IX Simposio de Teoría y Aplicaciones de la Minería de Datos (IX TAMIDA)

TAMIDA 2:
Modelos Descriptivos

Reglas de Asociación en Datos Multi-Instancia mediante Programación Genética Gramatical

José María Luna
Depto. Informática y Análisis Numérico,
Universidad de Córdoba
Email: jmluna@uco.es

Oscar Reyes
Depto. Informática y Análisis Numérico,
Universidad de Córdoba
Email: ogreyesp@gmail.com

María José del Jesus
Depto. Informática,
Universidad de Jaén
Email: mjjesus@ujaen.es

Sebastián Ventura
Depto. Informática y Análisis Numérico,
Universidad de Córdoba
Email: sventura@uco.es

***Abstract*—El estado del arte actual en minería de reglas de asociación es bastante prometedor en cuanto a la extracción automática de relaciones de interés entre patrones o elementos de grandes bases de datos. Por lo general, los datos se encuentran representados en forma tabular, donde cada fila define inequívocamente un registro de datos u objeto en el dominio de la aplicación. En ocasiones, estos registros de datos no pueden o deben ser considerados de forma aislada (por ejemplo, datos que comprenden hábitos de compra de clientes de manera que diferentes registros pueden describir un mismo cliente) y, por tanto, el problema no puede ser abordado como un problema de minería de reglas de asociación tradicional. Un primer intento de resolver este problema consideró un enfoque determinista sobre dominios discretos y requiriendo analizar todo el espacio de búsqueda (el tiempo de computo es exponencial respecto al número de elementos). Sin embargo, cuando nos enfrentamos a un problema real, la información se encuentra generalmente definida en dominios continuos, se requieren resultados en el menor tiempo posible, y se necesita satisfacer las expectativas de los usuarios (los resultados deben coincidir con la forma deseada). En este sentido, el objetivo de este trabajo es solucionar estos desafíos mediante un algoritmo de programación genética gramatical. Tanto sus fortalezas como defectos se analizan sobre un conjunto variado de datos en el estudio experimental.**

I. Introducción

Desde que se produjo la revolución digital, la cantidad de información que se almacena en cada dominio de aplicación ha crecido a un ritmo exponencial. En este sentido, se requiere que las diferentes técnicas de minería de datos mejoren su eficiencia para poder tratar con tales conjuntos masivos de datos en un tiempo considerable. Una de las tareas de minería de datos que se ha visto claramente afectada por la dimensionalidad de los datos es la minería de reglas de asociación (ARM por su término en inglés) [1]. Considerando un conjunto de datos con k elementos, el espacio de búsqueda en el que ARM debe buscar soluciones incluye un total de $3^k - 2^{k+1} + 1$ reglas de asociación diferentes obtenidas sobre un total de $2^k - 1$ patrones o combinaciones de elementos.

Incluso cuando se trata de un problema realmente complejo, especialmente para grandes conjuntos de datos, diferentes investigadores han propuesto algoritmos de ARM realmente eficientes. Dichas propuestas fueron especialmente relevantes gracias a las arquitecturas paralelas actuales y *frameworks* [**?**]. En este sentido, el problema de ARM se puede considerar como resuelto desde el punto de vista del rendimiento [**?**]. Sin embargo, el creciente interés en el almacenamiento de datos ha causado no solo un crecimiento en la dimensionalidad de los datos, sino también en la representaciones de dichos datos. A modo de ejemplo, consideremos una cadena de supermercados que quiere extraer conocimiento útil sobre los hábitos de compra de sus clientes para tomar decisiones correctas. Supongamos que la información se almacena en una tabla tradicional (cada fila indica una transacción diferente). Aquí, cada cliente tendrá un peso específico en la descripción de datos con respecto al número de transacciones que representa y, por lo tanto, esta representación de datos desvía el conocimiento extraído según el tipo de cliente (los clientes frecuentes son más importantes que los clientes esporádicos). Consideremos ahora que cada cliente se almacena como un ejemplo que comprende un número indefinido de transacciones, y los datos se describen sobre conjunto de ejemplos en lugar de sobre el conjunto de transacciones. De esta forma, el conocimiento extraído revela información explícita sobre los hábitos de los clientes independientemente de la cantidad de transacciones que realiza cada cliente. Esto implica que la forma en que se estructuran los datos es esencial para proporcionar el conocimiento correcto.

Este almacenamiento de información está por tanto relacionado con *multiple-instance learning* (MIL) [4], una generalización del aprendizaje supervisado tradicional. En MIL, los ejemplos son ambiguos y un solo ejemplo puede tener muchas instancias alternativas que lo describen. Esta ambigüedad se analizó recientemente [5] desde una perspectiva de descripción de datos mediante ARM sobre estructuras de datos de múltiples instancias. Sin embargo, este primer intento se basó en una metodología de búsqueda exhaustiva para establecer las bases en esta nueva representación para análisis descriptivo de datos y no se prestó atención al tiempo de ejecución, la metodología, el dominio de la aplicación y las

preferencias de los usuarios.

El objetivo de este trabajo es proponer un nuevo algoritmo de ARM para trabajar sobre datos de multi-instancia. El enfoque propuesto, conocido como MIAR-GePro, se basa en una metodología de programación genética gramatical y tiene como objetivo mejorar el estado del arte en este campo. MIAR-GePro puede trabajar tanto en dominios discretos como continuos, representando una importante ventaja respecto a la propuesta existente (requiere una discretización previa al proceso de extracción). Además, el modelo evolutivo propuesto es capaz de extraer reglas de asociación en un solo paso (la propuesta existente requiere extraer primero todos los patrones frecuentes y, a posteriori, obtener reglas de asociación sobre dichos patrones). Finalmente, los resultados obtenidos mediante MIAR-GePro están restringidos de acuerdo a una gramática predefinida por el usuario, por lo que es éste usuario final quien tiene la capacidad de definir la forma de las reglas que desea obtener. Todas estas características, así como algunos inconvenientes de esta propuesta se analizan en la etapa experimental al considerar un conjunto variado de datos.

El presente documento está organizado de la siguiente manera. La sección II presenta las definiciones más relevantes y el trabajo relacionado. La sección III describe el algoritmo MIAR-GePro propuesto. La sección IV presenta los conjuntos de datos utilizados en los experimentos, la configuración seguida, así como los resultados obtenidos. Finalmente, algunas observaciones finales se describen en la sección V.

II. Preliminares

En esta sección se describe por primera vez el problema de datos multi-instancia. A continuación, se define la tarea de minería de regla de asociación (ARM por su término en inglés) y se analiza el primer intento de extraer reglas de asociación en datos multi-instancia.

A. Datos Multi-Instancia

El problema de aprendizaje sobre datos multi-instancia fue introducido por primera vez por *Dietterich et al.* [4] en 1997 en el contexto de la predicción del comportamiento de diferentes fármacos. En este dominio de aplicación, los ejemplos son ambiguos y pueden incluir múltiples instancias alternativas que lo describen. De manera formal, supongamos que cada ejemplo e_i puede tener v_i variantes $e_{i,1}$, $e_{i,2}$, ..., e_{i,v_i}. Cada una de estas variantes estará representada por un vector de características distintas $V(e_{i,j})$, por lo que un ejemplo completo e_i se representa por lo tanto como un conjunto de vectores de características en la forma $\{V(e_{i,1}), V(e_{i,2}), ..., V(e_{i,v_i})\}$. Incluso cuando este problema se ha aplicado generalmente a las tareas de aprendizaje supervisado [6], realizar un análisis descriptivo de ese tipo de datos es de gran interés en muchos dominios. Considerando el conocido problema de análisis de la cesta de la compra, un único cliente puede describir diferentes compras, por ejemplo, en días diferentes de un mes. Como resultado, cada ejemplo (un cliente específico) incluye un número indefinido de registros (diferentes compras). En este escenario, el problema se puede definir como el análisis de un conjunto de datos en el que los datos se representan como multi-instancias, y se pueden aplicar diferentes hipótesis de trabajo.

La hipótesis de trabajo más simple fue propuesta por *Dietterich* [4], y considera que un determinado ejemplo se cumple si al menos una de las instancias incluidas en dicho ejemplo es satisfecha. En la minería de patrones, el problema se modela como encontrar un patrón P en el que $\exists j \ : \ P \subset V(e_{i,j})$ para una gran cantidad de ejemplos e_i. Teniendo en cuenta esta hipótesis en el análisis de la cesta de mercado, su objetivo es encontrar artículos (o conjunto de artículos) comprados al menos una vez por un gran número de clientes. Aquí, el objetivo es descubrir cuáles son los productos más populares o qué productos generalmente compran los clientes a la vez.

Múltiples autores han propuesto diferentes hipótesis de trabajo para tratar el problema de multi-instancia. *Weidmann et al.* [7] definió dos tipos adicionales de hipótesis que son casos especiales del propuesto por *Dietterich*. En la primera de estas hipótesis adicionales, el problema se modela como el de encontrar un patrón P en el que $\sum_{j=1}^{v_i} P \subset V(e_{i,j}) \geq m$ para un número alto de ejemplos e_i. Centrándose en el análisis de la cesta de la compra, esta hipótesis es realmente útil para determinar qué artículos (o conjunto de artículos) generalmente se compran más de m veces por un gran número de clientes. Aquí, no solo es importante conocer productos populares, sino que también se requiere que tales productos sean generalmente comprados por cada uno de los clientes. Finalmente, el segundo tipo de hipótesis definido por *Weidmann et al.* [7] establece que un patrón P se satisface con un número determinado de instancias en un ejemplo, y este número está entre un valor mínimo m y un valor máximo o. En otras palabras, $o \geq \sum_{j=1}^{v_i} P \subset V(e_{i,j}) \geq m$ en una gran cantidad de ejemplos e_i. Volviendo al análisis de la cesta de la compra, esta hipótesis es esencial para conocer el conjunto de clientes que consumen un producto específico un número predefinido de veces (dando un valor umbral mínimo y máximo).

