Revista Cubana de Ciencias Informáticas
Vol. 15, No. 3, Julio-Septiembre, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 153-174

Tipo de artículo: Artículo de revisión
Temática: Inteligencia Artificial
Recibido: 30/05/2021 | Aceptado: 07/07/2021

Reducción de tamaño en Modelos de Reglas de Asociación: Una revisión sistemática de la literatura

Size reduction in Association Rules Models: A systematic literature review

Julio César Diaz Vera [1]https://orcid.org/0000-0002-9868-1697
Guillermo Manuel Negrín Ortiz [1*]https://orcid.org/0000-0001-8637-9258
Carlos Molina [2]https://orcid.org/0000-0002-7281-3065
María Amparo Vila [3]https://orcid.org/0000-0002-2773-3306

[1] Profesor. Facultad 3. Universidad de Ciencias Informáticas, La Habana, Cuba. jcdiaz@uci.cu
[3] Universidad de Jaén, Jaén, España. carlosmo@ujaen.es
[4] Universidad de Granada, Granada, España. vila@decsai.ugr.es

*Autor para la correspondencia. (gmnegrin@uci.cu)

RESUMEN

Las Reglas de Asociación constituyen una de las tareas de minería de datos más estudiadas y aplicadas quizás porque su representación hace que sean fácilmente aceptadas e interpretadas por agentes humanos. Su

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 3, Julio-Septiembre, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 153-174

principal debilidad está asociada a la gran cantidad de reglas que son generadas para casos relativamente sencillos y que hacen imposible su análisis manual para determinar cuáles son las reglas relevantes. El objetivo de este trabajo es ejecutar una revisión sistemática de la literatura en el campo de la reducción del tamaño de los modelos de reglas de asociación con vistas a caracterizar y presentar el estado del arte de esta temática e identificar nuevas oportunidades de investigación. El análisis de los resultados muestra que la mayoría de los esfuerzos se enfocan hacia la eliminación de reglas redundantes pero este enfoque se está desplazando desde definiciones de redundancia asociadas a la estructura de las reglas hacia la inclusión del conocimiento de los usuarios dentro del proceso.

Palabras clave: reglas de asociación; reducción de modelos de reglas de asociación; revisión sistemática de la literatura.

ABSTRACT

Association Rules are one of the most studied and applied techniques in Data Mining. This is because they are easily accepted an interpreted by human agents. Association Rules main handicap is the great cardinality of models that even in simple datasets produce too many rules to be, manually, analyzed by experts in order to find those that are relevant ones. The objective of this paper is to carry out a systematic literature review in the field of size reduction in association rules models, to characterize and present the state of the art of this field. From the analysis of the results, it could be observed that most works focus on redundancy elimination but they are moving from redundancy definition associated to rule structure to redundancy definitions based on user knowledge and preference.

Keywords: association rules; size reduction in association rules models; systematic literature review.

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 3, Julio-Septiembre, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 153-174

Introducción

El volumen de datos que se genera diariamente ha crecido de manera drástica hasta alcanzar cotas superiores a los 2.5 billones de bytes diarios (Taluja, 2020). De igual forma el impacto de los datos en la sociedad y la economía es cada vez más significativo. La necesidad de procesar los datos para convertirlos en ventajas competitivas es un factor clave para el éxito de las organizaciones.

La minería de datos forma parte de las tecnologías utilizadas para analizar los datos y convertirlos en información útil que pueda ser transformada en acciones concretas que permitan potenciar el éxito. Dentro de la minería de datos una de las tareas más estudiadas y aplicadas es la minería de Reglas de Asociación (RA) quizás debido a que son muy similares a como los seres humanos representan el conocimiento.

El minado de reglas de asociación se define sobre una base de datos D que almacena las transacciones relevantes para un dominio particular. El dominio es representado a partir de un conjunto de elementos I. De modo que toda transacción $t \in D$ satisface que $t \subseteq I$. Una RA es una implicación de la forma $X \rightarrow Y$ donde X se denomina antecedente mientras que Y es el consecuente de la regla. Al mismo tiempo satisfacen las siguientes condiciones $X \cup Y \subseteq I$ y comúnmente pero no necesariamente, $X \cap Y = \emptyset$.

Las RA pueden interpretarse como reglas de la forma $si\ X\ entonces\ Y$ con la semántica asociada de que la aparición de X en una transacción de D implica la aparición de Y en la misma transacción. El problema del minado de RA consiste en encontrar todas las reglas $X \rightarrow Y$ que están presentes en las transacciones de D. De manera formal se expresa como $\{X \rightarrow Y \mid \exists t, t \in D \wedge X \cup Y \subseteq t\}$. Para $m = |I|$ la cantidad de subconjuntos de I que pueden aparecer en las transacciones sería $2^m - 1$ y para cada subconjunto de I con cardinalidad n la cantidad de reglas de asociación que pueden ser generadas es $2^n - 2$.

Las condiciones previas establecen dos grandes dificultades para el minado y utilización de RA.

La primera está dada por la complejidad exponencial lo que hace necesario la utilización de heurísticas para lograr algoritmos eficientes que permitan podar el espacio de búsqueda. En este sentido se ha establecido que las reglas interesantes deben ocurrir en la base de datos con una frecuencia mínima definida por el usuario. Esta frecuencia recibe el nombre de soporte y es la base de la heurística Apriori la cual utiliza la propiedad de clausura descendente del soporte (si un conjunto de ítems no satisface la frecuencia mínima cualquier super conjunto del mismo tampoco satisface esta condición) para podar el espacio de búsqueda. El desarrollo de

algoritmos eficientes para el minado de RA es un campo activo en el que aún se desarrollan muchas investigaciones y que tiene margen de mejora. Sin embargo, se han alcanzado resultados que permiten el minado de reglas en aplicaciones prácticas.

