemas se pueden atenuar tratando de objetivar las revisiones. En esta investigación se plantea el proceso de revisión por pares como un problema de clasificación, en el que se dispone de varios clasificadores (los revisores) que, ante una entrada (el trabajo a revisar), deben asignarle una determinada clase (aceptación o rechazo). Se proponen dos métricas: la tasa de aciertos y el grado de acuerdo, que se corresponden con los conceptos de validez y fiabilidad, pero adaptados a cada revisor por separado. Como caso práctico se analizan las revisiones de las últimas cinco JENUI. Por último, se plantean una serie de pautas para tratar de objetivar y mejorar el proceso de selección resultado de la revisión por pares.

Abstract

Peer review is the main method of selecting articles in conferences and journals. Although it is a widely accepted process, some authors have detected potential problems of validity and reliability. These problems can be mitigated by trying to objectify the reviews. In this research, the peer review process is considered as a classification problem, in which several classifiers (reviewers), before an input (the work to be reviewed), must assign a given class (accept or reject). Two metrics are proposed: the success rate and the agreement degree, which correspond to the concepts of validity and reliability, but adapted to each reviewer separately. As a practical case the reviews of the last five JENUI are analysed. Finally, some guidelines are proposed to try to objectify and improve the selection process resulting from the peer review.

Palabras clave

Revisión por pares, Tasa de acierto, Grado de acuerdo, Validez, Fiabilidad, Matriz de confusión, Clasificación automática.

1. Introducción

La revisión por pares es la forma más aceptada en la comunidad científica para diversos procesos de evaluación y valoración de la calidad de un trabajo [3]. Algunos ejemplos son la evaluación de proyectos y la revisión de artículos en revistas y conferencias. En general se considera un método fiable y eficaz [8]. Tanto es así que sin este tipo de revisión la calidad de muchos trabajos se pone en duda. En definitiva, los investigadores estamos dejando la evaluación de nuestros trabajos y, por consiguiente, el futuro de la ciencia, en manos de nuestros compañeros, expertos en cada una de nuestras áreas de investigación. En principio, parece razonable que sean nuestros colegas, tan expertos como nosotros en nuestra disciplina, quienes decidan si lo que estamos investigando y publicando supone un avance significativo para la ciencia. Es lo que podemos denominar control mutuo, puesto que nosotros mismos nos convertiremos en evaluadores de los trabajos de nuestros colegas.

Se ha destacado mucho la ventaja que supone el control mutuo, pero la práctica nos dice que muchas veces la revisión por pares como método de selección presenta algunos efectos perjudiciales [10]. Todos nos hemos encontrado con revisiones de nuestros trabajos que no son todo los rigurosas que esperábamos. Y nos surgen ciertas dudas: ¿son los revisores tan expertos como esperábamos? ¿se toman todos los revisores sus evaluaciones tan en serio como cabría esperar? ¿detectamos en las revisiones sesgos que no deberían existir? ¿están siendo los revisores realmente objetivos?

Hay una buena cantidad de trabajos que indican que estas cuestiones están despertando interés en la comunidad científica. Por ejemplo, el trabajo de Mulligan, Hall y Raphael [8] es un estudio

internacional a gran escala que recaba la opinión de 4000 investigadores sobre las bondades y problemas de la revisión por pares a través de una gran encuesta online. El estudio establece que este tipo de revisión es aceptada de forma abrumadora entre los investigadores como la forma esencial de velar por la calidad de las publicaciones científicas. Nueve de cada diez creen que además les ayudó a mejorar significativamente sus artículos. Sin embargo, también se detectan algunos problemas en el proceso. Se cree que se pueden introducir ciertos sesgos en las revisiones y que en algunos casos se producen favoritismos. Las soluciones a estos problemas pasan por introducir sistemas de revisión ciegos, tanto para el autor como para el revisor, y por la formación adecuada de los revisores de forma previa a su introducción como evaluadores en el sistema. También se apunta que las nuevas tecnologías pueden aportar avances para hacer el proceso de revisión por pares más efectivo y seguro.

Además de los posibles sesgos que pueden introducir los revisores, dos son las críticas fundamentales que se le hace a la revisión por pares. La primera es la posible falta de fiabilidad (se suele utilizar el término inglés *reliability*), que se produce cuando los diversos revisores de un mismo trabajo no coinciden en su evaluación y recomendación para un artículo, y la segunda es la falta de validez (*validity* en ingés), referida a la frecuente falta de relación entre los resultados de la revisión y los valores reales y objetivos de calidad.