B. Minería de Reglas de Asociación

La tarea de ARM fue definida formalmente por *Agrawal et al.* [8] en los años 90. Esta tarea tiene como objetivo extraer implicaciones de la forma $Antecedente \rightarrow Consecuente$ que sean de especial interés y desconocidas. Tanto $Antecedente$ (A) and $Consecuente$ (C) representan conjuntos disjuntos de elementos. De manera formal, se define $I = \{i_1, i_2, ..., i_n\}$ como el conjunto de n elementos incluidos en un conjunto de datos Ω, y $T = \{t_1, t_2, ..., t_m\}$ como el conjunto de transacciones o registros que cumplen $\forall t \in T : t \subseteq I \in \Omega$. Una regla de asociación es una implicación de la forma $A \rightarrow C$ donde $A \subset I$, $C \subset I$, y $A \cap C = \emptyset$. La frecuencia (también conocida como soporte) de una regla $A \rightarrow C$ es definida como el porcentaje de transacciones en las que se cumple $A \cup C \subseteq t : t \in T$. Este hecho se ha definido también en términos absolutos como el número de transacciones en las que aparece $A \cup C$, es decir, $|\{\forall t \in T : A \cup C \subseteq t, t \subseteq I \in \Omega\}|$. El significado de una regla de asociación es que si el antecedente

G = $(\Sigma_N, \Sigma_T, P, S)$ con:

S = Regla
Σ_N = {Regla, Antecedente, Consecuente, Condiciones, Condición, Condición_Nominal, Condición_Numérica}
Σ_T = {'Y', 'Atributo', 'valor', '=', '$\geq$','$\leq$'}
P = {Regla = Antecedente, Consecuente ;
Antecedente = Condiciones;
Consecuente = Condiciones;
Condiciones = 'Y', Condiciones, Condición | Condición ;
Condición = Condition_Nominal | Condición_Numérica ;
Condición_Nominal = 'Atributo', '=', 'valor' ;
Condición_Numérica = 'Atributo', '$\geq$', 'valor' | 'Atributo', '$\leq$', 'valor';}

Fig. 1. Gramática de contexto libre que representa reglas de asociación expresadas en notación BNF

de la regla se cumple, entonces es bastante probable que el consecuente de dicha regla también se cumpla.

El uso de ARM sobre representaciones de datos no estándar, por ejemplo datos multi-instancia, han sido recientemente estudiados debido a su relevancia. En un conjunto de datos multi-instancia Ω, cada bolsa de instancias o ejemplos e_i $in\Omega$ comprende un número variado l de transacciones o instancias, es decir, $e_i = \{t_{i,1}, t_{i,2}, ..., t_{i,l}\}$. Cada instancia única $t_{i,j}$ se define como un subconjunto de elementos tales que $t_{i,j} \in e_i$: j $leql, t_{i,j} \subseteq I$. La principal entre ARM clásica y la basada en multi-instancias radica en las medidas de calidad utilizadas para determinar el interés de las soluciones. En la forma más simple, una regla de asociación $A \rightarrow C$ cumplirá un ejemplo e_i si y sólo si al menos una de sus instancias $t_{i,j} \in e_i$ es satisfecha por la regla, es decir, $A \cup C \subseteq t_{i,j} : t_{i,j} \in e_i$.

La métrica conocida como soporte de una regla $A \rightarrow C$ se define como el número de ejemplos que la cumplen, es decir, $|\{\forall e_i \in \Omega \ : \ A \cup C \subseteq t_{i,j} \in e_i\}|$. El soporte puede definirse también en su forma relativa al tamaño del conjunto de datos tal que $|\{\forall e_i \in \Omega \ : \ A \cup C \subseteq t_{i,j} \in e_i\}|/|\Omega|$. Una característica importante del soporte en datos multi-instancia es la posibilidad de que tanto A como C satisfagan todas las transacciones, pero si se analizan juntas no satisfacen ninguna transacción. Esta afirmación no es posible en ARM clásico ya que considera cada transacción como una sola instancia. Centrándonos en otra medida de la calidad de las reglas en problemas multi-instancia, la confianza se define como la proporción del número de ejemplos que incluyen A y C entre todos los ejemplos que comprenden A. Otra métrica ampliamente utilizada entre las muchas existentes en ARM [10] es conocida como *lift* y permite establecer cuántas veces ocurren A y C juntas más a menudo de lo que se esperaría si fueran estadísticamente independientes, es decir $Lift(A \cup C) = Soporte(A \cup C)/(Soporte(A) \times Soporte(C))$.

III. MIAR-GePro

El algoritmo propuesto, conocido como MIAR-GePro (por sus siglas en inglés de *Multiple-Instance Association Rule Genetic Programming*), incorpora una gramática libre de contexto que se adapta tanto al problema en cuestión como al conocimiento subjetivo del usuario. Esta gramática permite predefinir la forma de las soluciones y, por lo tanto, el espacio de búsqueda se reduce a aquellas soluciones que cumplen con la gramática.

a) ***Criterio de codificación****:* en programación genética gramatical [11] cada solución al problema bajo estudio está representada por un genotipo (árbol de derivación basado en el lenguaje definido por una gramática G) y un fenotipo (significado de la estructura del árbol). Una gramática libre de contexto se define como $(\Sigma_N, \Sigma_T, P, S)$ donde Σ_T representa el alfabeto de los símbolos terminales y Σ_N el alfabeto de los símbolos no terminales, siendo ambos conjuntos disjuntos, es decir, $\Sigma_N \cap \Sigma_T = \emptyset$. Además, P establece el conjunto de reglas de producción que se utilizan para codificar una solución, y estas reglas de producción deben comenzar desde el símbolo inicial S. Una regla de producción se define como una regla $\alpha \rightarrow \beta$, donde $\alpha \in \Sigma_N$, y $\beta \in \{\Sigma_T \cup \Sigma_N\}^*$. Como se describió anteriormente, se aplica una serie de reglas de producción del conjunto P para obtener una solución que se representa como un árbol donde los nodos internos contienen solo símbolos no terminales y las hojas contienen solo terminales.

La Figura 1 muestra la gramática de contexto libre propuesta en el algoritmo MIAR-GePro. En esta gramática, el símbolo inicial S se representa con el término Rule (raíz del árbol de derivación), el cual siempre tiene como nodos hijos el antecedente y el consecuente de la regla. Estos dos nodos secundarios (antecedente y consecuente) pueden ser expandidos en una sola condición o un conjunto de condiciones que incluyen características tanto discretas como continuas. Como resultado y considerando la gramática propuesta (ver Figura 1), se obtiene el siguiente lenguaje de la gramática $L(G) = \{(Y \ Condicion)^n \ Condicion \ \rightarrow (Y \ Condicion)^m \ Condicion, \ n \geq 0, m \geq 0\}$ que cualquier solución válida debe satisfacer.

b) ***Proceso de evaluación****:* El proceso clave de cualquier algoritmo evolutivo es la evaluación de las soluciones, el cual guía el proceso de búsqueda a través de soluciones prometedoras dentro del espacio de búsqueda. Aunque se pueden utilizar muchas hipótesis de trabajo para evaluar las soluciones de MIAR-GePro (varias hipótesis se describieron en la sección II-A), en este trabajo hemos considerado la hipótesis tradicional de *Dietterich* donde un ejemplo específico se satisface si y sólo si se cumple al menos una instancia dentro de dicho ejemplo. Así pues, el número máximo de registros

que puede satisfacer una regla (solución al problema) vendrá dado por el número de ejemplos y no por el número total de instancias (considerando todos los ejemplos).

c) ***Operadores genéticos****:* MIAR-GePro incluye dos operadores genéticos para generar nuevas soluciones en cada generación del proceso evolutivo. Estos dos operadores se basan en el cruce y la mutación selectiva propuestos por *Whigham* [12]. El operador de cruce permite crear nuevas soluciones intercambiando subárboles aleatorios a partir de dos padres de forma que se garantizan soluciones correctas de acuerdo al lenguaje de la gramática (ver Figure 1). Para evitar el crecimiento incontrolado, se utiliza como parámetro una profundidad máxima de árbol. Con respecto al operador de mutación, éste selecciona de forma aleatoria un nodo no terminal del árbol y genera un nuevo subárbol a partir de él.

d) ***Esquema evolutivo****:* MIAR-GePro se basa en un algoritmo evolutivo generacional clásico con elitismo (ver Algoritmo 1). La élite está formada por un número máximo predefinido de soluciones, que son las soluciones que deben devolverse al finalizar la ejecución del algoritmo (ver línea 16, Algoritmo 1). Inicialmente, se genera aleatoriamente una población de individuos (soluciones en forma de árboles) (ver línea 3, Algoritmo 1) al considerar la gramática propuesta de contexto libre G y los atributos dentro del conjunto de datos Ω. Una vez que los individuos son evaluados con respecto a algunas medidas de calidad (ver *Evaluation Procedure*, líneas 17 a 44, Algoritmo 1) se ejecuta el ciclo principal, que comprende las siguientes operaciones (ver líneas 7 a 15, Algoritmo 1): (a) se selecciona un conjunto de individuos de la población pop para actuar como padres por medio de un selector por torneo de tamaño 2; (b) el conjunto de padres se cruzan y mutan para formar nuevos individuos (incluidos en el conjunto $newPop$); (c) la población general pop se actualiza mediante un reemplazo directo con elitismo, tomando las mejores soluciones de la élite $\mathcal{S}$ y del conjunto $newPop$. El tamaño máximo de $\mathcal{S}$ siempre es igual o menor que el tamaño de población principal (representado como $popSize$).