La segunda gran dificultad para la utilización de las RA radica en la alta cardinalidad de los modelos minados, que resulta inmanejable para los usuarios. Utilizar la frecuencia de ocurrencia de los patrones para disminuir la cantidad de reglas no es práctico debido a que las reglas que tienen frecuencias de aparición cercanas al 100% son usualmente conocidas por los especialistas o triviales y por tanto no tienen valor real.

Lo común es que las reglas realmente interesantes aparezcan con frecuencias mucho menores. En muchos casos se definen valores inferiores al 10% como umbral de frecuencia a la hora de considerar válida una regla. De esta forma los modelos contienen demasiadas reglas para poder ser tratadas por los especialistas, a este problema en la literatura científica (Bastide, 2000), (Balcazar, 2010) se le conoce como el problema de exposición de las reglas y es reconocido como el principal obstáculo para la utilización en la práctica para los modelos de reglas de asociación (Tirnauca, 2020).

En este trabajo se propone una revisión sistemática de la literatura de las técnicas utilizadas para reducir el tamaño de los modelos de Reglas de Asociación. La intención es identificar el estado del arte en esta temática y explorar sus potencialidades con vistas a facilitar la utilización práctica de las RA.

El resto del artículo se estructura de la siguiente forma: en la siguiente sección se discute la metodología seguida para la investigación. La sección 3 aborda los resultados y la discusión de los mismos. Finalmente se presentan las conclusiones en la sección 4.

Desarrollo

La revisión sistemática de la literatura sigue un protocolo preciso con vistas a obtener resultados contrastados desde el punto de vista científico. En este trabajo se utilizó la propuesta metodológica de (Keele, 2007) que define un proceso para identificar, analizar e interpretar los estudios relevantes disponibles en un área de conocimiento específica. El mismo cuenta con tres fases fundamentales:

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 3, Julio-Septiembre, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 153-174

1. Fase 1: Planificación de la revisión: en la que se define el protocolo que es utilizado para conducir la revisión.
2. Fase 2: Conducción de la revisión: en este momento se identifican y seleccionan los estudios relevantes, se evalúa la calidad de los estudios encontrados y se sintetiza la información.
3. Fase 3: Reporte de la revisión: En esta fase se crea el documento de revisión que presenta los resultados alcanzados (añadir un grupo de referencias).

Preguntas de investigación

La principal pregunta de investigación a responder en esta investigación es:
¿Cuál es el estado del arte en las publicaciones relacionadas con la reducción del tamaño de los modelos de reglas de asociación?

Dada la complejidad intrínseca de esta pregunta y con vistas a alcanzar una visión más específica de los estudios primarios más relevantes, fueron definidas otras cuatro preguntas de investigación. Con ello se pretende facilitar la tarea de responder la pregunta central. En la tabla 1 se presentan las preguntas de investigación adicionales, así como las razones que motivan cada una.

Tabla 1 – Preguntas de investigación.

Preguntas de investigación	**Motivación**
PI1: ¿Qué estrategias han sido definidas para reducir el tamaño de los modelos de RA?	Determinar qué tipos de estrategias (técnicas, modelos, metodologías, etc.) han sido establecidas para reducir el tamaño de los modelos de RA.
PI2: ¿Qué técnicas, en cada estrategia, han sido utilizadas para la reducción de tamaño en modelos de RA?	Identificar las técnicas que pueden facilitar la reducción de tamaño en los modelos de RA.
PI3: ¿Qué tipo de resultado científico presentan los investigadores?	Identificar los tipos de resultados que buscan los investigadores (modelos, marcos de trabajo, métodos etc.).
PI4: ¿Qué métodos científicos se han utilizado para validar las investigaciones?	Establecer los tipos de métodos empíricos que han sido empleados para validar los diferentes estudios.

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 3, Julio-Septiembre, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 153-174

Fuentes de datos y estrategia de búsqueda

Las fuentes a partir de las que se consideraran los estudios primarios son las bases de datos de la IEEE, la base de datos de Elsevier y la base de datos de Springer. Estas fuentes fueron seleccionadas ya que contienen la mayoría de las más importantes revistas y conferencias para diferentes áreas del conocimiento incluyendo el área en la que se centra esta investigación.

Para diseñar la cadena de búsqueda se utilizan operadores *OR* para enlazar elementos alternativos mientras que se utilizan operadores *AND* para enlazar los términos fundamentales. Los términos fundamentales de búsqueda pueden establecerse con facilidad y se seleccionó la expresión: *Associacion Rules size reduction techniques* los términos alternativos son más complejos debido a la variedad de formas en que puede ser expresado el problema del tamaño en los modelos de reglas de asociación. En la tabla 2 se presentan los términos utilizados para construir la cadena de búsqueda.

Tabla 2 - Cadena de búsqueda.

Términos principales	**Términos alternativos**
Association Rules size reduction techniques	*"Association Rules" AND ("large number" OR "immense quantity" OR "huge number" OR "too much" OR "huge size" OR "so large" OR "huge amount" OR "often huge" OR "overwhelming" OR "too many" OR "representative rules" OR "alternative for all rules")*

Finalmente, la cadena utilizada fue: *("Association Rules" AND ("large number" OR "immense quantity" OR "huge number" OR "too much" OR "huge size" OR "so large" OR "huge amount" OR "often huge" OR "overwhelming" OR "too many" OR "representative rules" OR "alternative for all rules"))*

Adicionalmente se aplicaron los siguientes filtros para refinar los resultados y obtener los de mayor calidad posible: 1- idioma: "*English*", 2- tipo de documento: (*"Conference Paper and Article and Article in Press"*), 3- Fecha: entre 2010 y febrero 2021.