Aunque hay algunos trabajos que proponen estudios de fiabilidad y validez, Topping [12] señala que en la mayoría de los casos se comparan las revisiones entre pares con evaluaciones hechas por profesionales y no con las de otros revisores. Falchikov y Goldinch [4] presentan un meta-análisis de los trabajos de varios autores, y obtienen conclusiones sobre los campos y niveles en los que la fiabilidad y validez es mayor, así como un conjunto de recomendaciones para los profesionales sobre la implementación de la revisión por pares basada en las conclusiones de este meta-análisis.

Se han utilizado diferentes medidas para calcular la fiabilidad y la validez. La medida más común es el coeficiente de correlación entre las evaluaciones de los revisores y la del editor [1]. Otras medidas no tan comunes son la proporción de revisores que dan un valor de evaluación en un rango de confianza con respecto a la del editor, el uso de un test T para comparar la evaluación media de los revisores y del editor, y el análisis de varianza (ANOVA) para determinar la fiabilidad entre los revisores [5, 7]. Bornmann, Mutz y Daniel [2] hacen un completo meta-análisis sobre diferentes estudios que tratan de calcular la fiabilidad y sobre los métodos estadísticos utilizados.

Todas estos trabajos se centran en casos específicos, tratando de obtener la fiabilidad y la validez de las evaluaciones en condiciones muy concretas. Además, las estadísticas que se obtienen son globales, es decir, dan una medida de la fiabilidad o validez de todo el proceso, pero no permiten un estudio individual del grado de acuerdo de cada revisión con la de los otros colegas o con la decisión del editor. Finalmente, la medida estadística más habitual en estos estudios es el coeficiente de correlación.

La principal contribución de este trabajo es proponer una analogía entre el proceso de revisión por pares y la utilización de clasificadores automáticos. Esto nos lleva a plantear un enfoque radicalmente diferente, que permite un estudio caso por caso y, por lo tanto, no sólo se obtiene la evaluación del trabajo revisado, sino también una medida de la fiabilidad y validez de la evaluación del revisor de ese trabajo. Se realiza, por lo tanto, una doble evaluación: la evaluación del trabajo y la evaluación del proceso de revisión.

Este trabajo pretende contribuir en la mejora de la calidad de los procesos de revisión por pares para proporcionar una valoración más objetiva del trabajo de los revisores. En el siguiente apartado presentamos nuestra propuesta de métricas para medir la calidad de las revisiones. El apartado 3 está dedicado a analizar el caso concreto del proceso de revisión de las JENUI, lo que nos proporciona una forma de caracterizar diferentes perfiles de revisores y proponer en el apartado 4 un conjunto de pautas para la mejora del proceso de selección de artículos. Por último, en el apartado 5, se presentan las conclusiones y posibles líneas de investigación futuras.

2. Métricas para medir la calidad de la revisión por pares

2.1. La revisión por pares y los clasificadores automáticos

En la revisión por pares, cada revisor debe evaluar el trabajo de sus colegas, asignando una calificación o una categoría. En otras palabras, a partir del trabajo a evaluar, el revisor produce un resultado en forma de clasificación. Para hacer esta clasificación, los revisores deben tener un conjunto de criterios (por ejemplo, una rúbrica) que les permitan llevar a cabo su tarea de la manera más objetiva posible.

Se puede establecer una analogía entre este proceso de revisión y el proceso que realizan los clasificadores automáticos. En un trabajo anterior ya establecimos esta analogía [9]. Un clasificador automático es un modelo computacional que asigna a un individuo, caracterizado por un conjunto de variables, una etiqueta entre varias posibles etiquetas asociadas con diferentes clases. El algoritmo utilizado para la clasi-

ficación establece los criterios para realizar esta asignación. Más allá de las diferencias obvias entre los dos procesos, ambos parten de un individuo que debe ser clasificado. En el caso de un clasificador automático, el individuo se caracteriza por un conjunto de variables de diferente tipo que se pueden manejar automáticamente. En el caso de un evaluador humano, el elemento que caracteriza al individuo es el trabajo a revisar, por lo que la información disponible es mucho más rica pero menos estructurada y más difícil de automatizar. En cualquier caso, y siguiendo con la analogía, en ambos casos a partir de estas entradas se debe aplicar un algoritmo de clasificación, basado en métodos computacionales en un caso, y basado en una rúbrica y una tarea subjetiva de aplicar esta rúbrica en el otro. Como resultado del algoritmo, en ambos casos se emite una etiqueta que identifica la clase en la que el individuo está clasificado.

La cuestión clave es, en todo caso, cuál es la calidad de la clasificación. En el caso de modelos informáticos, los investigadores han dedicado mucho esfuerzo a buscar la manera de comparar los clasificadores en función de los éxitos y fracasos que ocurren en la clasificación. Dado que este tipo de métricas se basa sólo en los resultados pero no en las características técnicas del algoritmo, ¿sería posible aplicarlo en el caso de una clasificación humana? Esta es la hipótesis de partida este trabajo.