Algorithm 1 Pseudocódigo del algoritmo MIAR-GePro

Require: $\Omega, popSize, maxGen, maxSol, \alpha, G$
Ensure: $\mathcal{S}$
1: $\mathcal{S} \leftarrow \emptyset$
2: $pop \leftarrow \emptyset$ ▷ Conjunto de soluciones de cada iteración
3: $pop \leftarrow$ createSolutions $(G, \Omega, popSize)$ ▷ Gramática G usada para generar soluciones
4: $pop \leftarrow$ evaluationProcedure $(\Omega, nExamples, pop, \alpha)$
5: $\mathcal{S} \leftarrow$ takeBestSolutions$(pop, maxSol)$ ▷ Selecciona los $maxSol$ mejores soluciones encontradas en pop
6: $generation \leftarrow 1$
7: **while** $generation \leq maxGen$ **do** ▷ El algoritmo itera un número específico de generaciones
8: $parents \leftarrow$ selectParents(pop)
9: $newPop \leftarrow$ applyGeneticOperators $(parents)$
10: $newPop \leftarrow$ evaluationProcedure $(\Omega, newPop, \alpha)$
11: $\mathcal{S} \leftarrow \mathcal{S} \cup newPop$
12: $pop \leftarrow$ takeBestSolutions$(\mathcal{S}, popSize)$ ▷ Selecciona las $popSize$ mejores soluciones encontradas en $\mathcal{S}$
13: $\mathcal{S} \leftarrow$ takeBestSolutions$(pop, maxSol)$ ▷ Selecciona las $maxSol$ mejores soluciones encontradas en pop
14: $generation \leftarrow generation + 1$
15: **end while**
16: **return** $\mathcal{S}$

17: **procedure** Evaluation Procedure(Ω, pop, α)
18: $nExamples \leftarrow$ getNumberExamples(Ω)
19: **for all** solutions $s \in pop$ **do** ▷ Soluciones en pop
20: $countS \leftarrow 0$ ▷ Ejemplos satisfechos por la regla
21: $countA \leftarrow 0$ ▷ Ejemplos satisfechos por el antecedente
22: **for all** examples $e_i \in \Omega$ **do**
23: **for all** instance $t_{i,j} \in e_i$ **do**
24: **if** $s \subseteq t_{i,j}$ **then**
25: $countS \leftarrow countS + 1$
26: **break**
27: **end if**
28: **if** antecedent$(s) \subseteq t_{i,j}$ **then**
29: $countA \leftarrow countA + 1$
30: **break**
31: **end if**
32: **end for**
33: **end for**
34: $support \leftarrow countS/nExamples$
35: $supportA \leftarrow countA/nExamples$
36: $confidence \leftarrow countS/countA$
37: **if** support $\geq \alpha$ **then**
38: assignFitness$(s, support \times confidence)$
39: **else**
40: assignFitness$(s, 0)$
41: **end if**
42: **end for**
43: **return** pop
44: **end procedure**

IV. Estudio experimental

En esta sección, analizamos el rendimiento del algoritmo MIAR-GePro considerando una serie de conjuntos de datos y teniendo en cuenta hipótesis tradicional propuesta por *Dietterich*. En primer lugar, se realizará una comparativa del rendimiento de la propuesta respecto a un algoritmo de búsqueda exhaustivo (única propuesta existente en este campo hasta la fecha). En segundo lugar, se estudiarán una serie de métricas para medir la calidad media de las soluciones. Todos estos experimentos se realizarán considerando los siguientes parámetros en MIAR-GePro: 50 individuos; 100 generaciones; 0.8 probabilidad de cruce; 0.2 probabilidad de mutación; elitismo con un número máximo de soluciones de 20; y 0.1 como umbral de soporte mínimo.

En este estudio experimental se consideran 10 conjuntos de datos multi-instancia generados artificialmente (ver descripción en la Tabla I). A pesar de que existen multitud de conjuntos de datos reales multi-instancia en la literatura, éstos han sido especialmente diseñados para tareas de clasificación y, por tanto, no proporcionan resultados interesantes para una tarea específica como ARM. Con el fin de poder ejecutar algoritmos de búsqueda exhaustiva como Apriori-MI, los conjuntos de datos que comprenden atributos numéricos se han discretizado previamente. No obstante, cabe destacar que MIAR-GePro se puede aplicar directamente tanto a atributos discretos como continuos, no requiriendo ningún paso previo de procesamiento de datos.

TABLE I
CONJUNTO DE DATOS UTILIZADOS EN EL ESTUDIO EXPERIMENTAL

Conjunto de datos	#Ejemplos	#Atributos	#Instancias	Tamaño medio ejemplos
artificial1	200	6	705	3.53
artificial2	200	8	690	3.45
artificial3	500	8	1,742	3.48
artificial4	500	8	7,406	14.81
artificial5	1,000	10	6,958	6.96
artificial6	1,000	10	12,522	12.52
artificial7	2,000	12	19,986	9.99
artificial8	2,000	16	20,213	10.11
artificial9	5,000	6	17,482	3.50
artificial10	5,000	8	37,748	7.55

A. Análisis del rendimiento

El análisis de la escalabilidad de MIAR-GePro con respecto al algoritmo Apriori-MI [5] se ha llevado a cabo por medio de dos formas diferentes (ver Figuras 2 y 3). En primer lugar, es importante destacar que la escalabilidad de los algoritmos de ARM ha sido ampliamente estudiada por diferentes investigadores [3], lo que demuestra que el tiempo de ejecución de los algoritmos de búsqueda exhaustiva aumenta exponencialmente con el número de atributos. En este estudio, queremos analizar el rendimiento de los algoritmos de búsqueda tanto evolutivos como exhaustivos cuando se aplican sobre conjuntos de datos multi-instancia, considerando diferentes números de ejemplos (Figura 2) y atributos (Figura 3). Como se muestra, el algoritmo Apriori-MI requiere un tiempo computacional más alto con el incremento del número de ejemplos (también conocidos como bolsas), obteniendo todas las posibles reglas de asociación que satisfacen los umbrales establecidos (valor mínimo de soporte de 0.1. Destacar que este valor tan bajo permite obtener casi todas las reglas de asociación existentes). En cuanto a la escalabilidad de los algoritmos para diferentes números de atributos (ver Figura 3), se obtiene que algoritmos de búsqueda exhaustivos como

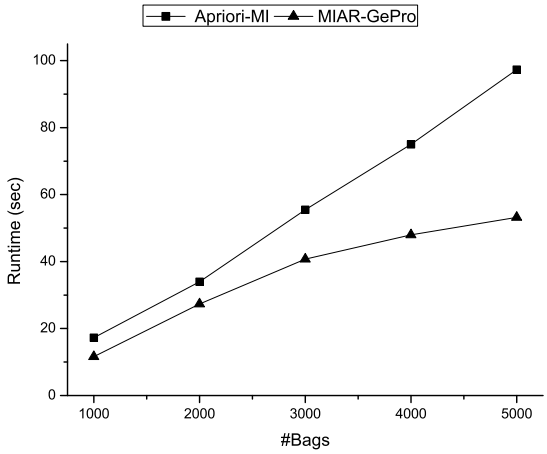


Fig. 2. Tiempo de ejecución en segundos para diferentes número de ejemplos (bolsas de instancias)

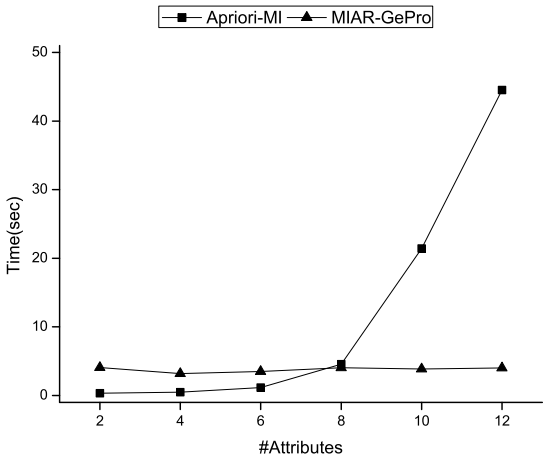


Fig. 3. Tiempo de ejecución en segundos para diferentes números de atributos

Apriori-MI aumentan exponencialmente su tiempo de cálculo con el incremento del número de atributos. Por el contrario, las propuestas evolutivas como MIAR-GePro no se ven afectadas por la cantidad de atributos.

B. Análisis de las medidas de calidad

La tabla II muestra los resultados obtenidos para diez conjuntos de datos diferentes discretizados en 5, 10 y 15 intervalos de igual anchura (la discretización es un requisito previo para poder ejecutar el enfoque de búsqueda exhaustiva existente). El algoritmo Apriori-MI se ejecuta utilizando un Soporte mínimo de 0.1, por lo que el algoritmo no tiene en cuenta ninguna regla que tenga un valor de Soporte inferior y calcula tres medidas de calidad diferentes (Soporte, Confianza y Lift). Analizando los resultados promedio representados en la Tabla II, el número de reglas descubiertas es mayor cuando el número de intervalos en los que se discretiza es menor. Esto se debe a la mayor cantidad de ejemplos que podrían satisfacerse cuando los atributos están discretizados en un pequeño número de intervalos (mayor rango de valores). Por el contrario, el mero hecho de considerar un número mayor de valores discretos (rango inferior de valores) implica que los atributos o patrones apenas se satisfacen, por lo que el número de ejemplos que comprenden patrones específicos es menor.

El mismo análisis anterior se ha realizado utilizando un enfoque evolutivo (ver Tabla III). En este caso específico, MIAR-GePro no requiere discretizar los conjuntos de datos en un número de intervalos, por lo que MIAR-GePro se ejecuta utilizando únicamente un umbral de soporte mínimo de 0.1. Además, MIAR-GePro permite extraer un número específico de reglas, por lo que hay que indicar dicho número que será el de las mejores reglas encontradas (se fija en 20 debido a que es el número de reglas generalmente utilizadas por los expertos en ARM con enfoques evolutivos [3]). El análisis de los resultados para la métrica de Confianza en el algoritmo evolutivo (no requiere preprocesado) determina mejores resultados que Apriori-MI. Es bastante interesante cómo cuanto mayor es el número de intervalos, mejores son

TABLE II
NÚMERO DE REGLAS OBTENIDAS Y VALORES MEDIOS PARA TRES MEDIDAS DE CALIDAD PARA EL ALGORITMO APRIORI-MI.

