Revista Cubana de Ciencias Informáticas
Vol. 14, No. 3, Mes Julio-Septiembre, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 18-40

Artículo original

Nuevo método para el descubrimiento de subgrupos no redundantes

A new method for not redundant Subgroup Discovery

Lisandra Bravo Ilisastigui[1*] https://orcid.org/0000-0002-8209-4121
Diana Martín Rodriguez[1] https://orcid.org/ 0000-0001-9188-3926
Milton García Borroto[1] https://orcid.org/0000-0002-3154-177X

[1]Facultad de Informática de la Universidad Tecnológica de la Habana José Antonio Echeverría, CUJAE. {lbravo, dianamartin85, mgarcia}@ceis.cujae.edu.cu

*Autor para la correspondencia. (lbravo@ceis.cujae.edu.cu)

RESUMEN

El descubrimiento de subgrupos es una tarea de la Minería de Datos que tiene como objetivo identificar subconjuntos de ejemplos con un comportamiento inusual con respecto a una característica de interés. Un problema que puede afectar la comprensibilidad de los modelos obtenidos por los métodos de descubrimiento de subgrupos, es la redundancia. En este artículo se presenta DINOS, un algoritmo para la extracción de subgrupos no redundantes y con alta inusualidad en forma de reglas cuantitativas. Para ello se emplea un algoritmo genético multiobjetivo que permite optimizar inusualidad, sensibilidad, confianza y comprensibilidad, mientras realiza un aprendizaje evolutivo de los intervalos de los atributos que intervienen en las reglas de subgrupos. Además, este algoritmo emplea criterios para determinar redundancia basados en cobertura y los intervalos de confianza del oddratio, para filtrar los subgrupos

redundantes. Un estudio experimental basado pruebas no paramétricas y una comparación pareada con los modelos obtenido por varios algoritmos del estado del arte demuestra la novedad y validez de la propuesta. Los resultados muestran que DINOS obtiene los mejores valores en las métricas estudiadas entre los algoritmos involucrados en el estudio.

Palabras claves: Descubrimiento de Subgrupos; redundancia; algoritmo genético.

ABSTRACT

Subgroup Discovery is a Data Mining task to identify descriptions of subsets of a data set that show an interesting behavior with respect to certain interestingness criteria. A major problem that affect the comprehensibility of the results is the redundancy feature. Dependencies between the non-target attributes lead to large numbers of variations of a particular subgroup. Since many descriptions can have a similar coverage of the given data. In this work a new algorithm for the description induction of not overlapped subgroups (DINOS) is proposed. This algorithms is able to automatically determine the intervals for the numerical attributes of a data set. The experimental study shows that DINOS obtains subgroups with high quality that improves the results reported in the literatura.

Keywords: Subgroup Discovery; redundancy; genetic algorithm.

Recibido: 06/04/2020

Aceptado: 05/05/2020

INTRODUCCIÓN

El descubrimiento de subgrupos es una tarea descriptiva de la Minería de Datos asociada a la inducción supervisada de reglas (Herrera et al. 2011), (Novak et al. 2009). Definido inicialmente por Klosgen en EXPLORA (Klösgen, 1996) y Wroble en MIDOS (Wrobel, 1997) como: “Dada una población de individuos y una propiedad específica de ese individuo que sea de interés, encontrar un subgrupo de esa población que sea estadísticamente más interesantes”, dígase lo más grande posibles, y con la distribución más inusual de la característica de interés (clase). Es por ello que los algoritmos de descubrimiento de subgrupos tienen como objetivo minar patrones locales, que describan conjuntos de ejemplos de una base de datos de forma tal que se maximice el tamaño y la inusualidad de la distribución estadística en este.

La calidad de un subgrupo está asociada a la inusualidad y a otras métricas clásicas en métodos de inducción de reglas, como son la confianza y la sensibilidad. Por esta razón se han propuesto métodos evolutivos multiobjetivo que permitan extraer subgrupos con un buen balance en las medidas de calidad (Berlanga et al., 2006; Carmona et al., 2010). Estas soluciones han probado una alta eficacia en el descubrimiento de subgrupos, la mayoría empleando etiquetas difusas para tratar con variables continuas. El trabajo con lógica difusa ha permitido ganar en precisión y comprensibilidad del conocimiento obtenido, pero el empleo de la misma necesita un preprocesamiento de la base de datos, que requiere de conocimiento del usuario para que los resultados sean realmente útiles. En otras tareas de inducción de reglas como Reglas de Asociación, se han desarrollado varios algoritmos para el minado de reglas cuantitativas, que permiten encontrar asociaciones entre intervalos de variables continuas sin información previa (Martin et al., 2014; Martín et al., 2016).

Un problema común de los métodos de obtención de subgrupos es el hecho de que los algoritmos pueden estar minando más de un descriptor para un mismo subgrupo. Estos descriptores se conocen como redundantes y pueden disminuir la comprensibilidad del modelo. También la evaluación del conjunto final se puede ver sesgada por la redundancia (van Leeuwen and Knobbe, 2012). En el estado del arte se presentan algunas estrategias para intentar eliminar la redundancia

entre descripciones de subgrupos (Boley and Grosskreutz,2009; Li et al., 2014). La mayoría de estas soluciones solo son capaces de trabajar con atributos nominales, por lo que las bases de datos con atributos numéricos deben ser discretizadas, lo que podría generar pérdida de información.

En este trabajo se presenta DINOS, un algoritmo para el minado descripciones de subgrupos no redundantes de alta calidad sobre datos numéricos sin discretización previa. La propuesta tiene como base un algoritmo genético multiobjetivo que optimiza, la inusualidad, confianza, sensibilidad y la comprensibilidad. Además, realiza un aprendizaje evolutivo de los atributos que describen al subgrupo, así como de los intervalos del dominio donde es válido el subgrupo. Este algoritmo incorpora un mecanismo novedoso para el filtrado de subgrupos redundantes que se basa en la combinación del análisis de la cobertura de los subgrupos, y la diferencia estadística que existe entre las distribuciones de las clases en los subgrupos.