Dos de las medidas más simples y habituales para evaluar la calidad de un clasificador son su tasa de éxito (exactitud[1]) y su tasa de error [11]. En realidad son medidas complementarias ya que exactitud=1-tasa de error.

Aunque la exactitud es un indicador muy popular y tiene la virtud de representar una medida de la calidad de un clasificador a través de un único valor, tiene el inconveniente de que presupone que el coste de una clasificación errónea es siempre el mismo, independientemente del error que se produzca. Asumir esto puede no ser aceptable. Tomemos un ejemplo para explicar el problema: Supongamos un clasificador que hace un diagnóstico de una enfermedad, es decir, a partir de un conjunto de valores relacionados con pruebas diagnósticas o determinados síntomas, clasifica a los pacientes indicando si se ven afectados o no por la enfermedad. El clasificador tiene una tasa de éxito del 95%, es decir, falla sólo en el 5% de los casos. La pregunta es: estas malas clasificaciones ¿se refieren a los pacientes que tienen la enfermedad, pero se clasifican como sanos, o a los pacientes sanos que se clasifican como enfermos? Obviamente, el coste de la clasificación errónea no puede ser el mismo, ya que en este caso un clasificador conservador que clasifique a todos los pacientes como enfermos incluso a costa de empeorar la exactitud es preferible a un clasificador más exacto pero que clasifica a algunos pacientes enfermos como sanos [6, 13].

Para completar la información escasa que proporciona la medida de exactitud, se han presentado otras medidas más completas que analizan otros aspectos de los clasificadores. Supongamos el caso más simple de un clasificador binario. Formalmente, para cada individuo o muestra, el clasificador adjudica al individuo una etiqueta *P* (clase positiva) o una etiqueta *N* (clase negativa). Supongamos que se conoce la clase real a la que pertenece el individuo, y se etiqueta como *p* o *n*, para distinguir las clases reales (minúsculas) de las estimadas por el clasificador (mayúsculas). Puede haber cuatro resultados posibles:

- La muestra es positiva (*p*) y se clasifica como positiva (*P*). Entonces estamos ante un verdadero positivo (VP).
- La muestra es positiva (*p*) y se clasifica como negativa (*N*). Entonces estamos ante un falso negativo (FN).
- La muestra es negativa (*n*) y se clasifica como negativa (*N*). Entonces estamos ante un verdadero negativo (VN).
- La muestra es negativa (*n*) y se clasifica como positiva (*P*). Entonces estamos ante un falso positivo (FP).

Estos resultados se representan habitualmente en una matriz de confusión o tabla de contingencia (Cuadro 1). Esta matriz es la base de muchas métricas comunes y puede extenderse a más de dos clases.

	p	*n*
P	VP	FP
N	FN	VN

Cuadro 1. Matriz de confusión para un clasificador binario

La diagonal principal muestra el número de clasificaciones correctas (verdaderos positivos y negativos), y la anti-diagonal muestra el número de errores o confusiones (falsos positivos y negativos). Se pueden calcular varias métricas a partir de la matriz de confusión:

- Exactitud = (VP+VN) / (VP+VN+FP+FN)
- Precisión = VP / (VP+FP)
- Sensibilidad o tasa de verdaderos positivos = VP / (VP+FN)
- Especificidad o tasa de negativos verdaderos = VN / (FP+VN)
- *Fall-out* o tasa de falsos positivos = FP / (FP+VN) = 1 – especificidad

[1] El término aceptado por la comunidad científica en inglés es *accuracy*. En español aparece traducido como exactitud o como precisión. Se ha elegido exactitud para distinguirlo de otra medida denominada *precision*, traducida, esta sí, como precisión.

- Tasa de error o tasa de falsos negativos = FN / (VP+FN)
- *F-score* = 2 / ((1/precisión) + (1/sensibilidad))

En el caso de clasificadores multiclase, sólo es posible distinguir entre positivos verdaderos, falsos positivos y falsos negativos. Es posible construir una matriz de confusión para un clasificador multiclase como una matriz cuadrada de tamaño $n \times n$, donde n es el número de clases y definir las anteriores métricas de forma análoga [11].

2.2. Propuesta de métricas para evaluar la revisión por pares

Una vez establecido el paralelismo entre el proceso de revisión y el de clasificación, se pueden diseñar medidas para evaluar la bondad del proceso de revisión por pares. El punto de partida es la matriz de confusión de cada revisor.

