Descubrimiento de reglas de asociación en bases de datos grandes mediante técnicas metaheurísticas distribuidas

Ricardo Di Pasquale[1]

[1] Universidad Católica Argentina, Argentina.

rdipasquale@uca.edu.ar

En este trabajo estamos interesados en el hallazgo de reglas de asociación en bases de datos grandes, particularmente, en ambientes que puedan ser clasificados como *Big Data* ya sea por su tamaño, o por su complejidad.

El problema de reglas de asociación se define en [Agrawal et al., 1994] de la siguiente manera: sea $I = \{i_1, i_2, ..., i_n\}$un conjunto de n atributos binarios (o items) y $D = \{t_1, t_2, ..., t_m\}$un conjunto de datos transaccionales almacenados en una base de datos. Cada transacción en D tiene un identificador único, y cada transacción contiene un subconjunto de items de I. Se define una regla como una implicación de la forma $X \implies Y, (X \cup Y) \subseteq I$, siendo X el antecedente de la regla, e Y, su consecuente. Se define el *soporte* de un conjunto $C \subseteq I$ de items en una base de datos D a la proporción de transacciones o registros que contiene el conjunto C de items, es decir:

$$sop(C) = \frac{|C|}{|D|}$$

Se define *confianza* de una regla $r = X \implies Y$ como:

$$conf(X \implies Y) = \frac{sop(X \cup Y)}{sop(X)} = \frac{|X \cup Y|}{|X|}$$

Se busca encontrar conjuntos de reglas del tipo $r = X \implies Y$, $(X \cup Y) \subseteq I$ de tal manera que se maximice $conf(r)$ para valores paramétricos de $sop(r)$. El objetivo del problema es, entonces, buscar conjuntos de reglas del estilo $r = X \implies Y, (X \cup Y) \subseteq I$ para los que fijando un parámetro de *soporte* ($0 < ps_1 \leq sop(r) \leq ps_2 \leq 1$), que puede ser ajustado según la naturaleza del conjuntos de datos, tengan un valor máximo de confianza. Conviene ajustar conve-

nientemente el *soporte* del conjunto de reglas objetivo en cada conjunto de datos para minimizar la posibilidad de encontrar reglas tautológicas que no son de interés en este problema.

Dado que las implementaciones clásicas de este tipo de algoritmos trabaja cargando el conjunto de datos en memoria, no es posible considerar su uso para volúmenes importantes de datos. Con el fin de obtener buenos resultados en conjuntos grandes de datos, se optó por implementar metaheurísticas en ambientes distribuidos compatibles con las nociones de *Big Data* y con *clusters* de computadoras de bajo costo. Es por eso que la distribución de los datos se ha implementado mediante *file systems* distribuidos en *Apache Hadoop (HDFS)*, y, los algoritmos, se han implementado en *Scala* con *Apache Spark como framework* de procesamiento distribuido a gran escala.

El estudio realizado presenta resultados comparativos con herramientas similares que utilizan otras técnicas, particularmente, con la técnica *"A Priori"* introducida en [Agrawal et al., 1994], que es utilizada habitualmente para deducir reglas de asociación a partir de conjuntos de datos. Cabe destacar que la implementación original de *"A Priori"* no fue preparada para procesar en paralelo o de manera distribuida. La implementación de referencia es *WEKA* de la Universidad de Waikato documentada en [Hall et al., 2011].

Los resultados presentados comparan tanto atributos funcionales como pertinencia de las soluciones, mantenibilidad y extensibilidad de la solución desarrollada; así como también atributos no funcionales, a saber, rendimiento general, sensibilidad al tamaño de la base de datos en función del enfoque y nivel de distribución y s*peedup.*

Referencias

1. [Agrawal et al., 1994] Rakesh Agrawal and Ramakrishnan Srikant Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.
2. [Araujo et al., 1999] D. L. A Araujo., H. S. Lopes, A. A. Freitas, "A parallel genetic algorithm for rule discovery in large databases" , Proc. IEEE Systems, Man and Cybernetics Conference, Volume 3, Tokyo, 940-945, 1999.
3. [Hall et al., 2011] Mark Hall, Ian Witten, Eibe Frank - "Data Mining: Practical Machine Learning Tools and Techniques." Third Edition. Morgan Kaufmann Publishers - 2011 - ISBN: 978-0-12-374856-0.