ANTECEDENTES

El objetivo de los algoritmos de descubrimiento de subgrupos es minar descriptores en formas de reglas del tipo *Cond* → *Clase*; donde *Cond* contiene las restricciones que definen al subgrupo y *Clase* representa el valor de la variable de interés con mejor representación en este subconjunto de datos, maximizando la inusualidad. Las métricas de evaluación para los subgrupos miden la calidad de cada subgrupo, de forma individual. Entre las más conocidas están:

Confianza: Representa la frecuencia con que se puede encontrar el valor de la *Clase* en los ejemplos que pertenecen al subgrupo descrito por la regla. Se calcula según la ecuación 1.

$$Conf\ (A \rightarrow B) = p(B|A) \quad (1)$$

Sensibilidad: Expresa la probabilidad que se cumplan las condiciones que describen al subgrupo en los ejemplos pertenecientes a la *Clase* Se calcula como indica la ecuación 2.

$$Sens\ (A \rightarrow B) = p(A|B) \quad (2)$$

Inusualidad: También conocida como precisión relativa ponderada, que mide el balance entre la ganancia de precisión y la cobertura de la regla. También se emplea como medida de interés y se calcula según la ecuación 3

$$WRAcc\ (A \rightarrow B) = p(A)(p(B|A) - p(B)) \quad (3)$$

Los métodos de comparación entre algoritmos se basan en estas métricas para determinar el que tiene mejor rendimiento en un marco determinado. La forma más empleada (García et al., 2010), realiza una validación cruzadas de 10 partes y se promedian los valores de cada métrica para los modelos obtenidos en cada partición. El valor de la métrica para cada modelo se obtiene a su vez promediando los valores de las métricas los descriptores minados. Con estos resultados se aplican métodos estadísticos no paramétricos para determinar si existen diferencias significativas entre los resultados de los algoritmos.

Por otra parte, en (Ilisástigui et al., 2019) se propone un método que brinda mayor información sobre el comportamiento de los algoritmos con apoyo visual de gráficos, que además permite medir y neutralizar el efecto de la redundancia en la comparativa. En este método también se emplea la validación cruzada y comienza por medir y eliminar la redundancia. A continuación, se realizan comparaciones por pares de los modelos obtenidos para la misma partición de la base de datos entre los algoritmos a estudiar. Para cada par se detectan aquellos descriptores que cubren el mismo subgrupo en ambos modelos, para determinar la cantidad de subgrupos que fueron capaces de detectar ambos (comunes) y los que fueron detectado solo por alguno de los dos. Esta prueba permite determinar la similitud entre los resultados de los algoritmos, además de reflejar la cantidad de conocimiento que uno de los algoritmos es capaz de extraer con respecto al otro.

Por otra parte, en (Ilisástigui et al. 2019) se propone un método que brinda mayor información sobre el comportamiento de los algoritmos con apoyo visual de algunos gráficos, además de medir y

tratar la redundancia. En este método también se emplea la validación cruzada y comienza por eliminar la redundancia. A continuación, se realizaron comparaciones por pares de los modelos obtenidos para la misma partición de la base de datos de DINOS y cada uno de los algoritmos seleccionados. Para cada par se detectan aquellos descriptores que cubren el mismo subgrupo en ambos modelos (comunes), para determinar cuántos subgrupos fueron capaces de detectar ambos y cuantos alguno de ellos. Esta prueba permite determinar la similitud entre los resultados de los algoritmos, además de reflejar la cantidad de conocimiento que uno de los algoritmos es capaz de extraer con respecto al otro.

Luego se realiza un análisis también pareado del comportamiento de las métricas. Para ello se determinan los valores máximo y mínimo de las métricas con las que se evalúa la calidad de los subgrupos de ambos modelos (generalmente Inusualidad, Confianza y Sensibilidad). Se establece un intervalo con estos valores que se divide en tres partes iguales. Luego se determina el número de patrones de cada modelo que están en cada porción. De esta forma se puede apreciar la calidad de los patrones obtenidos por cada uno de los modelos. Luego se promedian los resultados de las comparaciones de todas las particiones de la misma base de datos y se muestran en una gráfica de barras apiladas, representando los porcientos de patrones que se encuentran en cada intervalo del dominio de la métrica estudiada.

Detección de redundancia en subgrupos

La redundancia en el descubrimiento de subgrupos es la existencia de una o más reglas que están describiendo el mismo comportamiento de la variable objetivo en conjuntos de datos con alto grado de solapamiento. En (van Leeuwen and Knobbe, 2012) se plantea que pequeños cambios en las descripciones de los subgrupos implican subgrupos casi iguales. Las reglas redundantes en los modelos de subgrupos pueden ocasionar varias dificultades. En primer lugar, está el hecho de que se devuelve a los usuarios más información de la necesaria para identificar subgrupos. Este hecho puede generar interpretaciones erróneas del modelo, principalmente en la cantidad de subgrupos

encontrados. Por otra parte, una cantidad grande de reglas puede hacer ilegible o poco práctico el modelo. El segundo problema está relacionado con la selección de los “k” mejores descriptores. Los subgrupos de alta calidad suelen tener más de una relación de atributos que los describe por lo que sus descriptores van a quedar entre los k-mejores. De esta forma se van a devolver descripciones repetidas y se eliminarán aquellas con menor valor en las métricas de evaluación, aun cuando podrían ser interesantes. Al existir descripciones repetidas, lo métodos de comparación basados en el promedio de los valores de las métricas de los subgrupos encontrados se ven afectados al tomar en cuenta más de una vez los mismos subgrupos.

En (van Leeuwen and Knobbe, 2012) se plantea que existen tres niveles de complejidad para tratar la redundancia en los subgrupos, cada uno más restrictivo que el anterior. El primero se conoce como nivel de descripciones que solo analiza si las condiciones presentes en las reglas de subgrupos se solapan. Por ejemplo, las reglas Embarazada → diabetes y Mujer ∧ Embarazada → diabetes comparten atributos y ambas hablan del mismo conjunto de persona. El segundo es el nivel de cobertura y analiza el solapamiento en los ejemplos cubiertos por los subgrupos. Para ello se pueden tomar en cuenta solo reglas que cubren exactamente los mismos ejemplos, o se pueden relajar la restricción determinando un porciento de solapamiento. El tercero se denomina nivel de excepciones y tiene en cuenta que pueden existir subgrupos pequeños dentro de otros subgrupos más generales que describen una variación local de la probabilidad (Atzmueller, 2015).

En (Li et al., 2014) para determinar si dos subgrupos son redundantes, comprueba que tengan parte del ante- cedente igual. Luego determinan si las diferencias que presentan en el antecedente influyen en la distribución de los valores de la variable objetivo. Si la diferencia no es significativa entonces son redundantes. Para ello emplean el cálculo del oddratio (Tan et al., 2004) y de los intervalos de confianza para el mismo (Li et al.,2005).

El oddratio es muy empleada en estudios estadísticos para medir diferencias en la probabilidad de ocurrencia de fenómenos para dos poblaciones distintas. El cálculo de los intervalos de confianza

de los valores de oddratio de las poblaciones permite definir si la diferencia es significativa. Cuando los intervalos de confianza se solapan, la diferencias entre la variación de probabilidad de ambas poblaciones no es significativa (Fleiss et al., 2013), que en el caso de dos subgrupos significaría que son redundantes.