En primer lugar, definimos los siguientes conjuntos y elementos:

- Sea **A** el conjunto de artículos a revisar y a_i cada uno de esos artículos.
- Sea **R** el conjunto de revisores que evalúan los artículos y r_j cada uno de los revisores.
- Sea **C** el conjunto de clasificaciones canónicas de los artículos y c_i la clasificación canónica correspondiente al artículo i, es decir, es la clase a la que consideramos que pertenece el artículo a_i (por ejemplo, la opinión fundada de los expertos).
- Sea **E** el conjunto de evaluaciones y e_{ij} la evaluación que del artículo a_i hace el revisor r_j, es decir, e_{ij} es la clase a la que pertenece el artículo a_i según el revisor r_j.
- Sea n_j el número de revisiones realizadas por cada revisor r_j.

Proponemos dos métricas, ambas referidas a cada uno de los revisores: la tasas de aciertos (*TA*) y el grado de acuerdo (*GA*).

Tasa de aciertos

La tasa de aciertos del revisor r_j, TA_j, se define de la siguiente manera:

$$TA_j = \frac{\sum_{\forall\, e_{ij}} S_{ij}}{n_j}$$

donde

$$S_{ij} = \begin{cases} 1, & si\ e_{ij} = c_i \\ 0, & en\ otro\ caso \end{cases}$$

es decir, S_{ij} vale 1 si la evaluación del artículo a_i hecha por el revisor r_j (es decir, e_{ij}) coincide con la clasificación canónica c_i y, por lo tanto, TA_j es la proporción de revisiones en las que el revisor coincide con la clasificación canónica. Esta medida puede tomar valores en el intervalo [0,1], de manera que TA_j=1 indica completo acierto, es decir, el revisor clasifica todos los trabajos correctamente y no hay falsos positivos ni falsos negativos. Un valor de TA_j=0 indica que el revisor no ha coincidido nunca con la clasificación canónica.

Esta medida tiene un significado similar al de la validez, con la diferencia de que la validez es una medida global para todas las revisiones, y esta tasa es una medida particular para cada revisor. En el caso de un clasificador binario, la TA coincide con la exactitud.

Grado de acuerdo

El grado de acuerdo del revisor r_j con el resto de revisores, GA_j, se define a través de la siguiente ecuación:

$$GA_j = \frac{\sum_{\forall\, e_{ij}, e_{ik}} L_{ijk}}{n_j}$$

donde

$$L_{ijk} = \begin{cases} 1, & si\ e_{ij} = e_{ik} \\ 0, & en\ otro\ caso \end{cases}$$

es decir, L_{ijk} vale 1 si la evaluación del artículo a_i hecha por el revisor r_j (es decir, e_{ij}) coincide con la revisión del mismo artículo hecha por el revisor r_k (es decir, e_{ik}) y, por lo tanto, GA_j mide la proporción de revisiones en las que un revisor coincide con el resto de revisores del mismo artículo. Como en el caso anterior, esta medida puede tomar valores en el intervalo [0,1], de manera que si GA_j=1 indica completo acuerdo con el resto de revisores de ese artículo y si GA_j=0 indica que el revisor no ha coincidido nunca con sus colegas en las revisiones que ha realizado.

En este caso, el grado de acuerdo establece una medida similar a la de la fiabilidad pero, como en el caso anterior, se trata de una medida individual para cada revisor y no global.

3. Caso de estudio: JENUI

Como caso de estudio hemos elegido el proceso de revisión que se realiza en las Jornadas sobre la Enseñanza Universitaria de la Informática (JENUI), en las que se realiza una revisión por pares de los artículos enviados para su participación en el congreso. Disponemos de los datos correspondientes a 5 jornadas, desde el año 2012 al año 2016. En total, se han realizado 1588 revisiones, correspondientes a 379 documentos distintos, realizadas por 201 revisores diferentes. En el proceso de revisión de las JENUI, los artículos son revisados, de forma doblemente ciega (los revisores desconocen quiénes son los autores de los artículos que revisan, y los autores desconocen qué revisores evalúan sus trabajos) por al menos 3 revisores (en dos casos excepcionales sólo hay 2