Conjunto de datos	#Reglas	Soporte	Confianza	Lift
artificial1-5	686	0.155	0.319	0.653
artificial1-10	192	0.156	0.408	1.101
artificial1-15	160	0.125	0.404	1.115
artificial2-5	1,480	0.147	0.324	0.702
artificial2-10	262	0.154	0.407	1.014
artificial2-15	208	0.128	0.406	1.112
artificial3-5	2,978	0.127	0.307	0.626
artificial3-10	140	0.240	0.548	0.994
artificial3-15	210	0.167	0.502	0.994
artificial4-5	36,788	0.147	0.244	0.365
artificial4-10	13,966	0.129	0.250	0.424
artificial4-15	430	0.313	0.560	0.987
artificial5-5	7,392	0.213	0.359	0.567
artificial5-10	294	0.294	0.542	0.993
artificial5-15	276	0.299	0.586	1.008
artificial6-5	26,670	0.172	0.263	0.372
artificial6-10	12,964	0.120	0.221	0.369
artificial6-15	538	0.289	0.538	0.998
artificial7-5	19,998	0.191	0.334	0.586
artificial7-10	35,979	0.118	0.342	0.993
artificial7-15	422	0.176	0.359	1.002
artificial8-5	15,523	0.164	0.366	0.898
artificial8-10	12,033	0.126	0.322	0.921
artificial8-15	528	0.194	0.542	1.001
artificial9-5	1,916	0.122	0.304	0.623
artificial9-10	100	0.243	0.552	1.002
artificial9-15	150	0.169	0.507	1.002
artificial10-5	4,498	0.226	0.372	0.581
artificial10-10	276	0.277	0.518	0.982
artificial10-15	224	0.309	0.592	1.024

TABLE III
NÚMERO DE REGLAS Y VALORES MEDIOS PARA TRES MÉTRICAS DE CALIDAD OBTENIDOS CON MIAR-GEPRO.

Conjunto de datis	#Reglas	Soporte	Confianza	Lift
artificial1	20	0.748	0.998	1.005
artificial2	20	0.895	1.000	1.005
artificial3	20	0.909	0.914	1.002
artificial4	20	0.975	0.999	0.999
artificial5	20	0.989	1.000	1.000
artificial6	20	0.987	0.999	0.999
artificial7	20	0.991	0.999	0.999
artificial8	20	0.995	1.000	1.000
artificial9	20	0.981	0.999	1.000
artificial10	20	0.997	0.999	0.999

los resultados obtenidos para esta medida de calidad en el enfoque de búsqueda exhaustiva. Esto explica por qué MIAR-GePro obtiene los mejores resultados, ya que no requiere un número fijo de intervalos (funciona directamente sobre atributos continuos).

V. CONCLUSIONES

El creciente interés en el almacenamiento de datos ha causado no solo un crecimiento en la dimensionalidad de los datos, sino también en la variedad de sus representaciones. Estudios recientes han propuesto soluciones para extraer reglas de asociación en conjuntos de datos donde múltiples registros describen un único objeto. Sin embargo, las soluciones existentes para la minería de reglas de asociación en este tipo de datos están restringidas a dominios discretos y requieren analizar todo el espacio de búsqueda (el tiempo de ejecución aumenta exponencialmente con el número de atributos). Así pues, el objetivo de este documento es proponer un nuevo algoritmo que permita trabajar en dominios continuos, obtener resultados lo antes posible y satisfacer las expectativas de los usuarios (los resultados deben coincidir con la forma deseada).

La propuesta MIAR-GePro utiliza programación genética gramatical para ARM sobre datos multi-instancia. MIAR-GePro puede trabajar en dominios discretos y continuos, representando una ventaja importante con respecto a la propuesta existente (Apriori-MI). Además, el modelo evolutivo propuesto es capaz de extraer las reglas de asociación en un solo paso (la propuesta existente requiere extraer primero todos los patrones frecuentes y, luego, obtener reglas de asociación de dichos patrones). El análisis experimental llevado a cabo en 10 diferentes conjuntos de datos ha demostrado que la propuesta es altamente eficiente, requiriendo un tiempo de ejecución que es casi constante para cualquier cantidad de atributos.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el proyecto TIN2017-83445-P del Ministerio de Economía y Competitividad y Fondos FEDER.

REFERENCES

[1] C. C. Aggarwal and J. Han, *Frequent Pattern Mining*. Springer International Publishing, 2014.
[2] F. Padillo, J. M. Luna, F. Herrera, and S. Ventura, "Mining association rules on big data through mapreduce genetic programming," *Integrated Computer-Aided Engineering*, vol. 25, no. 1, pp. 31–48, 2018.
[3] S. Ventura and J. M. Luna, *Pattern Mining with Evolutionary Algorithms*. Springer International Publishing, 2016.
[4] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1–2, pp. 31 – 71, 1997.
[5] J. M. Luna, A. Cano, V. Sakalauskas, and S. Ventura, "Discovering useful patterns from multiple instance data," *Information Sciences*, vol. 357, pp. 23–38, 2016.
[6] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez Tarragó, and S. Vluymans, *Multiple Instance Learning - Foundations and Algorithms*. Springer, 2016. [Online]. Available: https://doi.org/10.1007/978-3-319-47759-6
[7] N. Weidmann, E. Frank, and B. Pfahringer, "A Two-level Learning Method for Generalized Multi-instance Problems," in *Proceedings of the 14th European Conference on Machine Learning*, 2003, pp. 468–479.
[8] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD Conference '93, Washington, DC, USA, 1993, pp. 207–216.
[9] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach," *Data Mining and Knowledge Discovery*, vol. 8, pp. 53–87, 2004.
[10] F. Berzal, I. Blanco, D. Sánchez, and M. A. Vila, "Measuring the accuracy and interest of association rules: A new framework," *Intelligent Data Analysis*, vol. 6, no. 3, pp. 221–235, 2002.
[11] R. McKay, N. Hoai, P. Whigham, Y. Shan, and M. O'Neill, "Grammar-based Genetic Programming: a Survey," *Genetic Programming and Evolvable Machines*, vol. 11, pp. 365–396, 2010.
[12] P. Whigham and D. O. C. Science, "Grammatically-based genetic programming," in *Proceedings of the Workshop on Genetic Programming*, Tahoe City, California, USA, 1995, pp. 33–41.

Aproximación al índice externo de validación de clustering basado en chi cuadrado

José María Luna-Romera, Jorge García-Gutiérrez, María Martínez-Ballesteros and José C. Riquelme Santos
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Sevilla, España

***Abstract*—El clustering es una de las técnicas más utilizadas en minería de datos. Tiene como objetivo principal agrupar datos en clusters de manera que los objetos que pertenecen al mismo clúster sean más similares que los que pertenecen a diferentes clusters. La validación de un clustering es una tarea que se realiza aplicando los llamados índices de validación. Estos índices miden la calidad de la solución del clustering y se podrían clasificar como índices internos, los cuales calculan la calidad del clustering en función de los propios clusters; e índices externos, que miden la calidad mediante información externa de los datos, como puede ser la clase. Los índices externos que nos encontramos en la literatura están sujetos a una interpretación que puede dar lugar a error, por ello, el objetivo de este artículo es presentar un nuevo índice de validación externa basado en el test estadístico de chi cuadrado que mide la calidad del clustering de forma exacta, sin necesidad de tener que ser interpretado. Se ha realizado una experimentación usando 6 datasets que podrían ser considerados big data y los resultados obtenidos son prometedores ya que mejoran la tasa de aciertos y porcentaje de error respecto a los índices de la literatura.**

***Index Terms*—Análisis de clustering, validación de clustering, índices de validación externa, Big Data**

I. Introducción

El clustering es una técnica de minería de datos que agrupa datos no supervisados en clusters de manera que las instancias que pertenecen al mismo clúster son similares. El clustering se ha usado en diferentes áreas de conocimiento como las ciencias sociales [1], la biología [2], la electricidad [3] o la agricultura [4].

Existen numerosos métodos de clustering en la literatura, y por lo general, cada uno genera una solución de clustering diferente. En algunos casos, se pueden obtener diferentes soluciones con el mismo método con tan solo cambiar alguno de parámetros de entrada. Una de las principales tareas del clustering es el análisis y evaluación de las distintas soluciones. Para medir la calidad de la solución de clustering, existen los llamados índices de validación de clustering (CVI).

Los CVI se podrían dividir en dos categorías: índices internos e índices externos. Los índices internos miden la calidad de la solución en función de la distribución de las instancias por los clusters, es decir, evalúan la separación que existe entre los clusters y la compacidad que hay entre las instancias que pertenecen al mismo clúster. Este tipo de índice es el único que se puede aplicar cuando el dataset no aporta ningún dato adicional. Por otra parte, los índices externos son aquellos que evalúan los clusters en función de algún atributo externo como puede ser la clase. Los índices de este tipo comparan el resultado del clustering con el de una solución global denominada *ground truth*. De esta forma los índices saben a priori la solución óptima así como el número óptimo de clusters del dataset ya que el *ground truth* contiene esta información. Por lo general, los índices de la literatura indican la solución óptima a través de una representación gráfica, y los resultados podrían ser interpretados de manera imprecisa. Los CVI se han usado en [5]–[8]. En este trabajo vamos a centrarnos en los índices de validación externos.

El objetivo de este artículo es presentar un nuevo CVI de clustering externo basado en el test estadístico de chi cuadrado cuyo resultado no necesite ser interpretado. La efectividad de este nuevo índice se ha comparado con 3 índices de la literatura en 6 datasets que podrían ser consideradas big data. La implementación del índice se ha realizado haciendo uso de las librerías MLlib de Spark [9] por lo que el índice estaría preparado para ejecutarse con datasets que podrían considerarse big data.

Este artículo se organiza de la siguiente forma: Sección II trata la literatura sobre los índices externos de validación. En la Sección III se introduce el nuevo índice propuesto en este trabajo. Sección IV presenta los experimentos, la metodología y los resultados obtenidos. Y por último, las conclusiones y trabajos futuros en la Sección V.

