CAPÍTULO III

Patrones Similares Frecuentes, un nuevo enfoque para describir los conjuntos de datos

Gretel Bernal Baró, Rosa María Valdovinos Rosas,
Ansel Yoan Rodríguez González y J. R. Marcial-Romero

Introducción

El minado de Patrones Frecuentes (PF) constituye una etapa fundamental en el proceso de búsqueda de información de muchas tareas de minería de datos. En este sentido, los PF son un conjunto de valores, atributos o subdescripciones de instancias que aparecen en los datos con una frecuencia mayor que un umbral de validación especificado por el usuario. Si el valor de la frecuencia es mayor o igual al umbral de mínima frecuencia establecido, se considera que el patrón es frecuente.

La mayoría de los algoritmos existentes en la literatura, utilizados en el minado de PF consideran la igualdad entre los valores de los atributos para contar su frecuencia [1]. Sin embargo, en las ciencias blandas, dos instancias se pueden considerar similares, aunque no sean idénticas. Por ejemplo, dado los intervalos de edad (0-3) y (4-7) años, un niño de 3 años 11 meses tiene intereses muy similares a los de un niño de 4 años. En este ejemplo podemos ver que a pesar de que las edades son equivalentes, al momento de discretizar los datos se omite la semántica de los datos cambiando su naturaleza. En problemas reales como la obtención de perfiles de usuarios, *modus operandi* de la población, síndromes comunes, tendencias de compras y ventas o factores de riesgo, algunos Patrones Frecuentes podrían perderse y, por tanto, se estaría obviando conocimiento relevante [2].

Como alternativa de solución a esta problemática, el concepto de similitud entre subdescripciones de instancias es incorporado para contar cuántas veces aparece una subdescripción en un conjunto de

datos. Además, cuando se utilizan funciones de similitud diferentes a las de la igualdad es descubierto el universo de todos los Patrones Frecuentes existentes en el conjunto de datos [1]. Los nuevos patrones descubiertos son denominados Patrones Similares Frecuentes (PSF). Un PSF es una combinación de valores de atributos de la instancia en el conjunto de datos, de manera que el número de veces que aparece ese patrón, más la suma de las ocurrencias de las subdescripciones similares (al patrón analizado) no es menor que un umbral de frecuencia especificado por el usuario [3].

Al respecto, Danger y Shulcloper [4] fueron los precursores en incorporar funciones de semejanza en el cálculo de la frecuencia, de tal manera que dos subdescripciones que no tienen valores idénticos se consideran semejantes, dando origen al minado de Patrones Similares Frecuentes. En este capítulo se realiza un estudio en torno a las ventajas del uso de los PSF con respecto a los PF del enfoque tradicional. Por tal motivo, la comparación experimental incluye el análisis de la calidad de los patrones minados por los algoritmos más citados para cada enfoque: el algoritmo STreeDC-Miner es utilizado para minar los Patrones Similares Frecuentes con respecto a los Patrones Frecuentes del enfoque tradicional minados por el algoritmo Apriori [5].

Trabajos relacionados

En la literatura, se pueden encontrar varios algoritmos diseñados para minar los PSF. Estos algoritmos tienen los siguientes aspectos en común:

Para cada atributo del conjunto de datos, es necesario definir un criterio de comparación que indica si el par de valores comparados debe ser considerado similar o no por el proceso de minería. A continuación, se muestra un ejemplo de este tipo de funciones:

$$C_r(x,y)=\begin{cases} Si\ |x-y|\leq \varepsilon, y\ 1 \\ sino\qquad\qquad 0 \end{cases} \tag{1}$$

En la Ec. (1), C_r es el criterio de comparación correspondiente al atributo r y dos valores en r se consideran similares si el valor absoluto de la diferencia es menor o igual a un umbral ε especificado.

- Para cada problema particular, a diferencia de los algoritmos tradicionales de minería de Patrones Frecuentes, se define una función de similitud entre subdescripciones. Esta función de similitud es la que permite definir si dos subconjuntos de valores de atributos deben ser considerados similares o no por el proceso de minería. A continuación, se muestra un ejemplo de este tipo de funciones:

$$f_s(O,O') = \begin{cases} Si\ \forall r \in S,\ C_r(O[r],O'[r]) = 1, & y\ 1 \\ sino & 0 \end{cases} \quad (2)$$

En la Ec. (2), f_s es una función de similitud para comparar dos descripciones de objetos con respecto a un conjunto de características S. Dadas dos subdescripciones $I_s(O)$ e $I_s(O')$, con $O, O' \in \Omega$, siendo Ω el conjunto de datos, $f_s(O, O') = 1$ significa que la instancia O es similar a la instancia O' respecto al subconjunto de atributos S y $f_s(O, O') = 0$ significa que O no es similar a O' respecto a S.

- La mayoría de los algoritmos de minería de PSF trabajan siguiendo una estrategia de búsqueda en profundidad, diseñando una estructura en forma de árbol, llamada STree [1], en la que cada rama en el árbol, desde la hoja hasta la raíz representa la subdescripción de una instancia. Las subdescripciones iguales se agrupan en la misma rama del árbol. Además, se almacenan las frecuencias de dichas subdescripciones y los enlaces a subdescripciones similares.

