

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/365064026>

Minería incremental de datos para extracción de reglas de asociación difusas en datos ambientales

Conference Paper · September 2022

CITATIONS

0

READS

53

5 authors, including:



[Ignacio J. Blanco](#)

University of Granada

63 PUBLICATIONS 818 CITATIONS

[SEE PROFILE](#)



[Carmen Martínez-Cruz](#)

University of Granada

44 PUBLICATIONS 487 CITATIONS

[SEE PROFILE](#)



[Alain Perez](#)

19 PUBLICATIONS 37 CITATIONS

[SEE PROFILE](#)



[Jose M. Serrano](#)

University of Jaén

63 PUBLICATIONS 962 CITATIONS

[SEE PROFILE](#)

Minería incremental de datos para extracción de reglas de asociación difusas en datos ambientales

1st Ignacio Blanco
Dep. CC. de la Computación e I. A.
Univ. de Granada
Granada, España
iblanco@ugr.es

2nd Luisa M. González-González
Dep. CC. de la Computación
Univ. Central “Marta Abreu” de las Villas
Santa Clara, Cuba
luisagon@uclv.edu.cu

3rd Carmen Martínez-Cruz
Dep. Informática
Univ. de Jaén
Jaén, España
cmcruz@ujaen.es

4th Alain Pérez-Alonso
Dep. Electrónica e Informática
Univ. Técnica Federico Santa María
Concepción, Chile
alain.perez@usm.cl

5th José M. Serrano
Dep. Informática
Univ. de Jaén
Jaén, España
jschica@ujaen.es

Resumen—En problemas del mundo real, las bases de datos no son estáticas, y la información se actualiza continuamente. Un ejemplo son los datos procedentes de entornos inteligentes. Precisamente, el análisis de dicha información mediante técnicas tradicionales de minería de datos puede verse dificultado por esta continua variación. En este trabajo se resume la propuesta de dos algoritmos de minería de datos incremental, siguiendo un enfoque inmediato y diferido, respectivamente, para la extracción de reglas de asociación difusas.

Palabras clave—minería incremental de datos, reglas de asociación difusas, sensores, entornos inteligentes

I. INTRODUCCIÓN

La minería de datos continúa vigente como un conjunto de técnicas útiles para el análisis de grandes volúmenes de información. En particular, la extracción de reglas de asociación, en sus múltiples vertientes, sigue siendo objeto de numerosos estudios. En un campo como el de la inteligencia ambiental, donde es habitual trabajar con flujos continuos de datos, procedentes de distintas fuentes, tales como sensores o contenidos en la nube, resulta especialmente interesante contar con herramientas que permitan hallar relaciones o patrones entre las diferentes medidas consideradas.

Sin embargo, la naturaleza continua de dichos flujos de información dificulta la aplicación directa de técnicas tradicionales de minería de datos. Cambios en la información pueden provocar que las reglas obtenidas queden obsoletas rápidamente. En general, si los datos de partida se alteran, es necesario volver a generar e interpretar los resultados desde el principio.

Para afrontar este problema, se han propuesto diferentes soluciones a lo largo de los últimos años, dentro del área conocida como minería de datos incremental. En el presente

trabajo, se resumen las aportaciones realizadas en [4], donde se proponen dos algoritmos de minería de datos incremental, mediante los que mantener la validez de un conjunto de reglas de asociación previamente extraído.

Nuestra propuesta no solo se limita a los dos algoritmos. Además, se propone una estructura de almacenamiento de las características de dichas reglas para optimizar el proceso de mantener su validez a lo largo del tiempo.

II. REGLAS DE ASOCIACIÓN

Dado un conjunto de transacciones T , una regla de asociación [1] se puede ver como una implicación de la forma $X \Rightarrow Y$, donde $X, Y \subset It$, siendo $It = \{it_1, it_2, \dots, it_m\}$ un conjunto de ítems tal que $It \subseteq T$, donde se cumple que $X \neq \emptyset$, $Y \neq \emptyset$ y $X \cap Y = \emptyset$. A X e Y se les denomina, respectivamente, antecedente y consecuente de la regla.