II. Trabajos relacionados

Los índices externos evalúan los resultados de un clustering comparándolo con el *ground truth*. Los índices externos se podrían clasificar dependiendo del criterio de comparación del clustering con el *ground truth* [10]. Los índices podrían clasificarse en: *set matching*, *pair-counting* y *information theory*.

- *Set matching* es la categoría que establece que la etiqueta de cada instancia se corresponde con un clúster. Alguno de los índices de la literatura son *purity* [11], *F-measure* [12] y *Goodman-Kruskal* [13].
- Los índices *pair-counting* se basan en la comparación entre el número de instancias con la misma etiqueta y el resultado del clúster. Esta categoría incluye índices como: *rand index* [14], *adjusted rand index* [15], *Jaccard* [16], *Fowlkes-Mallows* [17], *Hubert Statistic* [18] y *Minkowski score* [19].

- Los índices basados en *information theory* como la *entropy* [11], *variation of information* [20] and *mutual information* [21] también se han aplicado en la literatura.

En los últimos años se han publicado en la literatura numerosos estudios que proponen nuevos índices externos para la validación de clusters. En [7] encontramos un nuevo índice *pair-counting* basado en un enfoque probabilístico intuitivo, que se utiliza para comparar soluciones que pueden tener un cierto grado de solape. Este índice fue probado usando 4 datasets artificiales con 6 clases y 4 datasets reales del repositorio UCI [22].

También se presentó un nuevo CVI en [23], pero en este caso, el índice se basa en la distancia Max-Min usando lo que denominan información previa. Este índice externo podría clasificarse en la categoría de *Set matching*. El rendimiento se comparó con índices de tipo *Set matching* y *Counting pairs* utilizando 6 datasets artificiales y 2 datasets reales también del repositorio UCI.

Los autores del trabajo presentado en [24] propusieron un nuevo índice basado en un conjunto de clasificadores supervisados. Podemos clasificar este índice como índice *pair-counting*. Para los experimentos se utilizaron 50 datasets reales del repositorio de la UCI y los resultados se compararon con algunos índices internos.

En [25] se presentó un nuevo CVI *pair-counting* para comparaciones analíticas. Aplica una corrección por azar y normaliza para cada grupo por separado. Los experimentos se llevaron a cabo con datasets artificiales con 3 clases y 6000 instancias cada una. Este nuevo índice obtuvo mejores resultados que otros CVI externos, tales como *purity*, *adjusted rand index* o *mutual information*.

Otros autores sugirieron en [26] un nuevo CVI de concordancia de texto basado en una concepción del grado de libertad que mide el intervalo de decisión entre dos clases. Este índice mide la calidad del clustering comparándolo con índices internos y externos. Se utilizaron 14 datasets reales para probar el rendimiento del índice.

La mayoría de estos CVI se comprueban comparando los resultados de los clusters con algunos de los CVI de la literatura y utilizando datasets sintéticos. Sin embargo, ninguno de estos índices ha sido probado en entornos paralelos y distribuidos utilizando grandes datasets. Este trabajo pretende proporcionar un CVI que permita trabajar con datasets que podrían considerarse big data y basado el test estadístico de chi cuadrado.

A. Chi cuadrado

El test estadístico de chi cuadrado es un método que mide la diferencia entre los valores esperados y los valores observados en una distribución entre dos variables [27]. La siguiente ecuación se utiliza para verificar esta correlación:

$$\chi^2 = \sum_{i}^{r}\sum_{j}^{c}\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \quad (1)$$

donde r es el número de filas, c es el número de columnas, n_{ij} es el valor observado y E_{ij} es el valor esperado. E_{ij} viene dado por

$$E_{ij} = \frac{n_{ij}}{n} \quad (2)$$

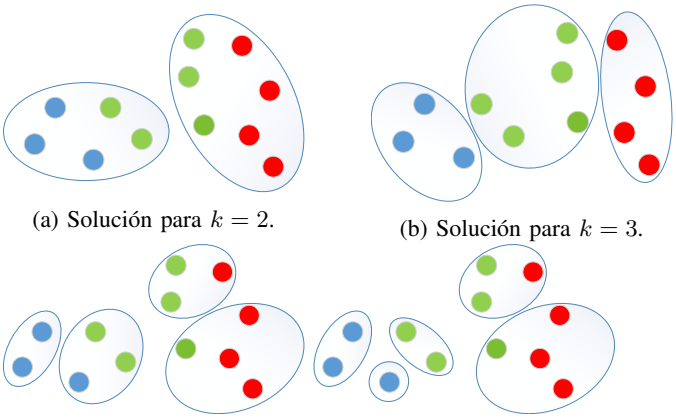
donde n es el número total de instancias. De manera que el valor de χ^2 estará más cerca de cero cuanto más se asemeje el valor observado al valor esperado.

III. Índice de validación de clustering externo basado en chi cuadrado

Supongamos una distribución de 12 instancias y 3 clases tal y como muestra la Figura 1, en la que cada punto representa una instancia y su color define la clase a la que pertenece.

Fig. 1: Representación de un dataset con 12 instancias y 3 clases donde los puntos representan a las instancias y los colores a las clases a las que pertenecen.

Antes de aplicar un método de clustering a este dataset tenemos que decidir el número de clusters (k) en el que lo vamos a dividir. Nuestro objetivo es encontrar el número óptimo de clusters de manera que las instancias que pertenecen a la misma clase queden agrupadas en los mismos clusters, y que los clusters tengan mayoritariamente instancias de una sola clase. La Figura 2 muestra las soluciones de clustering con $k = 2$ hasta $k = 5$, donde k es el número de clusters.



(a) Solución para $k = 2$.

(b) Solución para $k = 3$.

(c) Solución para $k = 4$.

(d) Solución para $k = 5$.

Fig. 2: Resultados de clustering para $k = 2$ hasta $k = 5$.

Para medir objetivamente la calidad de cada solución de clustering necesitaríamos un índice de validación externo. El

denominado *chi index* mide la calidad del clustering basandose en el test estadístico de chi cuadrado. Chi index aplica el test estadístico a la tabla de contingencia generada por la solución de clustering. Una tabla de contingencia muestra la distribución de frecuencias de las instancias, teniendo en cuenta las clases, en forma de matriz.

Siguiendo el ejemplo de la Figura 1, la Tabla I presenta la tabla de contingencia para de la solución de clustering para $k = 2$. Aquí podemos ver que el clúster 1 tiene 3 instancias de la clase azul y 2 instancias verdes, mientras que el clúster 2 tiene 4 instancias rojas y 3 verdes. Esta tabla puede ser analizada desde dos puntos de vista diferentes, teniendo en cuenta que queremos que cada clúster tenga instancias de la misma clase y que las instancias de una misma clase queden agrupadas en mismos clusters:

- Si la analizamos por filas, podemos decir que el clúster 1 está compuesto solo por instancias azules y verdes, sin instancias rojas. Y el clúster 2 está formado por instancias rojas y verdes.
- Si analizamos la tabla por columnas, podemos concluir que las instancias azules están sólo en el clúster 1, las instancias rojas están sólo en el clúster 2, y las verdes están repartidas en ambos clusters.

TABLE I: Tabla de contingencia para la solución de clustering con $k = 2$.

Clúster	Azul	Rojo	Verde	Total
1	3	0	2	5
2	0	4	3	7
Total	3	4	5	12

Estas lecturas se representan en las Tablas II y IIb, que son las tablas de contingencia expresadas con las frecuencias relativas al total de filas (Tabla IIa) y de columnas (Tabla IIb).

TABLE II: Tablas de contingencia relativas para $k = 2$.

(a) Frecuencias relativas al total de cada fila.

Clúster	Azul	Rojo	Verde	Total
1	60%	0%	40%	100%
2	0%	57%	43%	100%

(b) Frecuencias relativas al total de cada columna.

Clúster	Azul	Rojo	Verde
1	100%	0%	40%
2	0%	100%	60%
Total	100%	100%	100%

El siguiente paso sería obtener el valor de chi cuadrado para estas tablas, y realizar el mismo procedimiento en cada iteración.

En este ejemplo de juguete, hemos calculado el índice de chi para las soluciones de clustering desde $k = 2$ hasta $k = 6$. El objetivo es maximizar los valores del índice de chi en ambas tablas y minimizar la diferencia entre ellas. De esta manera, el resultado del índice de chi intentará que los valores observados y esperados sean lo más diferentes posible. Esto obligará a mantener la solución con el porcentaje más alto de clases en cada grupo.

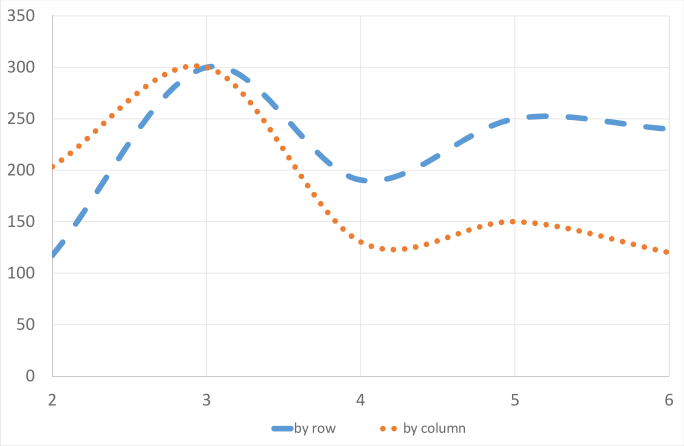


Fig. 3: Chi index result for $k = 2$ to 6.

La figura 3 muestra los resultados de nuestro ejemplo de juguete desde $k = 2$ hasta $k = 6$. Como se puede ver, en $k = 3$ las curvas alcanzaron un máximo, y también ambas curvas tienen el mismo valor. Por lo tanto, la mejor solución de clustering para este conjunto de datos es $k = 3$. Cabe destacar que este índice señala la solución cuando ambas curvas están en la distancia mínima o cruzadas. La interpretación de la solución es bastante simple porque sólo hay que mirar donde se cruzan ambas curvas.