NUEVO MÉTODO PARA EL DESCUBRIMIENTO DE SUBGRUPOS NO REDUNDANTES

DINOS es un algoritmo evolutivo multiobjetivo para el minado de descriptores de subgrupos no redundantes de alta calidad en atributos con dominios continuos. La propuesta tiene dos fases importantes, la obtención de un conjunto reducido de descriptores de subgrupos de alta calidad, y la selección de los no redundantes. En la primera fase se emplea una adaptación del modelo evolutivo del algoritmo de reglas de asociación MOPNAR (Martin et al., 2014) para el minado de subgrupos. De esta forma DINOS hereda la capacidad de obtener un conjunto reducido de reglas cuantitativas, positivas y negativas de alta calidad, que determina los intervalos de los atributos continuos durante el proceso evolutivo basado en el comportamiento de las métricas a optimizar. Estos intervalos pueden ser positivos o negativos, lo permite representar en un mismo descriptor intervalos discontinuos.

La fase de selección de reglas no redundantes emplea dos criterios, uno de cobertura y otro estadístico que permite filtrar los descriptores de subgrupos que cubren el mismo conjunto de datos, sin desechar los que pueden estar minando subgrupos excepcionales.

Selección de descripciones no redundantes

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 3, Mes Julio-Septiembre, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 18-40

Para realizar la selección de los descriptores no redundantes se debe identificar en primer lugar aquellos que representan el mismo conocimiento. Para ellos se emplean dos criterios: cubrimiento común de ejemplos, y redundancia estadística. El criterio basado en el cubrimiento de objetos cuantifica el solapamiento que puede existir entre los elementos que soportan dos descriptores. Para ellos se empleó una métrica de solapamiento presentada en (Martín et al., 2016) para la búsqueda de nichos. Esta métrica permite cuantificar la proporción de elementos de una regla que son cubiertos. Para ello determina el máximo entre las razones de los elementos comunes, y los elementos soportados por cada regla. Los valores de la métrica van de “0”, no se solapan, a “1” una regla es subconjunto de la otra, y se calcula como muestra la ecuación 4.

$$solp(S_a, S_b) = MAX\left[\frac{cov(S_a \wedge S_b)}{cov(S_a)}, \frac{cov(S_a \wedge S_b)}{cov(S_b)}\right] \quad (4)$$

Valores de solapamiento mayores al 50 % indican que una de las reglas está cubriendo más de la mitad de los ejemplos de la otra. En estos casos es más probables que los descriptores estén describiendo un comportamiento similar de la variable objetivo. Para la selección se emplea un umbral de solapamiento, *solev*, determinado por el usuario que permite regular el nivel de solapamiento que se va a considerar para la redundancia.

La métrica empleada para determinar la redundancia de cobertura, no distingue las excepciones dentro de un subgrupo. Para evitar descartar estos interesantes subgrupos se aplica el criterio de redundancia estadística para aquellos descriptores que sobrepasan el umbral de solapamiento. El criterio de redundancia estadística (Li et al. 2014) se basa en el cálculo del *OddRatio* que para un subgrupo ”P” se calcula como: $OR(P) = \frac{TP*TN}{FP*FN}$
sean *TP*, *FP*, *FN*, *TN* los términos de la tabla de contingencia. Los intervalos de confianza se calculan como

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 3, Mes Julio-Septiembre, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 18-40

$[OR(P)e^{-w}, OR(P)e^{w}]$, $w = z_{\alpha/2} * \sqrt{\frac{1}{TP} + \frac{1}{FP} + \frac{1}{FN} + \frac{1}{TN}}$ donde . El valor de corte de la prueba para un 95% de confianza es $z_{\alpha/2} = 1{,}96$.

Modelo Evolutivo

DINOS emplea el modelo evolutivo multiobjetivo basado en descomposición MOEA-D/DE (Tan et al., 2012). Este descompone el problema de optimización multiobjetivo en N subproblemas escalares, que optimizan una agregación distinta de todos los objetivos. Además, emplea alguna de las adaptaciones al modelo MOEA-D/DE que propone MOPNAR (Martin et al., 2014) para la extracción de reglas positivas y negativas.

Para promover diversidad en las soluciones y evitar el estancamiento en máximos locales, se emplea un mecanismo de reinicialización y una población externa (PE). Cuando el porciento de nuevas soluciones en una generación, es menor a un umbral α (α definido por el usuario, generalmente 5 %), se resguardan en la PE las soluciones no dominadas encontradas hasta el momento y se reinicia la población. Los nuevos individuos se generan a partir de los ejemplos no cubiertos por las soluciones en la PE. Nótese que en la PE van a mantenerse las soluciones no dominadas durante toda la ejecución del algoritmo y que serán las que conformen el resultado final.

En el proceso de selección de soluciones no dominadas se emplea el mismo mecanismo de análisis de cobertura que en el criterio de redundancia basado en la cobertura propuesto en la sección Detección de redundancia. En este caso se emplea para evitar descartar subgrupos que pueden ser interesantes, pero la distribución de los ejemplos que cubre en la base de datos no permite mejores valores en las métricas que las del mejor subgrupo encontrado. Para ello, las soluciones que no comparten al menos el 50 % de los ejemplos no se consideran que se dominen entre sí, aunque uno de ellos tenga mejores métricas. Además, las soluciones de la PE no deben ser redundantes entre sí, por los que se aplican los criterios de selección de reglas no redundantes a los elementos de las misma con solev = 0,75.

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 3, Mes Julio-Septiembre, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 18-40

Así el modelo comienza por generar un vector de pesos para cada subproblema, los cuales se emplean en el cálculo del valor de las soluciones en el enfoque de descomposición. Luego para cada vector se determinan su vecindad, formada por los T vectores de pesos más cercanos. A continuación, el algoritmo genera una población inicial, fija los valores de los objetivos a optimizar con los mejores encontrados hasta el momento y actualiza la PE con las reglas no dominadas encontradas en la población inicial. Luego por cada individuo se generan dos hijos aplicando los operadores de selección, mutación y reparación. La selección de la pareja se realiza de forma aleatoria con una probabilidad δ de entre las soluciones con vectores de pesos vecinos, o del resto de la población (δ definido por el usuario). A las nuevas soluciones se les asigna la clase que mejor representan para luego evaluarlas. De ser pertinente, los hijos pueden sustituir a los individuos de la generación previa con peores valores en el enfoque de descomposición, y actualizar los valores de referencia de los objetivos a optimizar por los mejores hasta el momento. Es importante tener en cuenta que el máximo número de soluciones que se pueden sustituir por una solución hija es limitado y debe ser mucho menor que T, para no afectar la diversidad en la vecindad. Por último, se verifica si es necesario reiniciar la población y actualizar la población externa. Este proceso de repite hasta alcanzar la cantidad de evaluaciones de la función objetivo deseada.