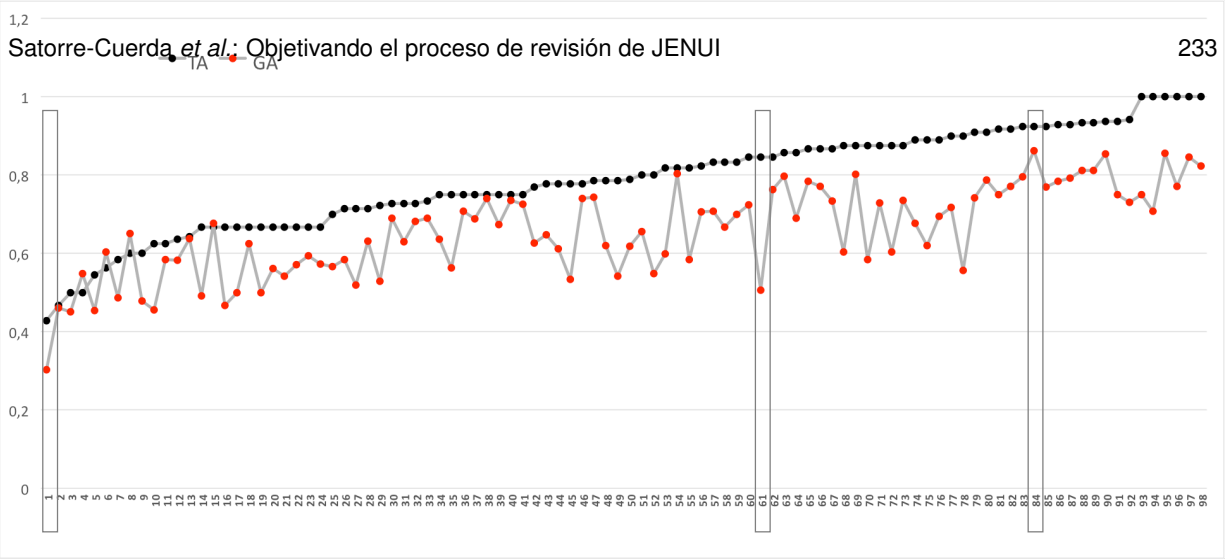

Figura 1. Valores de TA y GA para los 98 revisores que han realizado al menos 8 revisiones

revisiones), aunque de media cada artículo tiene 4,19 revisiones. Cada revisor clasifica el trabajo que ha revisado en una de las siguientes posibles clases: "aceptar", "probablemente aceptar", "probablemente no aceptar" y "no aceptar". El resultado final de la revisión (decisión del editor) es categórica y binaria, es decir, el resultado de la evaluación es "aceptar" o "no aceptar". El número de revisiones realizadas por cada evaluador es bastante diverso, pues depende del número de ediciones en las que ha participado y de los encargos que haya recibido en cada una. En todo caso, los revisores con menos evaluaciones han realizado 1 única revisión, y los que más han participado han llevado a cabo 20 revisiones. De media, cada revisor ha realizado 7,90 evaluaciones.

Uno de los aspectos más interesantes es la elección de la clasificación canónica. Esta clasificación es, por decirlo de alguna manera, la clasificación que cada trabajo debería obtener si fuéramos capaces de evaluarlo objetivamente y obtener una medida de su calidad intrínseca. Esta medida es en general muy difícil de lograr, si no imposible. Una opción, que es la que hemos considerado en este trabajo, es adjudicar a la decisión del editor el valor de clasificación canónica. Evidentemente la decisión del editor se toma en función de las revisiones, que son subjetivas, pero podemos asumir que el trabajo del editor es precisamente tomar una decisión pseudo-objetiva (o, al menos, con más elementos de juicio) a partir de evaluaciones subjetivas.

Con esta clasificación canónica ya podemos construir las matrices de confusión para cada revisor y calcular todas la métricas propuestas. Puesto que la clasificación final es binaria, se han asociado las clases "aceptar" y "probablemente aceptar" con el resultado final "aceptar", y las clases "probablemente no aceptar" y "no aceptar" con el resultado final "no aceptar". En la figura 1 presentamos los valores de *TA* y *GA* para 98 de los revisores participantes, aquellos cuyo número de revisiones está por encima de la media (8 o más revisiones). Se presentan estos resultados por dos razones: en primer lugar para simplificar la visualización de los resultados en la figura, y en segundo lugar porque de esta manera consideramos los revisores con más experiencia. Debemos añadir, además, que aunque el número de revisiones que realiza cada revisor es diferente, la formulación de las métricas está normalizada a valores entre 0 y 1 lo que permite la comparación entre ellos. Los valores de *TA* y *GA* se representan en el eje vertical y los identificadores de cada revisor en el eje horizontal. Para garantizar el anonimato, estos identificadores están alterados: los autores de este trabajo no conocemos la identidad de los revisores. Además, les hemos reasignado una etiqueta numérica aleatoria para que el comité de programa tampoco conozca a quién se refiere cada identificador. Por otro lado, los revisores están ordenados por orden creciente de *TA* para que el gráfico resulte más fácil de interpretar.

La figura 1 nos permite visualizar de un golpe de vista los valores de TA y GA para cada revisor y las diferencias de comportamiento entre ellos. Podemos observar una tendencia de incremento del GA conforme aumenta la TA (el coeficiente de correlación entre ambas variables es de 0,78), lo que tiene cierta lógica. Recordemos que el valor de TA se obtiene comparando la evaluación del revisor en cuestión con la clasificación canónica, es decir, con la dada por el editor, que a su vez está basada en las evaluaciones de los demás revisores, aunque pasadas por el filtro del editor que debe contribuir a hacerlas más objetivas. Por lo tanto, es normal que exista cierto grado de concordancia. Sin embargo, aunque la tendencia es esta, podemos observar que existen muchos casos en los que esto no es exactamente así. Consideramos que ambas pueden aportar información relevante.