Las tablas III muestran las tablas de contingencia relativas para la solución $k = 3$. Como se puede observar, cada clúster solo tiene instancias de una misma clase (Tabla IIIa), y cada clase está distribuida en un mismo clúster (Tabla IIIb).

TABLE III: Tablas de contingencia relativas para la solución de clustering $k = 3$.

(a) Frecuencias relativas al total de cada fila.

Clúster	Azul	Rojo	Verde	Total
1	**100%**	0%	0%	100%
2	0%	0%	**100%**	100%
3	0%	**100%**	0%	100%

(b) Frecuencias relativas al total de cada columna.

Clúster	Azul	Rojo	Verde
1	**100%**	0%	0%
2	0%	0%	**100%**
3	0%	**100%**	0%
Total	100%	100%	100%

Por lo tanto, *chi index* podría definirse como:

$$chi\,index = \min(\chi^2_{fila} - \chi^2_{columna}) \quad (3)$$

, donde

$$\chi^2_{fila} = \sum_i^r \sum_j^c \frac{(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{ij}}{n})}{\frac{n_{ij}}{n}} \quad (4)$$

$$\chi^2_{columna} = \sum_i^r \sum_j^c \frac{(\frac{n_{ij}}{n_{\cdot j}} - \frac{n_{ij}}{n})}{\frac{n_{ij}}{n}} \quad (5)$$

IV. Resultados

Esta sección describe los experimentos llevados a cabo para testear el CVI propuesto. Para realizar la comparativa se han usado 6 datasets públicos que podrían considerarse big data y se han comparado los resultados con 3 CVI de la literatura. Esta sección se compone de la Sección IV-A donde se detallan los datasets que se han utilizado en los experimentos. La Sección IV-B presenta el diseño de los experimentos seguidos. La Sección IV-D muestra los resultados de los experimentos realizados. La sección IV-D1 incluye un análisis estadístico para probar la efectividad de nuestro índice propuesto para los conjuntos de datos públicos. Finalmente, se incluye una discusión sobre los resultados en la Sección IV-D2.

A. Datasets

La tabla IV muestra los datasets utilizados para los experimentos. La tabla muestra las siguientes características: el nombre del dataset, el número de clases que va a ser usado como el número óptimo de clusters, el número de características y el número de instancias. Todos estos conjuntos de datos fueron descargados del repositorio de machine learning de UCI [22]. Todos los datasets incluyen la etiqueta de la clase, pero ésta no se ha utilizado para el proceso de clustering, solo se ha usado en la etapa de análisis.

TABLE IV: Descripción de los datasets.

Datasets	Clases	Atributos	Instancias
airlines	2	7	539,383
convtype	7	54	581,012
higgs	2	28	11,000,000
kddcup99	2	41	494,020
poker	10	10	829,202
susy	2	12	5,000,000

B. Diseño de experimentos

Para generar las soluciones de clustering, se han aplicado 3 métodos de clustering de Spark incluidos en la librería MLlib [9]: k-means, bisecting k-means y Gaussian mixture.

Estos métodos de clustering necesitan el número de clusters (k) en los que se va a particionar el dataset. Los experimentos se han realizado tomando los valores de k en el intervalo $[D_k-10, D_k+10]$ donde D_k es el número correcto de clusters del dataset siendo $k > 1$. Cada dataset se ha ejecutado con cada uno de estos 3 métodos de clustering con los que se han obtenido un total de 360 soluciones de clustering para probar los CVI. La comparativa se ha realizado entre 3 CVI de la literatura descritos en la Sección II y nuestro *chi index* propuesto.

C. Efectividad del CVI

La efectividad se mide en función de la cercanía a una solución ya dada (*ground-truth*) y realizar una comparativa de los resultados. El primer paso consiste en aplicar el algoritmo de clustering al dataset y obtener las múltiples soluciones. En el segundo paso se evalúan las soluciones de clustering aplicando los CVI. En el tercer y último paso se comparan los resultados del CVI y se selecciona el que mejor puntuación haya obtenido siguiendo.

Para hacer una comparativa entre los diferentes CVI se van a tener en cuenta los siguientes valores:

- Media de aciertos: este valor viene dado por la media de veces que el índice acierta el número óptimo de clusters por el total de datasets.
- Error medio al cuadrado: se calcula como la media de las distancias entre la predicción del índice I_i y el número correcto n_i por el total de datasets:

$$Error = \frac{\sum_{i \in n} d(I_i, n_i)^2}{n} \quad (6)$$

, donde n es el número total de datasets.

1) Test estadísticos: Por último, se ha aplicado un marco estadístico para probar el rendimiento de los CVI. Se ha seleccionado el test no paramétrico de Quade y el procedimiento post-hoc de Holm para validar estadísticamente las diferencias en los rangos medios de los *p-values* correspondientes alcanzados. Este análisis estadístico se realizó utilizando la plataforma de código abierto StatService [28].

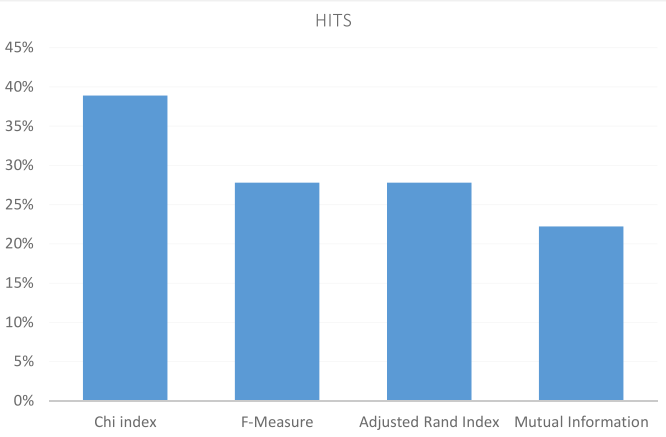
El test de Quade es una prueba estadística no paramétrica que evalúa las diferencias entre más de dos muestras estableciendo un ranking entre ellas. En nuestro caso, las muestras que vamos a evaluar son los CVI que vamos a comparar. Cuanto menor sea el *p-value*, mayor es la confianza de que un CVI funciona correctamente y, por lo tanto, se obtiene una mejor clasificación en el test de Quade.

Después de comprobar que los rankings medios son significativamente diferentes con un valor $\alpha = 0,05$, y siempre que el test de Quade rechace la hipótesis nula, se realizará el test post-hoc de Holm para evaluar el rendimiento relativo de las CVI estudiadas frente a un índice de control, en nuestro caso, tomaremos el que obtenga mejor puntuación en el ranking de Quade.

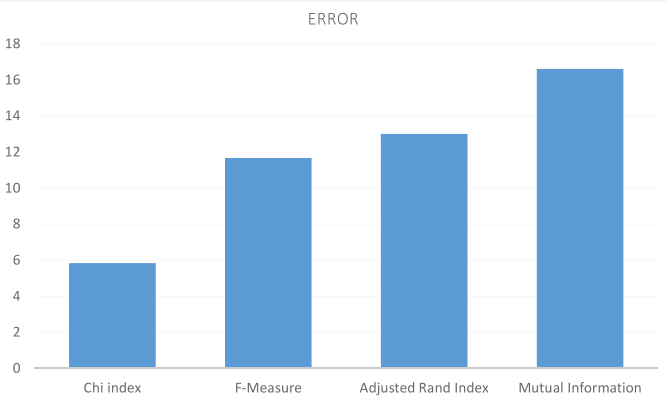
D. Resultados de los experimentos

Esta sección presenta los resultados realizados. La figura 4a muestra la media de aciertos para cada CVI en orden ascendente. *Chi index* ha alcanzado el valor más alto de aciertos (39%) con una diferencia significativa frente a sus competidores. Los índices de la literatura obtuvieron tasas similares de aciertos, que oscilaron entre el 28% en el caso de *F-Measure* y el 22% en el caso de *Mutual Information*.

Por otro lado, la Figura 4b presenta el error medio al cuadrado de cada CVI. Cabe señalar que *chi index* obtuvo el porcentaje de error más bajo (6%) quedando en segunda



(a) Media de aciertos para cada CVI.



(b) Error medio al cuadrado por CVI.

Fig. 4: Comparativa de resultados de los CVI.

posición *F-Measure* con casi el doble de puntos de error. Esto significa que *chi index* acierta el número óptimo de clusters la mayoría de las veces y en caso de error, el valor que indica no se aleja la solución.

1) Análisis estadístico: La Tabla V muestra el ranking del test de Quade para cada CVI. Como se muestra en el ranking, *chi index* ha quedado en primera posición con una puntuación de 1,807. El siguiente índice en el ranking fue *F-Measure* con una puntuación de 2,152, a 0,3 puntos del primer puesto. En última posición ha quedado *Mutual Information* con una puntuación en el ranking de 3,464.

TABLE V: Ranking del test de Quade.

CVI	Ranking
Chi index	1,807
F-Measure	2,152
Adjusted Rand Index	2,576
Mutual Information	3,464

El estadístico de Quade fue de 10,915, distribuida según una distribución F con 3 y 51 grados de libertad. El valor p de Quade fue 0,0 que fue inferior a 0.05. Por lo tanto, rechazó la hipótesis nula de que todos ellos se comportaron de manera similar con un nivel de significación de $\alpha = 0,05$.

Se ha realizado una prueba post-hoc por pares para verificar que nuestra propuesta es significativamente diferente al resto.

TABLE VI: Análisis post-hoc usando el procedimiento de Holm y tomando como índice de control a *chi index*.

CVI	p	z	α_{Holm}
Mutual Information	0,0058	2,760	0,0167
Adjusted Rand Index	0,2003	1,280	0,025
F-Measure	0,5530	0,574	0,050

La tabla VI muestra los *p-values*, el valor z y α_{Holm}, utilizando *chi index* como índice de control al ser el que mejor ranking obtuvo. Como puede verse, la hipótesis nula fue rechazada para todos los CVI. Por lo tanto, podemos concluir que *chi index* generó los mejores resultados (ya que obtuvo el mejor ranking) y fue estadísticamente diferente al resto de CVI.

