

SISTEMAS ESCALABLES PARALELOS Y DISTRIBUIDOS PARA EL DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS.

Carlos Rodríguez Lucatero

ITESM Campus Ciudad de México, Departamento de Computación,
Tlalpan Ciudad de México, México, 14380
e-mail lucatero@campus.ccm.itesm.mx

RESUMEN

Este trabajo es una propuesta para mejorar el desempeño de los sistemas de descubrimiento de conocimientos en grandes bases de datos aplicando algoritmos de clasificación de tipo PAC-learning en modo distribuido para extracción de conceptos expresables en FNC y explorar la posibilidad de aplicación de dicho método a conceptos representables en un lenguaje más expresivo a través del uso de la dimensión de Vapnik-Chervonenkis. A pesar de que se ha realizado una investigación considerable en el área de **KDD** la mayor parte de esta se ha dedicado a la concepción de algoritmos secuenciales y las paralelizaciones que se han implementado son variaciones de estos por lo cual la mejora en el desempeño es limitado.

Dichos algoritmos trabajan bastante bien en instancias muy pequeñas de problemas de clasificación pero frecuentemente en minería de datos se tiene que tratar con BD grandes y dicho enfoque resulta impráctico. En términos más precisos el problema de hallar un árbol de clasificación óptimo o el conjunto mínimo de reglas en el caso más general ha sido demostrado ser un problema NP-completo. En éste artículo proponemos un método de aprendizaje aproximadamente correcto en forma distribuida/paralela haciendo un compromiso entre precisión y velocidad de obtención de reglas.

Palabras claves: Minería de datos, PAC-learning, Algoritmos distribuidos, Paralelización, Agentes .