Entre todos los revisores, vamos a elegir tres revisores tipo y a estudiar con detenimiento los valores de las métricas para ellos y si es posible establecer una explicación a su comportamiento. En la figura 1 los

tres revisores elegidos como representantes de cada tipo están señalados con un recuadro. Podemos destacar tres comportamientos tipo que presentamos a continuación.

Revisores que presentan valores muy altos de TA y de GA

Este es el caso de revisores que se comportan de forma muy similar a la clasificación canónica y que, además, coinciden casi siempre con sus colegas. Tomemos como ejemplo el revisor 84. En el cuadro 2 se presenta la matriz de confusión correspondiente a este revisor.

	p	*n*
P	9	1
N	0	3

Cuadro 2. Matriz de confusión para el revisor 84

Este revisor ha evaluado 13 trabajos, coincidiendo con la clasificación canónica en 12 de ellos, de los que 9 se han aceptado y 3 se han rechazado. Su tasa de aciertos (TA) es muy alta (un valor de 0,92) y su alto grado de acuerdo con los otros revisores (GA) también (un valor de 0,86). Los valores de ambas métricas están relacionados y parecen indicar que este es un revisor con sólidos argumentos y suele comprender perfectamente el alcance de los trabajos que evalúa. Entre las otras métricas, vamos a destacar los siguientes valores:

- Sensibilidad = 1
- Especificidad = 0,75

El alto valor de la sensibilidad indica que el revisor es capaz de distinguir adecuadamente los casos positivos. El alto valor de especificidad indica que también lo es detectando los casos negativos. Combinando ambas métricas, podemos decir que este revisor es muy cuidadoso con los pequeños detalles que permiten la clasificación correcta de los trabajos más valiosos.

Revisores que presentan valores muy bajos de TA y de GA

El caso de estos revisores es todo lo contrario. Por ejemplo, en el cuadro 3 se presentan las métricas para el revisor 1, que puede considerarse como un ejemplo de este tipo de comportamiento.

	p	*n*
P	5	3
N	5	1

Cuadro 3. Matriz de confusión para el revisor 1

Este es un revisor con una tasa de acierto TA relativamente baja (un valor de 0,43) y un grado de acuerdo con los otros revisores también bajo (un valor de 0,30). Estos valores bajos pueden corresponder a un revisor nuevo o poco cuidadoso. Las otras métricas tienen los siguientes valores:

- Sensibilidad = 0,5
- Especificidad = 0,25

El bajo valor de la sensibilidad indica que el revisor no es capaz de identificar las sutiles diferencias entre las clases positivas. Lo mismo ocurre, aunque todavía en mayor medida con las clases negativas, según marca la muy baja especificidad. En definitiva, es un revisor que presta poca atención a los detalles que marcan la diferencia.

Revisores que presentan valores muy dispares de TA y de GA

Mención aparte merecen este tipo de revisores. En el caso de que exista esta disparidad siempre es por valores altos de TA y bajos de GA (en la figura 1 tan solo hay tres revisores entre los que el GA supera a la TA, y en pequeña cantidad), es decir, son revisores que tienden a evaluar canónicamente aunque suelen no coincidir con el resto de revisores. Este es el caso del revisor 61, cuya matriz de confusión se presenta en el cuadro 4.

	p	*n*
P	3	0
N	2	8

Cuadro 4. Matriz de confusión para el revisor 61

En este caso el valor de TA es considerablemente alto (0,85) mientras que el de GA es bastante menor (0,51). Además de los valores de TA y de GA, nos interesa conocer los valores de sensibilidad y especificidad:

- Sensibilidad = 0,6
- Especificidad = 1

Si bien la capacidad del revisor para clasificar las revisiones positivas no es muy alta (su valor de sensibilidad es relativamente bajo), presenta una alta capacidad para identificar las clasificaciones negativas (su especificidad es la mayor posible). Esta capacidad para detectar casos negativos y los valores de TA y GA obtenidos nos llevan a interpretar que se trata de un revisor que se sale de la norma, que es capaz de ver aspectos interesantes que no son obvios para el resto de revisores y, sobre todo, de detectar los casos negativo. Además, su alta TA indica que convence con sus argumentos a los editores, que consideran que deben tener en cuenta su opinión. Son revisores innovadores, con puntos de vista personales pero interesantes, y capaces de argumentar bien sus revisiones para convencer a los editores.