DINOS maximiza tres objetivos, el primero es la inusualidad medida con la precisión relativa ponderada (WRAcc), que se reporta en la literatura permite obtener subgrupos interesantes. Para garantizar la calidad de las soluciones solo se aceptan en la PE aquellas con *WRAcc* > 0. El segundo objetivo es una combinación entre confianza y sensibilidad, y se calcula $Obj_2 = Conf * Sens$; de esta forma se busca obtener subgrupos con equilibrio entre la precisión y la generalidad. El tercer objetivo es la comprensibilidad y se mide en términos de la cantidad de atributos que contiene el cuerpo de la regla. Para ello se emplea la expresión $\frac{1}{Atrib_{A \to B}}$, donde $Atrib_{A \to B}$ es la cantidad de atributos en el cuerpo de la regla. De esta forma se favorecen las reglas con menor cantidad de atributos, lo que permite que sean más fáciles de leer e interpretar.

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 3, Mes Julio-Septiembre, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 18-40

La codificación de los descriptores se realiza en un cromosoma de n genes, donde n es el número de atributos de la BD. Se emplea una codificación posicional en la que el i-ésimo gen codifica el i-ésimo atributo. Cada gen está descrito por 4 elementos: “ac”: indica si el atributo participa en la descripción [-1 no presente; 0 presente], “lb” y “ub” representan los límites de los intervalos del atributo, y “pn” que indica si el intervalo se incluye o no. La Figura 1 muestra un ejemplo.

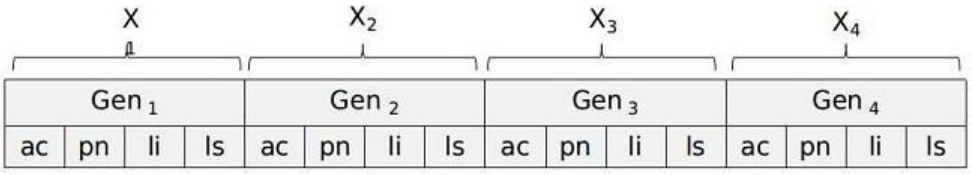


Fig.1 - Representación del cromosoma.

El atributo clase no se incluye en la codificación y se asigna después de ejecutar los operadores genéticos. Se escoge la clase que mayor frecuencia relativa tiene en los ejemplos cubiertos; de esta forma se pueden obtener reglas para todas las clases sin tener que ejecutar el algoritmo para cada clase. Otra ventaja del método de asignación es que promueve la inusualidad en los subgrupos, ya que selecciona la clase con mejor diferencia entre las probabilidades en el subgrupo y en la base de datos completa. Entonces para un subgrupo “s” su clase “C” se asigna según la ecuación 5, donde P se refiere a la probabilidad estimada en la base de datos. Esta fórmula es equivalente al cálculo de la métrica Lift (García - Borroto et al. 2017) lo que indica que se va a escoger la clase con mejor correlación positiva con respecto a la descripción del subgrupo.

$$s \rightarrow C_i | MAX \left[\frac{p(C_i|s)}{p(C_i)} \right] \quad (5)$$

Los operadores empleados son de cruce, mutación y reparación. Primero el operador de cruce genera dos hijos intercambiando aleatoriamente los genes de los padres. Después el operador de

mutación escoge con igual probabilidad un gen del cromosoma. De este gen, el azar, selecciona uno de los límites del intervalo y aumenta o disminuye su valor. Además, modifica los valores de "ac" y "pn" aleatoriamente. Por último, el operador de reparación corrige los cromosomas que tengan todos los atributos excluidos de la regla. Para ello elige al azar alguno de los atributos y cambia "ac" al estado presente. Este operador además decrementa el tamaño de los intervalos mientras que sigan cubriendo los mismos objetos.

Estudio experimental

En esta sección muestra un estudio experimenta en el cual se compara el comportamiento de DINOS con cuatro algoritmos del estado del arte de descubrimiento de subgrupos para contrastar el comportamiento del mismo. Se seleccionaron dos algoritmos exactos: APRIORI-SD (Kavsek and Lavrac, 2006) Y SD-MAP (Atzmueller and Puppe, 2006), y dos algoritmos evolutivos NMEEFSD (Carmona et al., 2010) y SDIGA (Romero et al., 2009). Para la experimentación se realizó con una validación cruzada de diez partes. Además, para los algoritmos, no exactos (DINOS, NMEEFSD, SDIGA) se realizaron tres ejecuciones en todas las bases de datos con el objetivo de descartar los efectos de la aleatoriedad. Adicionalmente para el caso de NMEEFSD y SDIGA que trabajan con particiones difusas se ejecutaron con 3 y 5 etiquetas difusas. En la Tabla 1 se muestran los parámetros de configuración de para los algoritmos estudiados. En la Tabla 2 se muestran las características de las 20 de bases de datos del repositorio UCI-ML (Dheeru and Karra Taniskidou, 2017) empleadas en el estudio, donde Var significa cantidad de variables, Disc: cantidad de variables discretas, Cont: cantidad de variables continuas, Inst: cantidad de instancias en la base de datos, Abrev: abreviatura empleada en el documento para referirse a la base de datos.

Tabla 1 - Parámetros.

Algoritmo	Parámetros
SDIGA	RulesRep = can;nLabels = 3,5; nEval = 10000; popLength = 100; crossProb = 0.6; mutProb = 0.01; minConf = 0.6; lSearch = yes; Obj1 = comp; Obj2 = fcnf; Obj3 = unus; w1 = 0.4; w2 = 0.3; w3 = 0.3

NMEEFSD	RulesRep = can; nLabels = 3,5; nEval = 10000; popLength = 50; crossProb = 0.6; mutProb = 0.1; diversity = crowding; ReInitCob = yes porcCob = 0.5; Obj1 = comp; Obj2 = unus; Obj3 = null; minCnf = 0.6; StrictDominance = yes
SD-Map	MinimumSupport = 0.1; minConf = 0.8; RulesReturn = 10
Apriori-SD	MinSupp = 0.03; minConf = 0.8; Number-of-Rules = 5; Postpruning-type = SELECT-N-RULES-PER-CLASS
DINOS	Neval = 50000; H= 13; Nobj = 3; Vec = 10; Pcross = 0.9; Pmut = 0.1; MaxSolRepl = 2 ; Fampl = 2; Dif = 5

Tabla 2 - Características de las Bases de Datos.