2) Discusión: Los resultados del análisis de los datasets del repositorio de la UCI muestran que nuestro índice externo mejora la tasa de aciertos en 11% (Figura 4a). Además, en el caso de no poder alcanzar el número correcto de clusters, nuestro índice obtuvo una tasa de 6 puntos menos que los índices de la literatura (Figura 4b). Basado en la prueba de Quade (Tabla V), nuestro índice propuesto mejora los resultados en 2 puntos.

Es interesante resaltar que *chi index* indica la solución de clustering óptima de manera directa y concisa. A diferencia de los CVI de la literatura que necesitan ser interpretados siguiendo el método del codo y localizando máximos o mínimos locales, *chi index* indica la solución de clustering óptimo de manera directa y concisa como vimos en la Sección III.

V. Conclusiones

En este trabajo, se ha propuesto un nuevo índice de validación de clustering externo implementado en Spark para ser aplicado en datasets sin importar su tamaño. Hemos mostrado las diferencias entre nuestra propuesta y los índices de la literatura. El índice propuesto se basa en el test estadístico de chi cuadrado.

El estudio experimental indica que nuestro índice externo es muy competitivo. Hemos probado su efectividad en datasets públicos con un tamaño que podrían ser considerados big data en los que varían el número de clusters, sus características y el número de instancias. Los principales logros obtenidos son los siguientes:

- Un CVI externo basado en el test estadístico de chi cuadrado.
- Nuestro índice nos permitió estimar el número óptimo de clusters basado en la clase del dataset.
- Los resultados de *chi index* son directos y no requieren ser interpretados.
- El índice propuesto está listo para trabajar con conjuntos de datos que pueden ser considerados big data.
- El software de esta contribución se puede encontrar como un paquete de Spark en http://spark-packages.org/package/josemarialuna/externalValidity.

- El código fuente del índice de chi y los otros índices de la literatura se pueden encontrar en https://github.com/josemarialuna/ExternalValidity

Actualmente estamos aplicando este *chi index* en el análisis de datos de empleo y los resultados son prometedores. *Chi index* también se está aplicando en datos eléctricos en colaboración con una compañía eléctrica española. Como trabajo futuro, sería interesante ampliar la aplicación del índice en bases de datos que tengan multiclase.

Agradecimientos

Este trabajo ha sido apoyado por el Ministerio de Economía y Competitividad bajo el proyecto TIN2014-55894-C2-R. J.M. Luna-Romera es becario FPI del Ministerio de Economía y Competitividad.

References

[1] M. Ghane'i-Ostad, H. Vahdat-Nejad, and M. Abdolrazzagh-Nezhad, "Detecting overlapping communities in lbsns by fuzzy subtractive clustering," *Social Network Analysis and Mining*, vol. 8, no. 1, 2018, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85044313063&doi=10.1007%2fs13278-018-0502-5&partnerID=40&md5=96282433476a98b7bfc029f892772fc7

[2] F. Ginot, I. Theurkauff, F. Detcheverry, C. Ybert, and C. Cottin-Bizonne, "Aggregation-fragmentation and individual dynamics of active clusters," *Nature Communications*, vol. 9, no. 1, 2018, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85042217142&doi=10.1038%2fs41467-017-02625-7&partnerID=40&md5=0d3b0fa19ff161e129139ac5de367f18

[3] R. Perez-Chacon, R. L. Talavera-Llames, F. Martinez-Alvarez, and A. Troncoso, "Finding electric energy consumption patterns in big time series data," in *Distributed Computing and Artificial Intelligence, 13th International Conference*, S. Omatu, A. Semalat, G. Bocewicz, P. Sitek, I. E. Nielsen, J. A. García García, and J. Bajo, Eds. Cham: Springer International Publishing, 2016, pp. 231–238.

[4] X. Wu, J. Zhu, B. Wu, J. Sun, and C. Dai, "Discrimination of tea varieties using ftir spectroscopy and allied gustafson-kessel clustering," *Computers and Electronics in Agriculture*, vol. 147, pp. 64–69, 2018, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85042208353&doi=10.1016%2fj.compag.2018.02.014&partnerID=40&md5=bbb11cb2c0d295517af2c497192a4d43

[5] J. Rojas-Thomas, M. Santos, and M. Mora, "New internal index for clustering validation based on graphs," *Expert Systems with Applications*, vol. 86, pp. 334 – 349, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417417304104

[6] J. Hämäläinen, S. Jauhiainen, and T. Kärkkäinen, "Comparison of internal clustering validation indices for prototype-based clustering," *Algorithms*, vol. 10, no. 3, 2017, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85029738018&doi=10.3390%2fa10030105&partnerID=40&md5=15f131f750705ed3aad57f2d3dbba8b1

[7] D. Campo, G. Stegmayer, and D. Milone, "A new index for clustering validation with overlapped clusters," *Expert Systems with Applications*, vol. 64, pp. 549 – 556, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417416304158

[8] Z. Zhang, H. Fang, and H. Wang, "A new mi-based visualization aided validation index for mining big longitudinal web trial data," *IEEE Access*, vol. 4, pp. 2272–2280, 2016, cited By 7. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84979828536&doi=10.1109%2fACCESS.2016.2569074&partnerID=40&md5=cb527399d9c0c9990be218434656d657

[9] A. Spark, "Clustering - Spark 2.2.0 Documentation," https://spark.apache.org/docs/2.2.0/ml-clustering.html, 2018, [Online; accessed 6-april-2018].

[10] M. Rezaei and P. Fränti, "Set matching measures for external cluster validity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2173–2186, Aug 2016.

[11] Y. Zhao and G. Karypis, "Criterion functions for document clustering: Experiments and analysis," Tech. Rep., 2002.

[12] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '99. New York, NY, USA: ACM, 1999, pp. 16–22. [Online]. Available: http://doi.acm.org/10.1145/312129.312186

[13] L. A. Goodman and W. H. Kruskal, *Measures of Association for Cross Classifications*. New York, NY: Springer New York, 1979, pp. 2–34. [Online]. Available: https://doi.org/10.1007/978-1-4612-9995-0_1

[14] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971. [Online]. Available: http://www.jstor.org/stable/2284239

[15] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Is a correction for chance necessary?" in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 1073–1080. [Online]. Available: http://doi.acm.org/10.1145/1553374.1553511

[16] R. Sokal and P. Sneath, *Principles of Numerical Taxonomy*, ser. Books in biology. W. H. Freeman, 1963. [Online]. Available: https://books.google.es/books?id=3Y4aAAAAMAAJ

[17] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 553–569, 1983.

[18] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, Dec 1985. [Online]. Available: https://doi.org/10.1007/BF01908075

[19] A. Ben-Hur and I. Guyon, *Detecting Stable Clusters Using Principal Component Analysis*. Totowa, NJ: Humana Press, 2003, pp. 159–182. [Online]. Available: https://doi.org/10.1385/1-59259-364-X:159

[20] M. Meilă, "Comparing clusterings by the variation of information," in *Learning Theory and Kernel Machines*, B. Schölkopf and M. K. Warmuth, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 173–187.

[21] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von mises-fisher distributions," *J. Mach. Learn. Res.*, vol. 6, pp. 1345–1382, Dec. 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1046920.1088718

[22] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[23] A. Alok, S. Saha, and A. Ekbal, "Development of an external cluster validity index using probabilistic approach and min-max distance," vol. 6, pp. 494–504, 06 2012.

[24] J. Rodríguez, M. Medina-Pérez, A. Gutierrez-Rodríguez, R. Monroy, and H. Terashima-Marín, "Cluster validation using an ensemble of supervised classifiers," *Knowledge-Based Systems*, vol. 145, pp. 1–14, 2018.

[25] M. Rezaei and P. Fränti, "Set matching measures for external cluster validity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2173–2186, Aug 2016.

[26] C. Liu, W. Wang, M. Konan, S. Wang, L. Huang, Y. Tang, and X. Zhang, "A new validity index of feature subset for evaluating the dimensionality reduction algorithms," *Knowledge-Based Systems*, vol. 121, pp. 83 – 98, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705117300291

[27] J. Antoch, "A Guide to Chi-Squared Testing : Greenwood, P.E. and Nikulin, M.S. New York: John Wiley & Sons, Inc.,.pp. 280 +XII, ISBN 0-471-55779-X. AMS 1991 Classification: 62-02, 62F03, 62H15," *Computational Statistics & Data Analysis*, vol. 23, no. 4, pp. 565–566, February 1997. [Online]. Available: https://ideas.repec.org/a/eee/csdana/v23y1997i4p565-566.html

[28] J. A. Parejo, J. Garcia, A. Ruiz-Cortes, and J. C. Riquelme, "Statservice: Herramienta de análisis estadístico como soporte para la investigación con metaheurísticas," in *Actas del VIII Congreso Expañol sobre Metaheurísticas, Algoritmos Evolutivos y Bio-inspirados*, 2012.

Atipicidad: Medida de calidad clave dentro del descubrimiento de reglas descriptivas supervisadas

C.J. Carmona[1], M.J. del Jesus[1], F. Herrera[2]
[1]*Departamento de Informática. Universidad de Jaén, España* ccarmona@ujaen.es, mjjesus@ujaen.es
[2]*Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada, España* herrera@decsai.ugr.es
[1,2]*Instituto Data Science and Computational Intelligence (DaSCI)*

***Resumen*—Esto es un resumen de nuestro trabajo publicado en la revista *Knowledge-Based Systems* [1] para su presentación en la Multiconferencia CAEPIA'18 Key Works.**

***Index Terms*—Descubrimiento de reglas descriptivas supervisadas, Descubrimiento de subgrupos, Conjuntos de contraste, Patrones emergentes, Atipicidad.**

I. Resumen

A lo largo de la literatura podemos encontrar un conjunto de técnicas que se encuentran a medio camino entre la predicción y la descripción, agrupadas bajo el nombre de descubrimiento de reglas descriptivas bajo aprendizaje supervisado (SDRD) [1], [2]. Este conjunto de técnicas intenta obtener reglas de una categoría o clase prefijada para describir información significativa y relevante del conjunto de datos.