4. Algunas propuestas para objetivar el proceso de revisión

En los apartados anteriores hemos presentado una serie de métricas que nos permiten caracterizar a los revisores según su histórico de comportamiento durante los procesos de revisión. Estas métricas serán útiles si son capaces de proporcionarnos herramientas para la mejora de los procesos de revisión futuros. Desde nuestro punto de vista mejoraremos las revisiones si conseguimos, de alguna manera, objetivar un proceso que es intrínsecamente subjetivo. Desde luego no podremos obtener nunca una evaluación completamente objetiva cuando tratamos con información compleja y no formalizable como un artículo científico, y con unos revisores humanos que, además de analizar la información, tienen una opinión y un punto de vista propios. Sin embargo, sí podemos analizar el comportamiento de los revisores a lo largo del tiempo y tratar de conocer sus sesgos habituales para contrarrestarlos. Las claves que damos a continuación son opinables y mejorables, pero al menos sirven para abrir un debate sobre el proceso de revisión que siguen congresos como las JENUI.

En primer lugar, proponemos utilizar el valor de TA como ponderador de las revisiones. Hasta ahora, las revisiones en JENUI (y en la mayoría de conferencias y otras publicaciones) permiten al revisor introducir en su evaluación un valor de confianza o nivel de experiencia en la disciplina, lo que solemos denominar *expertise* en inglés. En cualquier caso, este valor se suele utilizar para ponderar la revisión final, es decir, las revisiones correspondientes a verdaderos expertos deberían tener un mayor peso en la decisión final que las que provienen de colegas menos entendidos en el tema. Sin embargo, este nivel es una valoración propia que realiza el revisor y, por lo tanto, absolutamente subjetiva. Es decir, ponderamos una evaluación subjetiva (la revisión) utilizando un valor también subjetivo (el nivel de confianza). Desde nuestro punto de vista esta ponderación podría complementarse cor el valor de TA, un índice basado en la experiencia real del revisor en el proceso de revisión, observable de forma externa y que nos da una idea menos sesgada del comportamiento habitual de ese revisor. Con esta propuesta pasamos, por lo tanto, a ponderar una evaluación subjetiva mediante un valor basado en la experiencia y, quizás, más cercano a la objetividad.

Para poder aplicar esta propuesta todos los revisores deben contar con su TA correspondiente. Sin embargo los revisores noveles no tienen un valor inicial con el que ponderar. La propuesta es proporcionar un nivel inicial a todos los revisores y actualizar este valor con cada nueva revisión.

Por otro lado, el valor de GA también puede proporcionar información interesante. Hemos visto que combinado con valores altos de TA caracteriza a un tipo de revisor especialmente interesante para un congreso o publicación. Estos revisores son capaces de detectar diamantes en bruto en contra de la opinión de sus colegas más conservadores. Las revisiones proporcionadas por estos revisores deberían ser examinadas cuidadosamente por parte de los editores y ser tenidas en cuenta más allá de lo que pueda detectarse a primera vista.

5. Conclusiones y trabajos futuros

La revisión por pares se ha convertido en un elemento muy importante en los sistemas de evaluación. En algunos casos complementa otros métodos de medición (por ejemplo, la evaluación entre los estudiantes que normalmente complementa a la del profesor), pero en otros se convierte en el único elemento o al menos el principal del proceso de evaluación. Tal es el caso de revisiones para publicaciones o conferencias o el proceso de revisión de proyectos de investigación para obtener subvenciones. Los beneficios de la revisión por pares se han destacado en muchas áreas, pero en este tipo de evaluación sigue siendo un componente subjetivo inherente a los procesos con intervención humana. Este componente puede ser interesante desde varios puntos de vista, pero debe ser controlado adecuadamente. En resumen, es importante evaluar el propio proceso de evaluación. Esto nos ha llevado a considerar la pregunta clave propuesta al principio del documento: ¿Es posible establecer algunas pautas para objetivar el trabajo de los revisores en un sistema de evaluación por pares?

En este trabajo hemos tratado de responder a esta pregunta, avanzando en la analogía entre el trabajo de un revisor en un proceso de revisión por pares y el funcionamiento de un clasificador automático que planteamos en un trabajo anterior [9]. De esta manera podemos aprovechar las medidas habituales en la evaluación de la calidad de los clasificadores automáticos para establecer la calidad de un proceso de revisión, y así aprovechar el importante trabajo realizado en el área de la clasificación automática para abrir una nueva línea de estudio sobre los sistemas de revisión por pares.

Para ilustrar esta propuesta, se ha analizado el caso de la revisión por pares en la selección de trabajos para las JENUI. Para este caso de clasificación binaria (aceptar/rechazar), además de las matrices de confusión y las métricas habituales, se han descrito dos nuevos indicadores: la tasa de acierto TA (similar al concepto de validez) y el grado de acuerdo entre revisores GA, que tiene un significado similar al concepto de fiabilidad.