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 3, Julio-Septiembre, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 153-174

Selección de los trabajos

La selección de los trabajos se realizó a través de un proceso de tres fases:

1. Fase 1: Selección de los trabajos potenciales aplicando la cadena de búsqueda e inspeccionando en cada uno título, resumen y palabras claves.
2. Fase 2: Seleccionar los trabajos candidatos realizando el análisis de texto completo de los trabajos potenciales y aplicando los criterios de exclusión.
3. Fase 3: Aplicación de los criterios de calidad a los trabajos candidatos.

Como parte del protocolo se establecieron los siguientes criterios de inclusión y exclusión:
Criterios de inclusión (CI):

1. CI1- Artículos escritos en inglés referentes al tratamiento del tamaño en modelos de reglas de asociación.
2. CI2- Artículos con texto completo publicados en revistas o congresos y revisados por pares entre el 2010 y el 2021.

Criterios de exclusión (CE):

1. CE1- Artículos asociados a las reglas de asociación, pero no enfocados en la reducción del tamaño de los modelos o que la tratan de manera muy somera.
2. CE2- Artículos duplicados, dando preferencia a los más recientes y de mayor claridad.
3. CE3- Artículos que presentan revisiones de la literatura y meta análisis en reglas de asociación.
4. CE4- Artículos sin texto completo.

Criterios de calidad

Para evaluar la calidad de cada uno de los artículos se definieron dos objetivos. El primero, aborda la relevancia de lo reportado en el estudio con respecto al objetivo de la revisión sistemática; a estos objetivos

corresponden los criterios de evaluación C1 y C2. El segundo objetivo está asociado a la credibilidad de los resultados teniendo en cuenta los métodos de evaluación y la diseminación de la investigación; a este objetivo corresponden los criterios C3, C4 y C5. Cada estudio puede alcanzar una evaluación entre 0 y 6 puntos y representa solamente un indicador que permite valorar la calidad del artículo, pero no constituye un criterio de exclusión.

Criterios de calidad:

1. C1- ¿El artículo contiene una descripción detallada de la estrategia y propiedades que soportan el proceso de reducción de tamaño? Las posibles respuestas son: Si (+1), No (+0).
2. C2- ¿El artículo describe los dominios o situaciones en los que es factible utilizar el proceso de reducción? Las posibles respuestas son: Si (+1), No (+0).
3. C3- ¿Se valida el estudio de acuerdo al tipo de propuesta realizada? Las posibles respuestas son: Validación empírica por medio de caso de estudio o experimento (+1), no validado (+0)
4. C4- ¿El estudio presenta un plan de validación? Las posibles respuestas son: tiene un plan completo (+1), tiene un plan parcial (+0.5), no tiene plan de validación (+0)
5. C5- ¿Está el artículo publicado en una revista o conferencia prestigiosa? Se utilizó como referencia para las revistas el ranking JCR (JCR, 2019) y para las conferencias el CORE (CORE, 2020). Las posibles respuestas son: Q1 o A* (+2), Q2 o A (+1.5), Q3 o B (+1), Q4 o C (+0.5), no rankeado (+0).

Resultados y discusión

Luego de la aplicación de la cadena de búsqueda se encontraron más de 1383 artículos. En la Fig. 1 se presenta la distribución temporal de los artículos encontrados, el grueso de los artículos se concentra en los últimos años y se aprecia un pico de producción científica en el año 2020 que reafirma el interés en la temática.

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 3, Julio-Septiembre, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 153-174

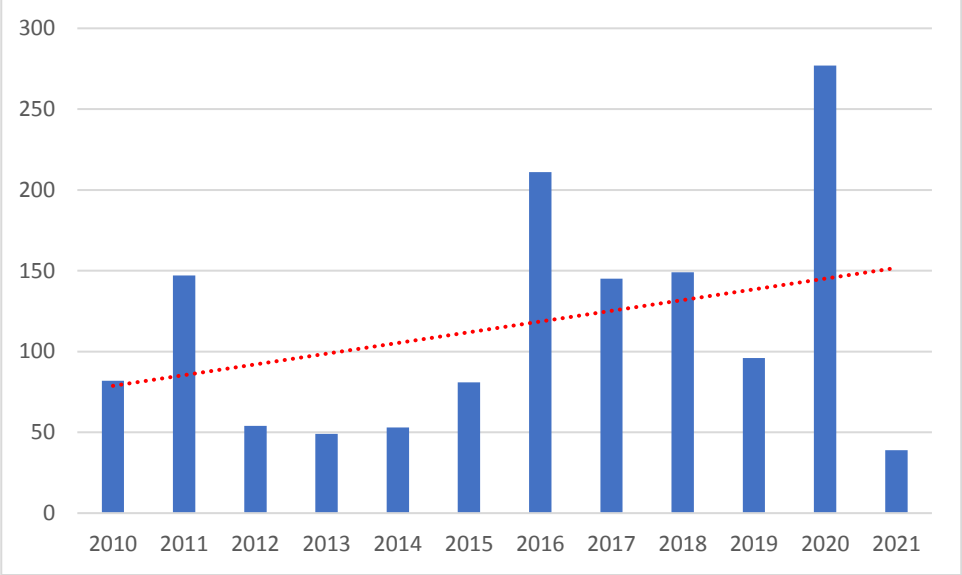

Fig. 1 - Distribución temporal de las publicaciones.

En la tabla 3 se muestra el resumen consolidado de los artículos estudiados y los seleccionados para la revisión. El total de artículos obtenidos al aplicar la cadena de búsqueda fue reducido a 327 al finalizar la fase 1. El análisis de cada uno de los artículos utilizando los criterios de inclusión y exclusión durante la fase 2 arrojó la selección de 42 artículos tal como puede observarse en la tabla 3. En la tabla 4 se listan los artículos finalmente seleccionados para la revisión identificándolos con el patrón [A + #] de esta forma se evita confundir los artículos que forman parte de la revisión con el resto de las referencias utilizadas en este trabajo.

