
Inferencia de lenguajes racionales estocásticos

Rafael C. Carrasco Jiménez

Tesis de Doctorado

Facultad: Escuela Politécnica Superior

Director: Dr. José Oncina Carratalá

1997

Universidad de Alicante
Departamento de Lenguajes
y Sistemas Informáticos



INFERENCIA DE LENGUAJES RACIONALES
ESTOCÁSTICOS

Rafael C. Carrasco Jiménez
Tesis doctoral

Mayo 1997

La presente memoria constituye la tesis doctoral presentada por Rafael C. Carrasco Jiménez para la obtención del título de Doctor en Informática y ha sido desarrollada bajo la dirección del Dr. Jose Oncina Carratalá, profesor del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Alicante

Para Carmen.

Agradecimientos: A todos aquellos que contribuyeron al desarrollo de esta tesis, bien con su dirección (Jose Oncina), sus sugerencias (Enrique Vidal, Mikel Forcada), o su insustituible aportación técnica (Emilio Corbí, Pepe Verdú). También a todos mis compañeros del Departamento de Lenguajes y Sistemas Informáticos, donde tan buena acogida he recibido.

*El hombre no se libera de la acción por
no emprenderla.*

Baghavat Gita

Índice General

1	Introducción	11
1.1	Aprendizaje inductivo	12
1.2	Inferencia gramatical	14
1.3	Identificación en el límite	15
1.4	Identificación de lenguajes estocásticos	18
1.5	Aprendizaje PAC	19
2	Autómatas finitos estocásticos y lenguajes	23
2.1	Alfabetos y lenguajes	25
2.2	Autómatas finitos deterministas	27
2.3	Gramáticas regulares	29
2.4	Autómatas finitos deterministas estocásticos	31
2.5	Autómatas finitos de árboles	33
2.6	Autómatas estocásticos de árboles	35
3	Distancia entre lenguajes estocásticos	37
3.1	Entropía de un lenguaje estocástico	38
3.2	Entropía relativa entre lenguajes	42
3.3	Entropía de un lenguaje de árboles	43
3.4	Entropía relativa entre lenguajes de árboles	47
3.5	Convergencia de una variable aleatoria	49
3.6	Aproximación contra identificación	52

4	Procedimientos clásicos de aprendizaje	59
4.1	Modelos de Markov ocultos	59
4.2	Modelos bayesianos y fusión de estados	65
5	Inferencia de lenguajes regulares	71
5.1	Antecedentes	72
5.2	Formalismo	73
5.3	Algoritmo de inferencia	79
5.4	Convergencia del algoritmo	82
5.5	Número de ejemplos necesarios para la convergencia	85
5.6	Resultados y discusión	86
6	Inferencia de lenguajes racionales de árboles	97
6.1	Antecedentes	97
6.2	Formalismo	99
6.3	Algoritmo de inferencia	107
6.4	Inferencia probabilística	109
6.5	Resultados	111
7	Inferencia estocástica con redes neurales	119
7.1	Antecedentes	119
7.2	Arquitectura de la red	121
7.3	Resultados y discusión	124
8	Conclusión y perspectivas	131

Presentación

En esta memoria se presentan una serie de algoritmos que permiten la identificación de forma eficiente de lenguajes racionales (esto es, lenguajes reconocidos por autómatas finitos) a partir de muestras aleatorias. El significado de la inferencia, en particular para los lenguajes estocásticos, los criterios de éxito en el aprendizaje y otros conceptos básicos son presentados en el capítulo 1. El capítulo 2 contiene una introducción a los autómatas finitos estocásticos. En el capítulo 3 se discuten distintos métodos para evaluar la calidad de los modelos obtenidos. Dichos métodos miden la distancia entre las distribuciones de probabilidad propuestas y las correctas basándose en la teoría de la información. Algunas técnicas y métodos tradicionales usados en el aprendizaje de lenguajes estocásticos son revisados en el capítulo 4. En el capítulo 5 se presenta un algoritmo nuevo para la identificación de lenguajes racionales de cadenas y en el capítulo 6, otro algoritmo para la identificación de lenguajes racionales de árboles. En el capítulo 7 se explora la posibilidad de que las redes neurales recurrentes de segundo orden identifiquen de forma robusta lenguajes racionales a partir de ejemplos estocásticos. Por último se realiza una discusión sobre los problemas abiertos y las posibles continuaciones de este trabajo.