A partir de las métricas obtenidas hemos propuesto dos formas de incrementar la objetividad de las

revisiones ponderando las evaluación con el valor de TA y utilizando el GA para detectar revisores con una visión innovadora. No obstante, estas métricas presentan utilidad más allá de la propuesta, por lo que queda abierta la línea de explorar otras formas de objetivar las decisiones a partir de los diferentes indicadores.

Esta experiencia es una mejora del trabajo anterior que era muy preliminar y planteaba diversos interrogantes. Hemos respondido a algunos ellos realizando un estudio más exhaustivo de los datos, incorporando los datos de más ediciones y planteando propuestas para objetivar el proceso de revisión. No obstante, aún quedan muchos caminos a estudiar. Además de explotar la información que nos proporcionan las métricas propuestas, pretendemos aplicar otras métricas comunes en el área de clasificadores automáticos, definir nuestras propias métricas y realizar un estudio sobre el significado exacto de cada indicador. Además, la propuesta de objetivación debería completarse con un análisis más profundo de los diferentes perfiles de revisor y su caracterización.

Otra cuestión interesante es estudiar el concepto de clasificación canónica y tratar de establecer un patrón de comparación más objetivo. Por ejemplo, el uso de medidas bibliográficas (tales como el volumen de citas) puede ser una fuente interesante de indicadores más objetivos del impacto final del artículo, como medida de la calidad del trabajo.

Otra línea interesante es realizar un estudio de la evaluación en el tiempo del comportamiento de los revisores, tratando de determinar si mejora, y si es conveniente, para la calidad de las revisiones, aprovechar la experiencia de los “grandes sabios” o introducir nuevos puntos de vista más frescos en el proceso.

Agradecimientos

Los autores queremos hacer constar nuestro agradecimiento a los responsables de JENUI por poner a nuestra disposición los datos anonimizados de revisión de estas jornadas en los últimos años.

Referencias

[1] Bornmann, L. 2015. Interrater reliability and convergent validity of F1000Prime peer review: nterrater Reliability and Convergent Validity of F1000Prime Peer Review. *Journal of the Association for Information Science and Technology*. 66, 12 (Dec. 2015), 2415–2426.

[2] Bornmann, L., Mutz, R. and Daniel, H.-D. 2010. A Reliability-Generalization Study of Journal Peer Reviews: A Multilevel Meta-Analysis of Inter-Rater Reliability and Its Determinants. *PLoS ONE*. 5, 12 (Dec. 2010), e14331.

[3] Campanario, J.M. 2002. The peer review system: many problems and few solutions. *Revista española de Documentación Científica*. 25, 3 (Sep. 2002).

[4] Falchikov, N. and Goldfinch, J. 2000. Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks. *Review of Educational Research*. 70, 3 (Jan. 2000), 287–322.

[5] Jackson, J.L., Srinivasan, M., Rea, J., Fletcher, K.E. and Kravitz, R.L. 2011. The Validity of Peer Review in a General Medicine Journal. *PLoS ONE*. 6, 7 (Jul. 2011), e22475.

[6] Kassirer, J.P. and Campion, E.W. 1994. Peer review. Crude and understudied, but indispensable. *JAMA*. 272, 2 (Jul. 1994), 96–97.

[7] Marsh, H.W., Bond, N.W. and Jayasinghe, U.W. 2007. Peer review process: Assessments by applicant-nominated referees are biased, inflated, unreliable and invalid. *Australian Psychologist*. 42, 1 (Mar. 2007), 33–38.

[8] Mulligan, A., Hall, L. and Raphael, E. 2013. Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology*. 64, 1 (Jan. 2013), 132–161.

[9] Satorre-Cuerda, R., Compañ-Rosique, P., Villagra-Arnedo, C., Gallego-Durán, F.J., Llorens-Largo, F. and Molina-Carmona, R. 2016. ¿Por qué no evaluamos la evaluación? Un esbozo para un sistema de evaluación entre iguales. *Actas del Simposio-Taller Comparte tu manera de innovar: aprendamos juntos* (Almería, 2016), 19–26.

[10] Smith, R. 1997. Peer review: reform or revolution? *BMJ*. 315, 7111 (Sep. 1997), 759–760.

[11] Sokolova, M. and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*. 45, 4 (Jul. 2009), 427–437.

[12] Topping, K. 1998. Peer Assessment Between Students in Colleges and Universities. *Review of Educational Research*. 68, 3 (Jan. 1998), 249–276.

[13] Yankauer, A. 1990. Who are the peer reviewers and how much do they review? *JAMA*. 263, 10 (Mar. 1990), 1338–1340.