Tabla 3 - Resumen de los artículos considerados en la revisión.

Base de datos	**Resultados**	**Fase 1**	**Fase 2**				
			CI1 y CI2	**CE1**	**CE2**	**CE3**	**CE4**
IEEE	476	112	11	84	3	13	1
Elsevier	427	101	22	70	1	8	0
Springer	480	114	9	86	5	11	3
Total	1383	327	42	283	9	32	4

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 3, Julio-Septiembre, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 153-174

Tabla 4 - Artículos incluidos en la revisión y su evaluación de calidad.

Id	Autor y título	Calidad	Referencia
[A1]	Shaharanee, Izwan Nizal Mohd; Jamil, Jastini Mohd. A framework for interestingness measures for association rules with discrete and continuous attributes based on statistical validity	2	(Shaharanee, 2015)
[A2]	Maamri, Ramdane; Said Hamani, Mohamed. Conceptual distance for association rules post-processing	2	(Maamri, 2011)
[A3]	Berka, Petr; Rauch, Jan. Meta-learning for Post-processing of Association Rules	2	(Berka, 2010)
[A4]	Fournier-Viger, Philippe; Tseng, Vincent S. Mining top-k non-redundant association rules	2	(Fournier, 2010)
[A5]	Mohammed, Mouhir. A new way to select the valuable association rules	1	(Mohammed, 2015)
[A6]	Marinica, Claudia; Guillet, Fabrice. Knowledge-based interactive postmining of association rules using ontologies	4	(Marinica, 2010)
[A7]	Liu, Huawen; Liu, Lei; Zhang, Huijie. A fast pruning redundant rule method using Galois connection	5	(Liu, 2011)
[A8]	Shaharanee, Izwan Nizal Mohd; Hadzic, Fedja; Dillon, Tharam S. Interestingness measures for association rules based on statistical validity	4	(Shaharanee, 2011)
[A9]	Ruiz, M. Dolores. Meta-association rules for mining interesting associations in multiple datasets	4	(Ruiz, 2016)
[A10]	Idoudi, Rihab. Ontology knowledge mining based association rules ranking	3	(Idoudi, 2016)
[A11]	Dimitrijevic, Maja; Bosnjak, Zita. Pruning statistically insignificant association rules in the presence of high-confidence rules in web usage data	2	(Dimitrijevic, 2016)
[A12]	Xu, Yue; Li, Yuefeng; Shaw, Gavin. Reliable representations for association rules	4.5	(Xu, 2011)
[A13]	Tran, Anh; Truong, Tin; LE, Bac. Simultaneous mining of frequent closed itemsets and their generators: Foundation and algorithm	4	(Tran, 2014)
[A14]	Madbouly, Magda M.; Abd el Reheem, e. M. A. N.; Guirguis, Shawkat K. Interval type-2 fuzzy logic using genetic algorithm to reduce redundant association rules	3	(Madbouly, 2021)
[A15]	Azzeddine, Dahbi; Jabri, Siham; Balouki Yousse, Gadi Taoufiq. The selection of the relevant association rules using the Electre method with multiple criteria	2.5	(Azzeddine, 2021)
[A16]	Liu, Xiangyu; Niu, Xinzheng; Fournier-VIGER, Philippe. Fast Top-K association rule mining using rule generation property pruning	3.5	(Liu, 2020)
[A17]	Szathmary, Laszlo. Closed Association Rules	2.5	(Szathmary, 2020)
[A18]	Azzeddine, Dahbi. Selecting, sorting and ranking association rules with multiple criteria using dominance relation	2	(Azzeddine, 2020)
[A19]	Parfait, Bemarisika; André, Totohasina. Elimination of Redundant Association Rules	2	(Parfait, 2020)

[A20]	Prakash, r. Vijaya; Sarma, s. S. V. N.; Sheshikala, M. Generating Non-redundant Multilevel Association Rules Using Min-max Exact Rules	3.5	(Prakash, 2018)
[A21]	Miani, Rafael Garcia Leonel; Junior, Estevam Rafael Hruschka. Eliminating Redundant and Irrelevant Association Rules in Large Knowledge Bases	3	(Miani, 2018)
[A22]	Ait-Mlouk, Addi; Gharnati, Fatima; Agouti, Tarik. Multi-criteria decisional approach for extracting relevant association rules	2	(Ait-Mlouk, 2017)
[A23]	Ait-Mlouk, Addi; Jiang, Lili. A Web-Based Platform for Mining and Ranking Association Rules	2.5	(Ait-Mlouk, 2020)
[A24]	Bemarisika, Parfait; Totohasina, André. An Efficient Method for Mining Informative Association Rules in Knowledge Extraction	2	(Bemarisika, 2020)
[A25]	Djenouri, Youcef, et al. Discovering strong meta association rules using bees swarm optimization	3.5	(Djenouri, 2018)
[A26]	Weidner, Daniel; Atzmueller, Martin; Seipel, Dietmar. Finding Maximal Non-redundant Association Rules in Tennis Data	3.5	(Weidner, 2019)
[A27]	Boudane, Abdelhamid. Enumerating non-redundant association rules using satisfiability	3.5	(Boudane, 2017)
[A28]	De Carvalho, Veronica Oliveira. Ranking Association Rules by Clustering Through Interestingness	2	(De Carvalho, 2017)
[A29]	Kryszkiewicz, Marzena. Representative Rule Templates for Association Rules Satisfying Multiple Canonical Evaluation Criteria	1	(Kryszkiewicz, 2011)
[A30]	Chen, Chun-Hao. Post-Analysis Framework for Mining Actionable Patterns Using Clustering and Genetic Algorithms	4	(Chen, 2019)
[A31]	Dong, Xiangjun. An efficient method for pruning redundant negative and positive association rules	4	(Dong, 2020)
[A32]	Chiclana, Francisco. Arm–Amo: an efficient association rule mining algorithm based on animal migration optimization	4	(Chiclana, 2019)
[A33]	Shukla, Shekhar; Mohanty, B. K.; Kumar, Ashwani. A Multi Attribute Value Theory approach to rank association rules for leveraging better business decision making	2	(Shukla, 2017)
[A34]	Ait-Mlouk, Addi; Agouti, Tarik; Gharnati, Fatima. Mining and prioritization of association rules for big data: multi-criteria decision analysis approach	5	(Ait-Mlouk, 2017)
[A35]	JIN, Maozhu; WANG, Hua; ZHANG, Qian. Association rules redundancy processing algorithm based on hypergraph in data mining	2.5	(Jin, 2018)
[A36]	Kryszkiewicz, Marzena. A lossless representation for association rules satisfying multiple evaluation criteria	3	(KryszkiewicZ, 2016)
[A37]	Kryszkiewicz, Marzena. Dependence factor for association rules	2	(Kryszkiewicz, 2015)