El principal objetivo de las técnicas dentro de SDRD no es extraer un modelo para clasificar nuevas instancias, sino obtener un modelo que permita comprender, describir o encontrar fenómenos subyacentes de interés en los datos. Dentro de este área de investigación se agrupan todas aquellas técnicas que, mediante el uso de reglas y un modelo de aprendizaje supervisado, intentan obtener conocimiento descriptivo de los datos que sea significativo, inusual y de interés para el usuario, como el descubrimiento de subgrupos [3], la minería de patrones emergentes [4] y los conjuntos de contraste [5], entre otros. A continuación se describen brevemente estas técnicas.

Descubrimiento de subgrupos: Se define como el descubrimiento de subgrupos de la población estadísticamente interesantes, esto es, tan grandes como sea posible y con una distribución estadística de la propiedad de interés lo más atípica posible respecto al conjunto de la población. La medida de calidad para medir esta atipicidad [6] en una regla R se define como:

$$WRAcc(R) = \frac{p+n}{P+N}\left(\frac{p}{p+n} - \frac{P}{P+N}\right) \quad (1)$$

Conjuntos de contraste: Es una técnica de exploración para contrastar diferencias de grupos dentro de un conjunto de datos, es decir, descubrir conjuntos de ejemplos con amplias diferencias de soporte entre grupos del conjunto de datos y se mide mediante la diferencia de soporte [5]:

$$DS(R) = \left|\frac{p}{P} - \frac{n}{N}\right| \ \geq \ \delta \quad (2)$$

Patrones emergentes: Se centra en buscar conocimiento relacionado con los valores de una clase, donde el número de instancias cubiertas por un patrón sea muy elevado para un valor de la variable objetivo y muy bajo o nulo para el resto; es decir, que el mismo patrón tenga un soporte muy alto para una clase y muy bajo para las demás clases del problema, y viene dado por el índice de crecimiento [4] que se representa a continuación:

$$GR(R) = \frac{\frac{p}{P}}{\frac{n}{N}} > 1 \quad (3)$$

Centrándonos en la importancia de la medida de atipicidad hay que destacar que mide el equilibrio entre generalidad y precisión. En concreto, el segundo factor de la atipicidad es el factor que mide la ganancia de precisión, y tal y como se puede observar, es posible que la aticipidad de una regla sea inferior a cero cuando la regla tiene una baja calidad ya que el porcentaje de ejemplos negativos cubiertos es superior al de los positivos. En este sentido, una regla de interés para el experto debería tener siempre un valor positivo, es decir, se debe cumplir la siguiente desigualdad para obtener una regla de interés:

$$\frac{p}{p+n} > \frac{P}{P+N} \quad (4)$$

Si descomponemos esta desigualdad [1] podemos decir que para que una regla obtenga una atipicidad positiva y ser un subgrupo de interés, obtenemos que el producto escalar entre $p\overline{n}$ debe ser superior que $\overline{p}n$. En consecuencia, esta descomposición nos indica que el producto de ejemplos cubiertos y no cubiertos correctamente debería ser superior al producto de ejemplos cubiertos y no cubiertos incorrectamente:

$$p\ \overline{n} > \overline{p}\ n \quad (5)$$

En el trabajo publicado en [1] se presenta tras este primer análisis, la compatibilidad entre el descubrimiento de subgrupos, los conjuntos de contraste y los patrones emergentes gracias al uso de esta medida de calidad.

Por ejemplo, en el caso de los patrones emergentes (según definición) podemos afirmar que una regla es emergente cuando su índice de crecimiento es superior a 1, es decir:

$$GR(R) = \frac{\frac{p}{P}}{\frac{n}{N}} > 1 \quad (6)$$

Si aplicamos una descomposición a esta formulación descubrimos que:

$$\frac{\frac{p}{P}}{\frac{n}{N}} > 1$$
$$p\ \overline{n} > \overline{p}\ n \quad (7)$$

Según esta demostración podemos afirmar que la descomposición de la atipicidad positiva y del índice de crecimiento son similares, es decir, si un subgrupo es de interés también es un patrón emergente. Además, por otro lado se puede observar una relación directa entre la atipicidad y la diferencia de soporte en [2] que dice:

$$DS(R) = \frac{WRAcc(R)}{p(PIS) \cdot p(NIS)} \geq \delta \quad (8)$$

donde $p(PIS)$ es el porcentaje de ejemplos positivos para la clase del conjunto de datos a analizar y $p(NIS)$ es el porcentaje de ejemplos negativos para el conjunto de datos.

Al mismo tiempo, es importante tener en cuenta que el dominio de la atipicidad para un problema depende del porcentaje de los ejemplos para la clase o variable objetivo [7], y por tanto, el dominio se puede determinar mediante las siguientes ecuaciones para el límite inferior (LB_{WRAcc}) y para el superior (UB_{WRAcc}):

$$LB_{WRAcc} = (1 - p(PIS)) \cdot (0 - p(PIS)) \quad (9)$$

$$UB_{WRAcc} = p(PIS) \cdot (1 - p(PIS)) = p(PIS) \cdot p(NIS) \quad (10)$$

Es decir, el valor de DS depende de las propiedades del problema ya que está directamente relacionado con el porcentaje de ejemplos de la clase a analizar, y en resumen:

$$DS(R) = \frac{WRAcc(R)}{p(PIS) \cdot p(NIS)} = \frac{WRAcc(R)}{UB_{WRAcc}} \geq \delta \quad (11)$$

Una regla se considera conjunto de contraste con un $\delta = 0{,}10$ cuando:

$$WRAcc(R) \geq 0{,}10 \cdot UB_{WRAcc} \quad (12)$$

Sin embargo, tenemos la necesidad de estandarizar este valor de $WRAcc$ a $WRAcc$ normalizada ($WRAccN$) ya que debemos evitar la utilización de un valor δ que esté condicionado a este porcentaje de ejemplos de la clase a analizar. Esta mejora se consigue con respecto a la homogeneización de esta medida clave dentro del descubrimiento de reglas descriptivas supervisadas. La $WRAccN$ se normaliza en el intervalo [0,1] mediante la siguiente ecuación donde una regla es subgrupo de interés cuando su valor sea superior a 0.5.

$$WRAccN(R) = \frac{WRAcc(R) - LB_{WRAcc}}{UB_{WRAcc} - LB_{WRAcc}} \quad (13)$$

Considerando un δ igual a 0.10, y el intervalo positivo de $WRAccN$ igual a (0.5, 1.0], una regla se considera conjunto de contraste con un $\delta = 0{,}10$ cuando:

$$WRAccN(R) \geq 0{,}50 + 0{,}10 \cdot (1{,}00 - 0{,}50)$$
$$WRAccN(R) \geq 0{,}55 \quad (14)$$

, es decir, cuando el valor de la $WRAccN$ es superior o igual que el 10 % de la parte positiva de $WRAcc$.

En conclusión, la atipicidad es un factor clave a medir dentro del descubrimiento de reglas descriptivas supervisadas donde una regla con un valor de $WRAccN$ superior a 0.50 se considera de interés para el descubrimiento de subgrupos, y emergente para la extracción de patrones emergentes. Además, si la $WRAccN$ es superior o igual a 0.55 se considera también regla de contraste. Esta demostración muestra la relación directa entre todas estas técnicas que se han agrupado dentro del descubrimiento de reglas descriptivas supervisadas y muestra la importancia de medir la calidad de una regla con respecto a la $WRAcc$ en un problema que se desea analizar. Por otro lado, es interesante indicar que solo con el análisis de esta medida de calidad, los expertos serán capaces de determinar si la regla representa un subgrupo, una regla emergente y/o una regla de contraste. De la misma forma, esta medida debe ser clave para el diseño de nuevas propuestas dentro del descubrimiento de reglas descriptivas supervisadas.

Agradecimientos

Este trabajo ha sido subvencionado por el Ministerio de Economía y Competitividad de España bajo el proyecto TIN2014-916 57251-P y TIN2015-68454-R (Fondos FEDER).

Referencias

[1] C. J. Carmona, M. J. del Jesus, and F. Herrera, "A Unifying Analysis for the Supervised Descriptive Rule Discovery via the Weighted Relative Accuracy," *Knowledge-Based Systems*, vol. 139, pp. 89–100, 2018.

[2] P. Kralj-Novak, N. Lavrac, and G. I. Webb, "Supervised Descriptive Rule Discovery: A Unifying Survey of Constrast Set, Emerging Pateern and Subgroup Mining," *Journal of Machine Learning Research*, vol. 10, pp. 377–403, 2009.

[3] W. Kloesgen, "Explora: A Multipattern and Multistrategy Discovery Assistant," in *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, 1996, pp. 249–271.

[4] G. Z. Dong and J. Y. Li, "Efficient Mining of Emerging Patterns: Discovering Trends and Differences," in *Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, 1999, pp. 43–52.

[5] S. D. Bay and M. J. Pazzani, "Detecting group differences: Mining contrast sets," *Data Mining and Knowledge Discovery*, vol. 5, no. 3, pp. 213–246, 2001.

[6] N. Lavrac, P. A. Flach, and B. Zupan, "Rule Evaluation Measures: A Unifying View," in *Proc. of the 9th International Workshop on Inductive Logic Programming*, ser. LNCS, vol. 1634. Springer, 1999, pp. 174–185.

[7] C. J. Carmona, V. Ruiz-Rodado, M. J. del Jesus, A. Weber, M. Grootveld, P. González, and D. Elizondo, "A fuzzy genetic programming-based algorithm for subgroup discovery and the application to one problem of pathogenesis of acute sore throat conditions in humans," *Information Sciences*, vol. 298, pp. 180–197, 2015.