Por otro lado, los algoritmos de minado de PSF se clasifican según los valores devueltos por las funciones de similitud permitidas y su monotonía. Las funciones de similitud pueden ser booleanas o no booleanas. Una función de similitud booleana (Ec. (2)) devuelve los

valores 0 o 1, es decir, las subdescripciones comparadas serán consideradas similares o no por el proceso de minería (STreeDC-Miner [1], CFSP-Miner [3], RP-Miner [6]). Una función de similitud no booleana es aquella cuyos valores están dentro del intervalo [0,1], es decir, las subdescripciones serán similares en mayor o menor medida dependiendo del valor devuelto por la función de similitud. Los valores cercanos a 1 indican una similitud más notable entre las subdescripciones (STree*DC-Miner [7], STree*NDCMiner [7], RP*-Miner [7]).

La monotonía de las funciones de similitud puede ser no creciente o creciente. Por un lado, una función de similitud es monótona no creciente si y sólo si para cualquier par de objetos, la similitud respecto a un conjunto de características es mayor o igual que la similitud respecto a cualquier superconjunto de características. Por otro lado, la monotonía no creciente de la función de similitud es una propiedad relevante porque implica que todas las súper-descripciones de un patrón similar no frecuente tampoco son frecuentes. Esta propiedad, conocida como f_s-Clausura Descendente, permite podar el espacio de búsqueda de los PSF [7].

De los algoritmos existentes en el estado del arte para el minado de PSF, el algoritmo STreeDC-Miner [1], presenta ventajas operacionales con respecto a sus homólogos, debido a que requiere menor tiempo computacional para minar los Patrones Similares Frecuentes existentes en el conjunto de datos. En STreeDC-Miner se introduce la estructura de árbol *STree*. Para cada conjunto de características *A*, es generada una estructura de árbol $STree_A$, cada hoja de la estructura representa una subdescripción respecto al conjunto de características *A* y almacena todas sus repeticiones, así como las similitudes con otras. Las ramas del $STree_A$ contienen los prefijos comunes de las almacenadas. STreeDC-Miner establece una definición explícita de un orden total sobre el conjunto de características. A partir de cada conjunto de una sola característica *A*, y siguiendo una estrategia de búsqueda en profundidad, un procedimiento recursivo añade a *A* en cada llamada, una nueva característica mayor que las características en *A*.

Además, en cada llamada recursiva, se construye una estructura de árbol $STree_A$, se calcula la frecuencia de las subdescripciones en $STree_A$ y se obtienen los PSF. Si el conjunto actual de característica

A sólo contiene una, el árbol $STree_A$ se realiza a partir del conjunto de datos. En caso contrario, se hace a partir de la estructura de árbol obtenida en la llamada recursiva anterior. El caso base del procedimiento recursivo se produce cuando no hay Patrones Similares Frecuentes para el conjunto de características *A* o no hay ninguna otra característica que añadir. En el proceso de construcción del *STree*, la similitud entre dos subdescripciones sólo se calcula si éstas, en la estructura de árbol resultante de la llamada recursiva anterior, son similares en consecuencia, así se reduce el número de evaluaciones de la función de similitud y el esfuerzo computacional para calcular la frecuencia de cada una.

Patrones Similares Frecuentes vs Patrones Frecuentes

Un Patrón Similar Frecuente (PSF) es una combinación de valores de atributos de una instancia tal, que la suma de la frecuencia de sus Patrones Similares no es menor que un umbral de frecuencia especificado por el usuario. Por ejemplo, en los estudios sociológicos se puede considerar que dos personas son similares en términos de su edad si pertenecen a la misma generación, lo que equivale a considerar que dos edades son similares si el valor absoluto de su diferencia es como máximo de 5 años. También se puede definir que dos personas son similares en términos de su salario si el valor absoluto de la diferencia, entre los salarios, es máximo de 5000 pesos.

Dado el conjunto de datos mostrado en la Tabla I para un umbral de frecuencia mínima igual a 0.7, es decir, una subdescripción es frecuente si aparece al menos en 2 de las instancias del conjunto de datos. Si se considera la igualdad estricta al contar la frecuencia entre los valores de los atributos, solo (Estado Civil = Soltero) sería un Patrón Frecuente. Sin embargo, si se consideran los criterios de semejanza anteriormente mencionados entre los atributos Edad y Salario, alguno de los PSF minados son: *(Edad = 23)*, *(Edad = 25, Salario = 7500)*, etc.

Tabla I. Conjunto de datos mezclado [7].

Objeto	Edad	Salario	Estado Civil
O_1	23	5000	Soltero
O_2	25	7500	Soltero
O_3	29	7300	Soltero

Como se pudo observar en el ejemplo, cuando es utilizada la igualdad estricta para contar la frecuencia con que aparecen las subdescripciones o patrones en el conjunto de datos, podrían no ser considerados algunos Patrones Frecuentes, lo que conlleva a la pérdida de conocimiento.