[A38]	De Padua, Renan; REZENDE, Solange Oliveira; DE CARVALHO, Veronica Oliveira. Post-processing association rules using networks and transductive learning	2.5	(De Padua, 2014)
[A39]	De Padua, Renan; DE CARVALHO, Veronica Oliveira; REZENDE, Solange Oliveira. Post-processing Association Rules: A Network Based Label Propagation Approach	2	(De Padua, 2016)
[A40]	Cifarelli, Bruno Braga; MIANI, Rafael Garcia Leonel. Weakly Supervised Learning Algorithm to Eliminate Irrelevant Association Rules in Large Knowledge Bases	2	(Cifarelli, 2020)
[A41]	Heraguemi, Kamel Eddine; KAMEL, Nadjet; DRIAS, Habiba. Multi-objective bat algorithm for mining interesting association rules	2	(Heraguemi, 2016)
[A42]	Gireesha, O.; Obulesu, O. Tkar: Efficient Mining of Top-k Association Rules on Real—Life Datasets	2	(Gireesha, 2016)

Resultados de las preguntas de investigación

En este acápite se reportan los resultados de la investigación en función de las preguntas planteadas luego de analizar cada uno de los trabajos seleccionados.

PI1: ¿Qué estrategias han sido definidas para reducir el tamaño de los modelos de RA?

Se han utilizados varias estrategias asociadas a la reducción de tamaño en los modelos de RA:

1. Representación concisa: enfocada en la obtención de grupos de reglas reducidos a partir de los que se puede generar u obtener el conjunto total de las reglas.
2. Reducción de redundancia: explota determinadas propiedades y relaciones entre las reglas que permitan excluir del modelo reglas que no aportan nueva información.
3. Post-procesamiento: una vez obtenido el modelo de reglas realiza tareas de ordenamiento, filtrado y eliminación de reglas atendiendo a diferentes criterios. Dentro de esta categoría se han utilizado varias técnicas diferentes e incluso la combinación de varias.
4. Métricas de interés: se enfocan en proponer nuevas variantes para evaluar el interés de una regla con vista a entregar a los usuarios aquellas que tengan mayor valor para él.
5. Top k rules: una variante de obtención de reglas que se enfoca en la obtención de las mejores k reglas de acuerdo a la confianza, pero sin especificar un nivel de soporte. Logran una reducción de

tamaño considerable pero no pueden garantizar que no exista perdida de información sensible producto de la reducción.

En la tabla 5 se agrupan los artículos de acuerdo a las estrategias utilizadas mientras que en la Fig. 2 se puede apreciar la distribución de los trabajos por cada una de las categorías. Claramente se aprecia que las estrategias que realizan post-procesamiento de las reglas minadas son las más favorecidas por los investigadores.

Tabla 5 - Estrategias para la reducción de tamaño utilizada en cada artículo.

Estrategia	**Artículos**
Representación concisa	A12, A13, A17, A20, A24, A27, A29, A36
Reducción de redundancia	A14, A19, A35
Post-procesamiento	A2, A3, A5, A6, A7, A9, A10, A18, A21, A23, A25, A26, A30, A31, A32, A33, A34, A38, A39, A40
Métricas de interés	A1, A8, A11, A15, A22, A28, A37, A41
Top k rules	A4, A16, A42

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 3, Julio-Septiembre, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 153-174