Nombre	Var	Disc	Cont	Cla	Inst	Abrev	Nombre	Var	Disc	Cont	Cla	Inst	Abrev
Appendicitis	7	0	7	2	106	Apen	Glass	9	0	9	6	214	Glass
Australian	14	8	6	2	690	Aust	Haberman	3	0	3	2	306	Habe
Balance	4	0	4	3	625	Blce	Heart	13	6	7	2	270	Heart
Brest Cancer	8	7	1	2	256	BrCr	Hepatitis	19	13	6	2	155	Hepa
Bridges	7	4	3	2	102	Brdg	Ionosphere	34	0	34	2	351	Ionos
Bupa	6	0	6	2	309	Bupa	Iris	4	0	4	3	150	Iris
Cleveland	13	0	13	5	303	Clev	Led	7	0	7	10	500	Led
Diabetes	8	0	8	2	768	Diab	Primary Tumor	17	17	0	21	303	PryTmr
Echo	6	1	5	2	131	Echo	Vehicle	18	0	18	4	846	Veh
German	20	13	7	2	1000	Germ	Wine	13	0	13	3	178	Wine

En la tabla 3 se muestran los valores promedios de las métricas confianza, sensibilidad e inusualidad en las particiones de control para cada base de datos. Se resaltan en negrita los mejores valores de cada métrica en las bases por base de datos. A partir de estos datos se aplica la prueba de Friedman 1XN con el Post Hoc de Holms con $\alpha = 0{,}05$, para determinar si las diferencias encontradas son significativas. Los resultados de las pruebas descritas se muestran en la tabla 4, donde se muestran los algoritmos de control para cada métrica, y en orden descendente el lugar en el *ranking* de los algoritmos para la prueba representados por la colimna (*i*). Además, se muestran los p-valores para Friedman y el Post Hoc de Holms y se señala si la hipótesis nula se rechaza o no.

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 3, Mes Julio-Septiembre, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 18-40

Tabla 3 - Valor medio de la Confianza por algoritmo en cada Base de Datos.

	CONFIANZA					SENSIBILIDAD					INUSUALIDAD				
BD	MAP	APRIORI	NMEEF	SDIGA	DINOS	MAP	APRIORI	NMEEF	SDIGA	DINOS	MAP	APRIORI	NMEEF	SDIGA	DINOS
Apen	**1.0000**	0.7924	0.9094	0.8294	0.7796	0.5655	0.3187	0.7144	0.5761	**0.7722**	0.0448	0.0428	0.0865	0.0671	**0.0873**
Aust	0.8374	**0.8844**	0.7801	0.6108	0.7020	0.5693	0.4088	**0.8235**	0.6608	0.7200	0.0423	0.0914	**0.1144**	0.0714	0.1022
Blce	0.6544	**0.8350**	0.7342	0.7575	0.7107	**0.6494**	0.0697	0.4642	0.3002	0.5349	0.0714	0.0154	0.0764	0.0434	**0.0806**
BrCr	0.7013	0.8098	0.7703	**0.8687**	0.4167	0.6028	0.1243	**0.8507**	0.3697	0.4405	0.0445	0.0194	**0.0524**	0.0190	0.0277
Brdg	**1.0000**	0.8404	0.9167	0.8107	0.6148	0.5383	0.2807	**0.7007**	0.5471	0.6823	0.0353	0.0220	**0.0434**	0.0342	0.0406
Bupa	**0.8339**	0.0750	0.6917	0.5567	0.5880	0.6039	0.0061	0.3529	0.2729	**0.6819**	**0.0705**	0.0006	0.0318	0.0199	0.0331
Clev	**0.8672**	0.8250	0.7301	0.5770	0.4795	0.5576	0.3160	**0.8123**	0.4201	0.4223	0.0405	0.0667	**0.1079**	0.0459	0.0645
Diab	0.8244	**0.9476**	0.7333	0.5115	0.6750	0.5839	0.0839	0.7514	0.4605	**0.7901**	0.0354	0.0170	0.0560	0.0390	**0.0793**
Echo	**0.9874**	0.8040	0.7303	0.6824	0.6580	0.5608	0.1402	**0.7642**	0.6780	0.7275	0.0441	0.0236	0.0406	0.0394	**0.0538**
Germ	0.6385	**0.8768**	0.7627	0.5922	0.6215	0.5933	0.2263	0.5454	**0.7440**	0.7392	**0.0412**	0.0349	0.0321	-0.0001	0.0160
Glass	**0.9587**	0.8325	0.6079	0.5294	0.6461	0.7215	0.2238	0.5222	0.3881	**0.7904**	0.0662	0.0342	0.0595	0.0269	**0.0945**
Habe	**0.8555**	0.7849	0.7755	0.7175	0.6439	0.5734	0.0573	**0.8793**	0.6129	0.7515	**0.0393**	0.0048	0.0334	0.0162	0.0392
Heart	**0.8700**	0.8569	0.7239	0.7061	0.5874	0.5694	0.3352	**0.7817**	0.6204	0.5724	0.0434	0.0727	**0.1081**	0.0899	0.0910
Hepa	0.7426	0.7616	**0.8048**	0.7557	0.6517	0.6563	0.3556	0.5930	0.4738	**0.6973**	0.0525	0.0387	0.0276	0.0299	**0.0572**
Ionos	**0.9509**	0.9339	0.8117	0.6264	0.8501	0.5838	0.3071	**0.7977**	0.5363	0.7647	0.0468	0.0680	0.1130	0.0299	**0.1159**
Iris	**0.9524**	0.8750	0.9223	0.9463	0.9452	0.6374	0.4888	0.8622	0.7606	**0.9414**	0.0611	0.1070	0.1820	0.1621	**0.2025**

Led	0.6466	**0.8997**	0.6620	0.6562	0.4680	**0.8959**	0.6479	0.8173	0.7349	0.7293	0.0577	0.0604	**0.0677**	0.0583	0.0566
PryTmr	0.6487	0.7104	**0.8929**	0.1147	0.3618	0.9043	0.4445	**0.9619**	0.6678	0.5682	0.0580	**0.0721**	0.0259	0.0048	0.0352
Veh	0.6935	**0.7263**	0.0000	0.3092	0.5025	**0.6439**	0.1451	0.0000	0.4493	0.5980	0.0560	0.0241	0.0000	0.0263	**0.0703**
Wine	**1.0000**	0.8874	0.8171	0.6817	0.8580	0.6916	0.4364	0.7608	0.3371	**0.8555**	0.0676	0.0922	0.1478	0.0572	**0.1707**

Con el objetivo de estudiar la redundancia en los resultados de los algoritmos, así como tener una perspectiva clara sobre la calidad de cada subgrupo de forma individual, se emplea el método comparativo propuesto en (Ilisástigui et al. 2019) descrito en la sección de “Antecedentes”. Se comienza cuantificar el porciento de patrones redundantes en los resultados de cada algoritmo, lo cual se grafica en la figura 2. Para determinar los patrones redundantes se emplea los mismos criterios de cobertura y estadísticos que se describen en la sección “Selección de descripciones no redundantes”.