Aunque a lo largo de los años se han propuesto diferentes medidas de la relevancia y el interés de una regla de asociación, las medidas más extendidas son su soporte, definido como la proporción de transacciones sobre el total de T que contienen a $X \cup Y$, y su confianza, obtenida como la fracción de transacciones que contienen a $X \Rightarrow Y$ sobre las transacciones que contienen a X . En general, todas las medidas de interés de una regla de asociación comparten una serie de características comunes, relacionadas con el antecedente o consecuente de la regla, excepciones a la misma, etc. Dichas propiedades permiten combinarse para obtener otras métricas en función de ellas.

Tradicionalmente, las reglas de asociación se aplican sobre conjuntos de transacciones. Posteriores estudios han extendido la definición para operar sobre datos cuantitativos o con una alta granularidad. En [2], se presenta una propuesta para extraer reglas de asociación difusas (FAR), donde se consideran transacciones difusas como subconjuntos difusos de ítems, y que generaliza la definición clásica de regla de asociación.

Este trabajo ha sido parcialmente financiado por el proyecto B-TIC-744-UGR20 ADIM: Accesibilidad de Datos para Investigación Médica. Convocatoria FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades.

III. MANTENIMIENTO DEL CONOCIMIENTO EXTRAÍDO

En problemas del mundo real, los datos no suelen ser estáticos, y varían como resultado de operar sobre la base de datos. En el ámbito de las bases de datos activas, dichas operaciones se representan como eventos, primitivos (PSE, *primitive structural event*) o compuestos (como una combinación de varios PSE). Es habitual considerar tres tipos de PSE: inserción, borrado y actualización de una tupla. Analizando los PSE que afectan a una base de datos, es posible registrar los diferentes estados por los que va pasando la misma.

En [4], se detalla cómo las FAR obtenidas se almacenan como un elemento más de la base de datos. Cada regla cuenta con una serie de características (FMP, *fuzzy measure part*) que, combinadas, nos permiten calcular métricas de interés. Un ejemplo es la confianza de una regla, obtenida como la fracción entre la proporción de casos donde aparecen antecedente y consecuente, y la proporción de casos en los que aparece el antecedente de la regla. Estas características nos permiten obtener diferentes métricas de forma eficiente. Por ejemplo, en [3] se demuestra que con solo 5 características distintas es posible obtener hasta 20 métricas de interés para una regla de asociación.

De acuerdo con lo anterior, en [4] se proponen dos algoritmos para el mantenimiento incremental de FAR en una base de datos. Ambos tienen la ventaja de que permiten mantener la validez de las FAR sin tener que volver a generarlas de cero, y sin necesidad de volver a revisar la base de datos al completo.

El primer algoritmo sigue un enfoque inmediato. Cada vez que se registra un PSE, el sistema considera si es preciso actualizar las FMP para cada regla almacenada en la base de datos. Si el número de eventos es muy alto (es decir, si la información se modifica con mucha frecuencia), el rendimiento del algoritmo puede verse perjudicado. Es por ello que el segundo algoritmo propone un enfoque diferido, donde se registran igualmente las ocurrencias de distintos PSE, pero en una estructura secundaria, y la actualización de las FMP de las reglas se lleva a cabo, en un solo paso, en una etapa posterior. Por ejemplo, si se inserta una nueva fila y más adelante se elimina, la operación en conjunto no tendría efecto sobre las reglas. El enfoque diferido permite obviar aquellos PSE que no afectan a la base de reglas. En el peor de los casos, habrá que considerar un número de operaciones igual al caso del enfoque inmediato. La figura 1 muestra un esquema del funcionamiento de ambos enfoques.