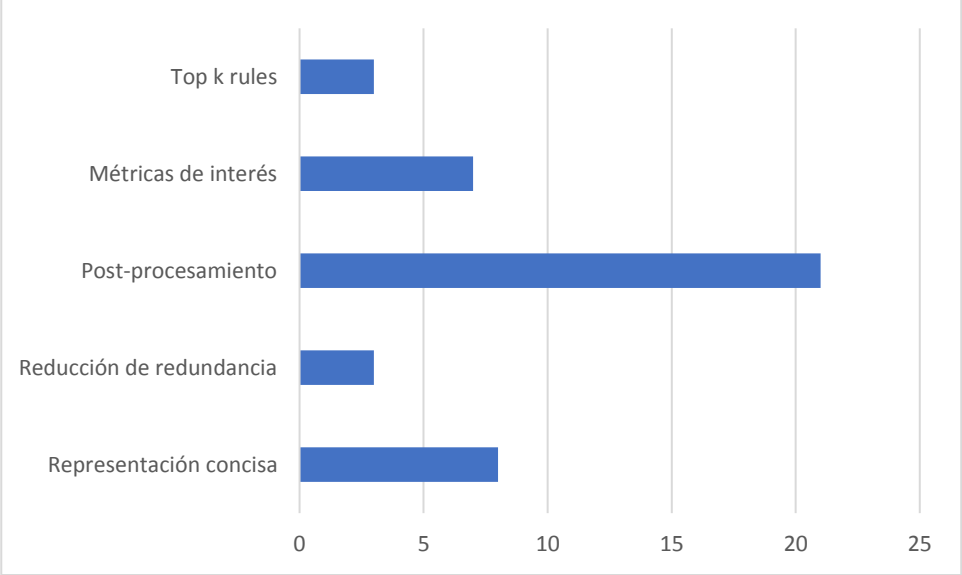

Fig. 2 - Distribución de artículos por categorías.

PI2: ¿Qué técnicas, en cada estrategia, han sido utilizadas para la reducción de tamaño en modelos de RA?

En la tabla 6 se presentan las diferentes técnicas que han sido utilizadas en los artículos. Es necesario precisar que varios artículos utilizan más de una técnica en su desarrollo y en esos casos se utilizó la clasificación dentro de la técnica considerada predominante.

Tabla 6 – Estrategias y técnicas involucradas en el desarrollo de los artículos.

Estrategia	**Técnicas**	**Artículos**
Representación concisa	Itemsets cerrados	A12, A13, A17, A20, A24, A27, A29, A36
Reducción de redundancia	Lógica difusa	A14
	Itemsets cerrados	A19
	Hiper grafos	A35
Post-procesamiento	Conocimiento del usuario	A2, A6, A10, A23
	Meta aprendizaje	A3, A9, A25
	Relación de preferencia	A5, A18, A40

	Eliminación de redundancia	A7, A21, A26, A31
	Metaheurísticas	A30, A32,
	Teoría de valor multi atributo	A33
	Decisión multicriterio	A34
	Análisis de redes	A38, A39
Métricas de interés	Estadística	A1, A8, A11, A37, A41
	Decisión multicriterio	A15, A22
	Clustering	A28
Top k rules	Frecuencia de las reglas	A4, A16, A42

En la tabla 6 puede observarse que la técnica más utilizada en la reducción del tamaño en modelos de RA está asociada a la eliminación de redundancia ya sea directamente o en la fase de post-procesamiento. Este detalle se acentúa si se considera que todos los trabajos que abordan las representaciones concisas a partir de ítemsets cerrados clasifican a las reglas que son eliminadas como reglas redundantes. Un total de 15 trabajos que representan el 35.71% de los estudiados siguen este enfoque.

Las definiciones de redundancias utilizadas en los artículos analizados comparten una misma base (BASTIDE, 2000). Semánticamente plantea que una regla de asociación es redundante si cubre la misma información o una información más específica que la cubierta por una regla de la misma utilidad y relevancia. Mientras que formalmente la define como: Una regla de asociación $R_1: X_1 \rightarrow Y_1$ es no redundante si no existe una regla de asociación $R_2: X_2 \rightarrow Y_2$ tal que $soporte(R_1) = soporte(R_2)$, $confianza(R_1) = confianza(R_2)$ y $X_2 \subseteq X_1 \wedge Y_1 \subseteq Y_2$.

Algunos de los artículos estudiados introducen algunas variaciones a la definición de redundancia.

1. (XU, 2011) propone que una regla $R_1: X_1 \rightarrow Y_1$ es redundante con respecto a $R_2: X_2 \rightarrow Y_2$ si $X_2 \subseteq X_1 \wedge Y_1 \subseteq Y_2$ y además $confianza(R_1) \leq confianza(R_2)$.

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 3, Julio-Septiembre, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 153-174

2. (PARFAIT, 2020) utiliza una métrica denominada $M_{\{GK\}} = \begin{cases} \frac{P(Y|X)-P(Y)}{1-P(Y)} \; si \; P(Y|X) > P(Y) \\ \frac{P(Y|X)-P(Y)}{P(Y)} \; si \; P(Y|X) \leq P(Y) \end{cases}$ y a partir de ella define que una regla $R_1: X_1 \rightarrow Y_1$ es redundante con respecto a $R_2: X_2 \rightarrow Y_2$ si $X_2 \subseteq X_1 \wedge Y_1 \subseteq Y_2$ y además $soporte(R_1) = soporte(R_2)$ y $M_{\{GK\}}(R_1) = M_{\{GK\}}(R_2)$.
3. (MIANI, 2018) plantea que una regla $R_1: X_1 \rightarrow Y_1$ es redundante con respecto a $R_2: X_2 \rightarrow Y_2$ si $Y_1 = Y_2 \wedge X_2 \subseteq X_1$.
4. (WEIDNER, 2019) propone que una regla $R_1: X_1 \rightarrow Y_1$ es redundante con respecto a $R_2: X_2 \rightarrow Y_2$ si $X_2 \subseteq X_1 \wedge Y_1 \subseteq Y_2$ y además $confianza(R_2) = 1$.
5. (DONG, 2020) plantea que una regla $R_1: X_1 \rightarrow Y_1$ es redundante con respecto a $R_2: X_2 \rightarrow Y_2$ si $X_1 = X_2 \wedge Y_2 \subset Y_1$.

La capacidad de reducción del tamaño de los modelos de reglas de asociación siguiendo este tipo de definición de redundancia fue estudiada en (BALCAZAR, 2010) demostrando que su límite teórico coincide con el de las reglas representativas. Estos modelos aún son demasiado complejos teniendo en cuenta la cantidad de reglas que contienen. Para continuar trabajando en este sentido se necesitan nuevas nociones de redundancia. Otra de las técnicas que ha sido explotada con éxito es la incorporación de conocimiento del usuario para descartar aquellas reglas que no sean relevantes. Sin embargo, ninguno de los trabajos ha intentado combinar ambas estrategias con el fin de alcanzar niveles de reducción más elevados o en última instancia simplificar la complejidad de los algoritmos involucrados en el proceso de reducción.