En la figura 3 se puede observar el resultado del estudio de similitud entre DINOS y SD-MAP, APRIORI-SD, NMEEFSD, SDIGA. La cantidad de subgrupos solo encontrados por DINOS se representan en color “blanco”, los comunes en color “gris” y los de cada algoritmo en “negro” respectivamente. En las Figura 4 se puede observar las proporciones de las cantidades de patrones presentes en cada partición de las tres métricas estudiadas. En “blanco” se representa el tercio superior de los valor de las métricas, en” gris2 el central y en” negro” inferior.

Tabla 4 - Resultados del método Post Hoc de Holms.

Métricas	***Alg_Control***	**I**	**Algoritmos**	***p***	**Holm**	**Hipótesis**
CONF	MAP	5	DINOS	0.000017	0.000068	Rechazada
		4	SDIGA	0.000216	0.000647	Rechazada
		3	NMEEF	0.089131	0.178262	No Rechazada
		2	APRIORI	0.764177	0.764177	No Rechazada

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 3, Mes Julio-Septiembre, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 18-40

SENS	DINOS	5	APRIORI	0	0	Rechazada
		4	SDIGA	0.006934	0.020802	Rechazada
		3	MAP	0.133614	0.267229	No Rechazada
		2	NMEEF	1	1	No Rechazada
WRACC	DINOS	5	SDIGA	0.000027	0.000107	Rechazada
		4	APRIORI	0.000318	0.000955	Rechazada
		3	MAP	0.021448	0.042896	Rechazada
		2	NMEEF	0.36812	0.36812	Rechazada

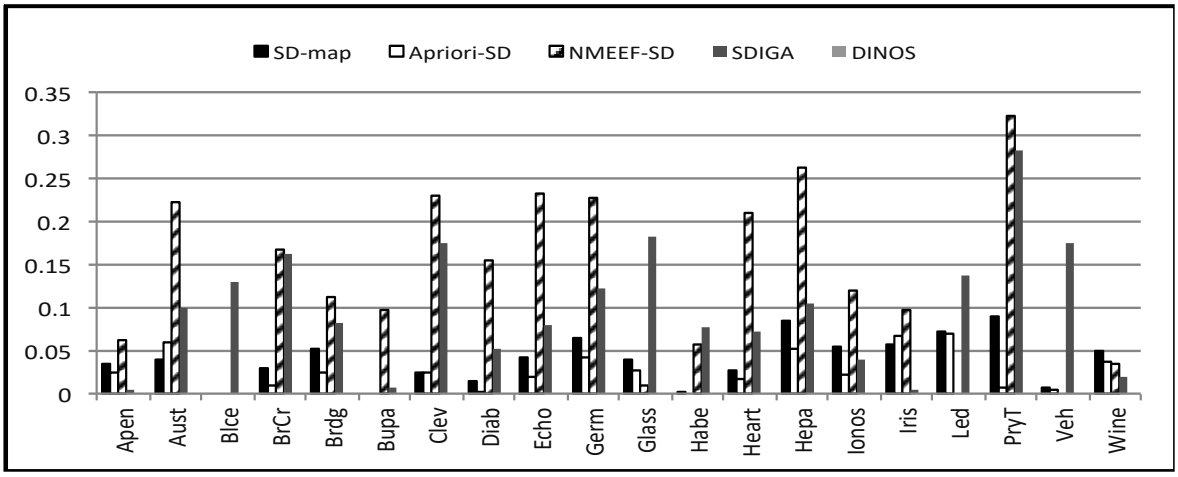
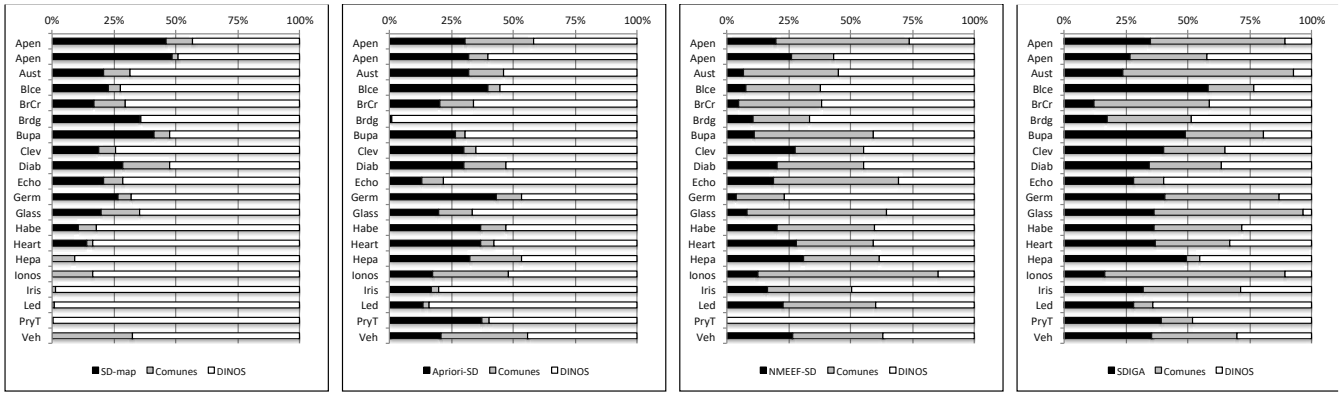


Fig. 2 - Porciento de Patrones Redundantes.

A partir de los resultados mostrados para las pruebas estadísticas en la tabla 4 se puede observar que DINOS obtiene buenos valores para las métricas de sensibilidad e inusualidad, donde es el algoritmo con mejor ranking, y en la última tiene diferencias significativas con todos los algoritmos comparados. En el caso de la confianza, DINOS obtiene los valores medios más bajos del estudio, lo que puede estar ocasionado por el empleo de WRAcc y la sensibilidad en la evaluación de las nuevas soluciones en el proceso evolutivo, ya que estas tienden a tener una correlación inversa. No obstante, los valores de WRAcc altos suponen una buena precisión relativa de los descriptores encontrados, por los que los valores de confianza no implican un bajo nivel de calidad. La figura 2

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 3, Mes Julio-Septiembre, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 18-40

muestra que DINOS no obtiene patrones redundantes mientras que en el resto de los algoritmos se detectaron redundancias en la mayoría de las bases de datos. Sólo SD-MAP y APRIORI-SD se mantiene por debajo del 10 % de patrones redundantes para todas las bases de datos. En la comparación pareada de DINOS con respecto a los algoritmos del estado del arte podemos observar:



(a) SD-MAP vs DINOS (b) APRIORI-SD vs DINOS (c) NMEEFSD vs DINOS (d) SDIGA vs DINOS

Fig. 3 - Proporción de patrones Comunes y por DINOS en comparación con algoritmos en estudio.