IV. EXPERIMENTOS Y CONCLUSIONES

En [4], se compara el rendimiento de los algoritmos propuestos con el de algoritmos tradicionales como fuzzy Apriori y fuzzy FP-growth, disponibles en KEEL [5]. Los experimentos se ejecutaron a partir de 3 datasets del UCI Machine Learning Repository: *diabetes*, *color texture* y *color moment*. Se extrajo un total de 7 FAR y se midió el tiempo transcurrido tras realizar 5000 operaciones sobre los datos, con el correspondiente mantenimiento de las reglas. En la figura 2 puede comprobarse cómo el tiempo de ejecución

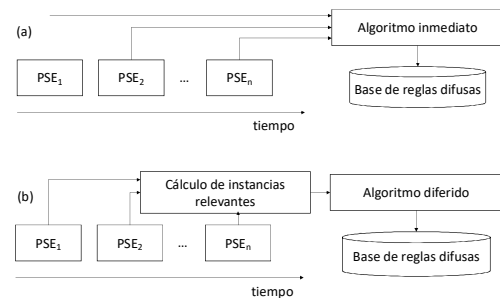


Figura 1. Mantenimiento incremental inmediato (a), y mantenimiento incremental diferido (b).

de nuestras propuestas es notablemente inferior al de los algoritmos tradicionales. Estos resultados ponen de manifiesto las ventajas de nuestra propuesta de minería incremental en problemas del mundo real, en particular, en su aplicación al análisis de bases de datos procedentes de sensores, aspecto que será estudiado en profundidad en un próximo trabajo.

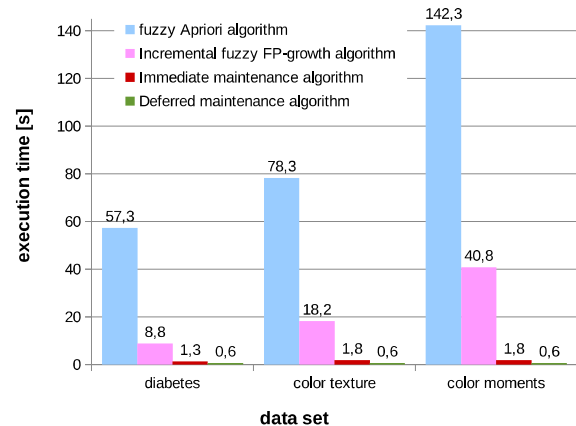


Figura 2. Comparativa entre algoritmos para el mantenimiento de FAR en diferentes datasets.

REFERENCIAS

- [1] R. Agrawal, T. Imieliński, and A. Swami. "Mining association rules between sets of items in large databases". SIGMOD Rec. 22, 2, 207–216, (1993). <https://doi.org/10.1145/170036.170072>
- [2] M. Delgado, N. Marin, D. Sanchez and M.A. Vila, "Fuzzy association rules: general model and applications", in IEEE Transactions on Fuzzy Systems, vol. 11, 2, 214–225, (2003). <https://doi.org/10.1109/TFUZZ.2003.809896>
- [3] P. Lenca, P. Meyer, B. Vaillant, S. Lallich, "On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid". Eur. J. Operational Research, vol 184, 2, 610–626, (2008). <https://doi.org/10.1016/j.ejor.2006.10.059>.
- [4] A. Pérez-Alonso, I.J. Blanco, J.M. Serrano et al. "Incremental maintenance of discovered fuzzy association rules". Fuzzy Optim Decis Making 20, 429–449 (2021). <https://doi.org/10.1007/s10700-021-09350-3>
- [5] I. Triguero, S. González, J.M. Moyano, S. García, J. Alcalá-Fdez, J. Luengo, A. Fernández, M.J. del Jesus, L. Sánchez, F. Herrera. "KEEL 3.0: An open source software for multi-stage analysis in data mining". Int. Journal of Computational Intelligence Systems 10, 1238–1249 (2017). <https://doi.org/10.2991/ijcis.10.1.82>