Los trabajos enfocados en las métricas de interés intentan decidir la mejor métrica a utilizar, pero sus resultados no han sido conclusivos. No se ha podido determinar una métrica que supere al resto en todos los casos e incluso los modelos de evaluación de la calidad de las métricas demuestran que estas no cumplen todas las propiedades deseables (Sudarsanam, 2020). Por ello el interés en este tipo de investigación se ha desplazado hacia la obtención de modelos de reglas que satisfagan al mismo tiempo varias métricas de calidad. De esta forma se logra minimizar los problemas asociados a la calidad de las métricas y al mismo tiempo se

Revista Cubana de Ciencias Informáticas
Vol. 15, No. 3, Julio-Septiembre, 2021
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 153-174

alcanzan grados de reducción del modelo que si bien no son suficientes pueden considerarse un punto de partida para continuar las investigaciones.

PI3: ¿Qué tipo de resultado científico presentan los investigadores?

Como puede observarse en la tabla 7 la mayoría de los trabajos han sido clasificados como métodos por parte de los investigadores. Este término, unido al de marco de trabajo, proyecta la intención del investigador de proyectar el trabajo sobre varias de las actividades del proceso de extracción de reglas de asociación. En contraste de las investigaciones clasificadas como algoritmos o bases representativas que se concentran en una etapa muy concreta del proceso de extracción. De esta forma la tendencia seguida por los investigadores se conduce a la presentación de un modelo de proceso y la descripción de las actividades y herramientas necesarias para culminar con éxito cada etapa.

Tabla 7 – Tipos de resultados de los artículos.

Tipo de resultado	**Artículos**
Marco de trabajo	A1, A6, A8, A30, A34
Algoritmo	A4, A9, A13, A14, A16, A18, A19, A25, A32, A41, A42
Método	A2, A3, A5, A7, A10, A11, A12, A15, A21, A22, A23, A24, A26, A27, A28, A31, A33, A35, A37, A38, A39, A40
Base representativa	A17, A20, A29, A36

PI3: ¿Qué métodos científicos se han utilizado para validar las investigaciones?

Solo el 7% de los artículos en el estudio no tienen alguna forma de validación, lo que puede considerarse como evidencia del grado de maduración en esta temática. En la tabla 8 se muestra en detalle la clasificación de los métodos de validación utilizados en los artículos. El método más empelado es el experimento, utilizado en más del 70% de los casos. Sin embargo, solo en (LIU, 2020) cuentan con un plan de validación lo que permite inferir que la temática precisa de definir protocolos de validación con vistas a incrementar el rigor y la credibilidad de los resultados alcanzados.

Tabla 8 – Distribución de los artículos teniendo en cuenta el método de evaluación.

Tipo de validación	**Artículos**
Experimento	A1, A2, A3, A4, A5, A6, A7, A8, A11, A12, A13, A14, A15, A16, A17, A18, A19, A20, A21, A24, A27, A28, A30, A31, A32, A38, A39, A40, A41, A42
Caso de estudio	A9, A10, A22, A25, A26, A33, A34
Teórica	A29, A36
Sin validación	A23, A35, A37

Conclusiones

Este trabajo ha presentado una revisión sistemática de la literatura asociada a la reducción de tamaño en modelos de Reglas de Asociación; con vistas a identificar, analizar y describir el estado del arte de esta temática. Luego de realizar la búsqueda de los trabajos potenciales y de seleccionar los trabajos relevantes, se identificaron las estrategias utilizadas en los diferentes artículos siendo la estrategia de post-procesamiento la más utilizada por los investigadores. Los trabajos también fueron clasificados atendiendo a la técnica utilizada para alcanzar el objetivo planteado. En este apartado se destaca la eliminación de redundancia que fue utilizada en el 35% de los trabajos tanto en la etapa de minado como en el post-procesamiento. Las acotaciones realizadas por los autores se encaminan a manejar definiciones de redundancia que se aparten de la clásica y que en la medida de lo posible incorporen conocimiento de los usuarios con vistas a entregar modelos que satisfagan en mayor grado sus expectativas.

La mayoría de los autores presentan sus resultados como métodos y generalmente están diseñados para manejar más de una actividad en el proceso de extracción de reglas de asociación y para incluir técnicas de otras áreas del conocimiento para mejorar los resultados. Sin embargo, se detectaron carencias metodológicas a la hora de establecer la definición de método como resultado de una investigación. Ninguno de los autores establece cuáles son las características metodológicas asociadas a la propuesta realizada en su investigación.

Los resultados muestran la existencia de interés alrededor de la temática y la madurez y rigor científico de la misma que se puede constatar a partir de que más del 90% de los artículos han sido validados mediante experimentos o casos de estudio. Aunque no se logró apreciar la existencia de un protocolo de experimentación estándar reconocido dentro de la temática.

Referencias

Taluja, S; Bhupal, J; Krishnan, S. A Survey Paper on DNA-Based Data Storage. En: International Conference on Emerging Trends in Information Technology and Engineering: IEEE, 2020, p. 1-4.

Tirnauca, C., Balcazar, J. L., Gomez-Perez, D. Closed-Set-Based Discovery of Representative Association Rules. International Journal of Foundations of Computer Science, 2020, 31(1): p. 143 – 156.

Keele, Staffs, Guidelines for performing systematic literature reviews in software engineering. Technical report, Ver. 2.3 EBSE Technical Report. EBSE, 2007.