1. **SD-MAP:** En la figura 3(a) se observa que, en 18 de las 20 bases de datos, DINOS descubre más de la mitad de los subgrupos que fueron minados por ambos algoritmos, en 4 de ellas mina todos los subgrupos encontrados. En cuanto a la calidad de los descriptores, en la figura 4 se puede observar que para la confianza DINOS tiene descriptores en los tres intervalos de la métrica, y que al menos el 25 % están en el intervalo superior. Si se analiza en conjunto con la distribución de la sensibilidad y de la insualidad, donde DINOS tiene mayor cantidad de descriptores en los intervalos superiores mientras SD-MAP tiene muy pocos, podemos confirmar que los valores más bajos de la confianza en DINOS están influenciados por valores superiores de WRAcc y la sensibilidad. Por lo que se considera que DINOS tiene mejor desempeño que SD-MAP, pues encuentra un mayor número de subgrupos con mejor calidad.

2. **APRIORI-SD:** El análisis de los valores de las pruebas estadísticas con respecto a DINOS es similar al anterior. Sin embargo, A PRIORI-SD si logra encontrar patrones distintos a los de DINOS en todas las bases de datos, aunque DINOS sigue obteniendo más del 50 % de los subgrupos en todas las bases de datos. En el caso de las métricas de calidad se vuelve a apreciar un comportamiento similar al explicado anteriormente, donde DINOS tiene descriptores en todo el dominio de las métricas y se comporta mejor para sensibilidad e inusualidad. Por esta razón se afirma que DINOS también en este caso muestra mejores resultados.

3. **NMEEFSD:** No tiene diferencias significativas con DINOS en cuanto a la sensibilidad, de hecho es el segundo algoritmo con mejor ranking. En el caso de la inusualidad si se logran apreciar diferencias significativas que llevan a concluir que Dinos obtiene mejores resultados en esta métrica en la mayoría de las bases de datos. NMEEFSD es el algoritmo con mayor número de descriptores redundantes, en alguna base de datos sobrepasa el 20 % y 30 % de los encontrados. En cuanto a la similitud se observa que encuentran un gran número de subgrupos en común, aunque DINOS mina mayor cantidad de forma individual por lo que podemos decir que DINOS encuentra más subgrupos. La calidad de los descriptores muestra que NMEEFSD obtiene valores de confianza ligeramente superiores, sin embargo, DINOS vuelve a tener valores de inusualidad mucho mejores, sin dejar de tener elementos en todos los intervalos de las métricas. La sensibilidad tiene un comportamiento muy similar en ambos algoritmos. El modelo de conocimiento que devuelve DINOS tiene mejor usabilidad que el de NMEEFSD, pues tiene un mayor número de descriptores no redundantes y obtiene más subgrupos con mejor calidad en la mayoría de las métricas.

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 3, Mes Julio-Septiembre, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 18-40

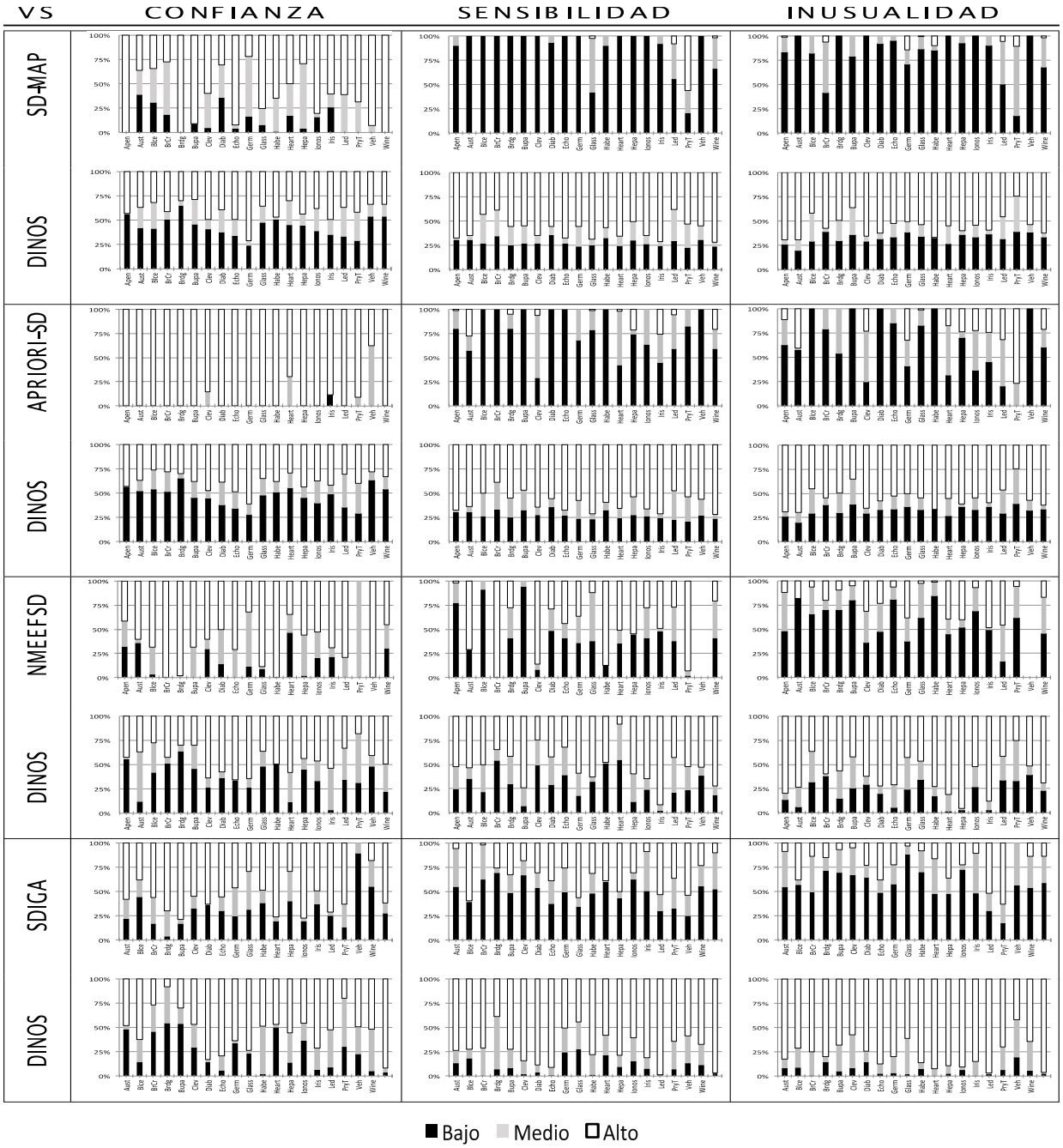


Fig. 4 – Distribución de los valores de las métricas de calidad de los conjuntos de subgrupos obtenidos por los algoritmos en cada base de datos.