Shaharanee, Izwan Nizal Mohd; Jamil, Jastini Mohd. A framework for interestingness measures for association rules with discrete and continuous attributes based on statistical validity. En: IFIP International Conference on Artificial Intelligence in Theory and Practice. Springer, Cham, 2015. p. 119-128.

Bastide, Yves, et al. Mining minimal non-redundant association rules using frequent closed itemsets. En International Conference on Computational Logic. Springer, Berlin, Heidelberg, 2000. p. 972-986.

Balcazar, Jose L., Redundancy, Deduction Schemes, and Minimum-Size Bases for Association Rules, Logical Methods in Computer Science, vol. 6, no.2, pp. 1-33, 2010.

Sudarsanam, Nandan, et al. Rate of change analysis for interestingness measures. Knowledge and Information Systems, 2020, vol. 62, no 1, p. 239-258.

Djenouri, Youcef, et al. Discovering strong meta association rules using bees swarm optimization. En Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2018. p. 195-206.

Weidner, Daniel; Atzmueller, Martin; Seipel, Dietmar. Finding Maximal Non-redundant Association Rules in Tennis Data. En Declarative Programming and Knowledge Management. Springer, Cham, 2019. p. 59-78.

Boudane, Abdelhamid, et al. Enumerating non-redundant association rules using satisfiability. En Pacific-Asia conference on knowledge discovery and data mining. Springer, Cham, 2017. p. 824-836.

De Carvalho, Veronica Oliveira. Ranking Association Rules by Clustering Through Interestingness. En Mexican International Conference on Artificial Intelligence. Springer, Cham, 2017. p. 336-351.

Kryszkiewicz, Marzena. Representative Rule Templates for Association Rules Satisfying Multiple Canonical Evaluation Criteria. En Asian Conference on Intelligent Information and Database Systems. Springer, Cham, 2018. p. 550-561.

Chen, Chun-Hao. Post-Analysis Framework for Mining Actionable Patterns Using Clustering and Genetic Algorithms. IEEE Access, 2019, vol. 7, p. 108101-108115.

Dong, Xiangjun. An efficient method for pruning redundant negative and positive association rules. Neurocomputing, 2020, vol. 393, p. 245-258.

Chiclana, Francisco. Arm–Amo: an efficient association rule mining algorithm based on animal migration optimization. Knowledge-Based Systems, 2018, vol. 154, p. 68-80.

Shukla, Shekhar; Mohanty, B. K.; Kumar, Ashwani. A Multi Attribute Value Theory approach to rank association rules for leveraging better business decision making. Procedia computer science, 2017, vol. 122, p. 1031-1038.

Ait-Mlouk, Addi; Agouti, Tarik; Gharnati, Fatima. Mining and prioritization of association rules for big data: multi-criteria decision analysis approach. Journal of Big Data, 2017, vol. 4, no 1, p. 1-21.

Jin, Maozhu; Wang, Hua; Zhang, Qian. Association rules redundancy processing algorithm based on hypergraph in data mining. Cluster Computing, 2019, vol. 22, no 4, p. 8089-8098.

Kryszkiewicz, Marzena. A lossless representation for association rules satisfying multiple evaluation criteria. En Asian Conference on Intelligent Information and Database Systems. Springer, Berlin, Heidelberg, 2016. p. 147-158.

Kryszkiewicz, Marzena. Dependence factor for association rules. En Asian Conference on Intelligent Information and Database Systems. Springer, Cham, 2015. p. 135-145.

De Padua, Renan; Rezende, Solange Oliveira; De Carvalho, Veronica Oliveira. Post-processing association rules using networks and transductive learning. En 2014 13th International Conference on Machine Learning and Applications. IEEE, 2014. p. 318-323.

De Padua, Renan; De Carvalho, Veronica Oliveira; REZENDE, Solange Oliveira. Post-processing Association Rules: A Network Based Label Propagation Approach. En International Conference on Current Trends in Theory and Practice of Informatics. Springer, Berlin, Heidelberg, 2016. p. 580-591.

Cifarelli, Bruno Braga; Miani, Rafael Garcia Leonel. Weakly Supervised Learning Algorithm to Eliminate Irrelevant Association Rules in Large Knowledge Bases. Journal of Information and Data Management, 2020, vol. 11, no 2.

Heraguemi, Kamel Eddine; Kamel, Nadjet; DRIAS, Habiba. Multi-objective bat algorithm for mining interesting association rules. En International Conference on Mining Intelligence and Knowledge Exploration. Springer, Cham, 2016. p. 13-23.

Gireesha, O.; Obulesu, O. Tkar: Efficient Mining of Top-k Association Rules on Real—Life Datasets. En Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications. Springer, Singapore, 2017. p. 45-54.

CORE. Computing research & education. 2020

JCR. SJR: Scientific Journal Rankings. 2019

Conflicto de interés

Los autores autorizan la distribución y uso de su artículo.

Contribuciones de los autores

1. Conceptualización: María Amparo Vila
2. Curación de datos: Guillermo Manuel Negrín Ortiz
3. Análisis formal: Julio César Díaz Vera
4. Adquisición de fondos: Carlos Molina Fernández
5. Investigación: Guillermo Manuel Negrín Ortiz
6. Metodología: Julio César Díaz Vera
7. Administración del proyecto: Carlos Molina Fernández
8. Recursos: María Amparo Vila
9. Software: Julio César Díaz Vera

10. Supervisión: María Amparo Vila
11. Validación: Guillermo Manuel Negrín Ortiz
12. Visualización: Julio César Díaz Vera
13. Redacción – borrador original: Guillermo Manuel Negrín Ortiz
14. Redacción – revisión y edición: Julio César Díaz Vera