4. **SDIGA:** En el análisis estadístico DINOS tiene mejores resultados de forma significativa en la inusualidad y la sensibilidad. Además, SDIGA mina un gran número de descriptores redundantes, sobrepasando el 15 % en 4 bases de datos. En la similitud es interesante como los algoritmos encuentran un gran número de patrones comunes, los cuales llegan a ser el 50 % de los encontrados en 6 bases de datos. También resulta interesante que en las bases de datos donde DINOS encuentra más descriptores de forma individual, es justo donde SDIGA encuentra pocos y viceversa. En algunos casos las cantidades de subgrupos detectados de forma individual es igual para ambos. En el caso de las métricas de calidad el comportamiento es bastante similar. Pero tanto para la sensibilidad y la inusualidad se ve una clara diferencia, pues los porcientos de descripciones en la parte superior de la métrica son mayores en DINOS. Sería interesante en algunos entornos emplear ambos algoritmos juntos para obtener un resultado diverso y de alta calidad.

Conclusiones

En este trabajo se ha presentado DINOS un algoritmo genético multiobjetivo con enfoque de descomposición para el minado de subgrupos no redundante y de alta calidad en bases de datos con atributos continuos sin discretización previa. Se optimizan la inusualidad, comprensibilidad, confianza y sensibilidad mientras se efectúa el aprendizaje evolutivo de los atributos y sus intervalos. Como forma de promover la diversidad en las soluciones, emplea una estrategia de reinicializacion de la población basada en la cobertura y una población externa para conservar las mejores soluciones encontradas hasta el momento. Además, incluye un novedoso método de filtrado de subgrupos redundantes basado en cobertura y similitud estadística de la distribución de clases en los subgrupos. Por lo que se obtiene un conjunto de subgrupos muy útil, pues son interesantes, inusuales, sin redundancias y comprensibles.

Se realiza un estudio experimental comparativo entre DINOS y cuatro algoritmos bien conocidos del estado del arte. Se emplearon para ello dos métodos que permiten estudiar la calidad desde perspectivas diferentes, así como la incidencia y efecto de la redundancia en los conjuntos de subgrupos obtenidos. El estudio demuestra que DINOS mina subgrupos de alta calidad mejorando los resultados de los algoritmos del estado del arte.

Referencias

ATZMUELLER, M. Subgroup discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2015, vol. 5, no. 1, p. 35–49.

ATZMUELLER, M AND PUPPE, F. SD-Map – A fast algorithm for exhaustive subgroup discovery. In Proceedings of the 10th European conference on Principles and Practice of Knowledge Discovery in Databases. Springer, Berling, Heidelberg 2006. p. 6–17.

BERLANG, F. ET, al. Multiobjective evolutionary induction of subgroup discovery fuzzy rules: A case study in marketing. In Proceedings 6th Industrial Conference on Data Mining. 2006. p. 337–349.

BOLEY, M. AND GROSSKREUTZ, H. Non-redundant subgroup discovery using a closure system. In Proceedings Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer ,Berling, Heidelberg, 2009. p 179–194.

CARMONA, C. J. ET AL. NMEEF-SD: nondominated multiobjective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. IEEE Transactions on Fuzzy Systems, 2010, vol. 18, no. 5, p. 958–970.

DHEERU DUA AND EFI KARRA TANISKIDOU. UCI machine learning repository, 2017. URL http://archive.ics. uci.edu/ml.

FLEISS JOSEPH L, BRUCE LEVIN, AND MYUNGHEE CHO PAIK. The odds ratio and its logarithm. In: John Wiley & Sons, Statistical methods for rates and proportions. John Wiley & Sons, 2013.

GARCÍA, S. ET al. Advanced nonparametric tests for multiple comparisons in the desing of experiments in computational intelligence and data mining: Experimental analysis of power. Information Ciences, 2010, vol. 180, no. 10, p. 2044-2064.

GARCÍA-BORROTO, M. et al. Evaluation of quality measures for contrast patterns by using unseen objects. Expert Systems with Applications, 2017, vol. 83, p. 104-113.

HERRERA, F. et al. An overview on subgroup discovery: foundations and applications. Knowledge and information systems, 2011, vol. 29, no. 3, p. 495–525.

ILISÁSTIGUI, L. B. et al. A new method to evaluate subgroup discovery algorithms. In Proceedings Iberoamerican Congress on Pattern Recognition. Springer, Cham, 2019. p. 417–426.

KAVSEK B AND N. LAVRAC. Apriori-sd: Adapting association rule learning to subgroup discovery. Applied Artificial Intelligence, 2006, vol. 20, no. 7, p. 543–583.

KLÖSgen WILLI. Explora: A multipattern and multistrategy discovery assistant. In Proceedings Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, 1996. p. 249–271.

LAVRAC, N. et al. Subgroup discovery with cn2-sd. Journal of Machine Learning Research, 2004, vol. 5, p. 153–188.

LI, HAIQUAN et al. Relative risk and odds ratio: A data mining perspective. In Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2005. p. 368–377.

LI JIUYONG et al. Discovering statistically non-redundant subgroups. Knowledge-Based Systems, 2014, vol. 67, p. 315–327.

MARTÍN, D ET AL. NICGAR: A niching genetic algorithm to mine a diverse set of interesting quantitative association rules. Information Sciences, 2016, vol. 355, p. 208–228.

MARTIN, D.et al. A new multiobjective evolutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules. IEEE Transactions on Evolutionary Computation, 2014, vol. 18, no. 1, p. 54–69.

NOVAK, P. K. et al. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. Journal of Machine Learning Research, 2009, vol. 10, no. Feb, p. 377–403.

ROMERO, C ET al. Evolutionary algorithms for subgroup discovery in e-learning: A practical application using moodle data. Expert Systems with Applications, 2009, vol. 36, no. 2, p. 1632–1644.

TAN NING - PANGet al. Selecting the right objective measure for association analysis. Information Systems, 2004, vol. 29, no. 4, p. 293–313.

TAN YAN-YAN et al. A modification to moea/d-de for multiobjective optimization problems with complicated pareto sets. Information Sciences, 2012, vol. 213, p. 14–38.

VAN LEEUWEN MATTHIJS AND ARNO KNOBBE. Diverse subgroup set discovery. Data Mining and Knowledge Discovery, 2012, vol. 25, no. 2, p. 208–242.

WROBEL STEFAN. An algorithm for multi-relational discovery of subgroups. In Proceedings European Symposium on Principles of Data Mining and Knowledge Discovery, Springer, 1997. p. 78–87.

Conflicto de interés

No existe

Contribuciones de los autores

Lisandra Bravo Ilisástigui: Contribución asociada al desarrollo e implementación del algoritmo propuesto y desarrollo del manuscrito.

Diana Martín Rodríguez: Contribución asociada con el marco teórico de la investigación y asesoría en temas de optimización y algoritmos descriptivos de minería de datos.

Miltón García Borroto: Contribución asociada con marco teórico de la investigación y asesor en temas de algoritmos supervisados de minería de datos.