Metodología

En esta sección se describe la estrategia metodológica propuesta. Primeramente, son seleccionados los conjuntos de datos que van a ser usados para evaluar la cantidad de Patrones Frecuentes minados por ambos enfoques. Luego, para el caso de los algoritmos de minado de PSF es necesario definir el criterio de comparación asociado a cada atributo y la función de semejanza a utilizar. Posteriormente, se establecen los umbrales de frecuencia (parámetro común para ambos algoritmos) y se lleva a cabo el proceso de minado haciendo uso de los algoritmos STreeDC-Miner [1] y Apriori [5] para minar los PSF y los PF respectivamente. Por último, es evaluada la calidad de los patrones PSF y los PF en correspondencia con la precisión obtenida por un clasificador, que utiliza los patrones minados, para clasificar nuevas instancias.

Conjuntos de datos

Para realizar la comparación entre los dos enfoques de minado de Patrones Frecuentes se utilizaron 19 conjuntos de datos del repositorio

de la Universidad de California *(https://archive.ics.uci.edu/ml/index.php)*. Estos conjuntos de datos fueron seleccionados debido a que presentan una gran diversidad entre los valores de atributos presentes. En la Tabla II se muestra la cantidad de instancias, la cantidad de atributos numéricos y no numéricos existentes. Los conjuntos de datos fueron ordenados dependiendo de la cantidad de instancias presentes.

Tabla II. Conjunto de datos de prueba[3].

Nombre	Objetos	Numéricos	No Numéricos
Glass Identifition	146	1	9
Iris	150	1	4
Teaching Assistant Evaluation	151	3	3
Wine	178	1	13
Heart Disease	270	1	13
Liver Disorders	345	1	6
Auto MPG	392	3	5
Metadata	528	2	17
Balance Scale	576	1	4
Indian Liver Patient	579	2	9
Breast Cancer Wisconsin	693	1	9
Credit Approval	690	9	7
Pima Indians Diabetes	768	1	8
Vehicle Silhouettes	846	1	18
Auto	1000	4	36
Contraceptive Method Choice	1473	2	8
Car	1728	2	5
Abalone	4177	2	7
Census	32561	6	9

Elaboración basada en https://archive.ics.uci.edu/ml/index.php

[3] *https://archive.ics.uci.edu/ml/index.php.*

Criterios de comparación y función de semejanza

Como se mencionó anteriormente, los algoritmos de minado de PSF requieren de la especificación de un criterio de comparación, para cada atributo, que indica si el par de valores debe considerarse semejante o no. Además, es necesario especificar la función de similitud o semejanza, la cual permite definir si dos patrones o subdescripciones van a ser considerados similares o no por el proceso de minería. El algoritmo STreeDC-Miner es diseñado para el uso de funciones de similitud booleana monótonas no creciente. Por tanto, la función de semejanza utilizada fue la definida en la Ec. (2) y los criterios de comparación utilizados fueron los siguientes:

$$C_r(x,y)=\begin{cases} Si & x-y=0, y\ 1 \\ sino & 0 \end{cases} \tag{3}$$

$$C_r(x,y)=\begin{cases} Si & \frac{|x-y|}{maxR-minR} \leq \varepsilon, y\ 1 \\ sino & 0 \end{cases} \tag{4}$$

La Ec. (3) es utilizada para todos los atributos no numéricos y nos indica que dos valores de un atributo no numérico se van a considerar similares si y sólo si son iguales. Para los atributos numéricos es usada la Ec. (4), en dicha ecuación *maxR* y *minR* representan el valor máximo y mínimo respectivamente del atributo *r*, el valor definido para el umbral ε fue de 0.05.

Proceso de minado de Patrones Frecuentes

En esta etapa se llevó a cabo el proceso de búsqueda de PF y PSF. Para ello, se definieron los umbrales de frecuencia utilizados por STree-DC-Miner [1] y Apriori [5] con el objetivo de evaluar la eficiencia de

los algoritmos de minado de PSF con respecto a los algoritmos del enfoque tradicional. Se considera más eficiente al enfoque que encuentre mayor cantidad de patrones. Los valores de frecuencia establecidos fueron 0.1 y 0.2, por ser los umbrales más utilizados en el estado del arte, es decir que las subdescripciones deben aparecer en el 10% y 20% de las instancias del conjunto de datos para ser consideradas frecuentes.

Clasificación

En la experimentación se utilizaron tres conjuntos de datos: Car, Contractive, Census. De cada conjunto de datos se obtuvieron los Patrones Frecuentes de ambos enfoques para los umbrales de frecuencia comprendidos en el intervalo de 0.01 a 0.06. Para cada conjunto de datos y umbral de mínima frecuencia, se utilizó validación cruzada con 10 repeticiones, en las que el 80% de las instancias fueron utilizadas para el entrenamiento y el 20% restante para clasificar.

Para revisar la calidad de PSF y PF obtenidos con el proceso de minería se seleccionaron aleatoriamente tres de los conjuntos de datos utilizados durante la experimentación (Car, Contractive, Census). De cada uno de ellos se calcularon los patrones frecuentes de ambos enfoques para los umbrales de frecuencia comprendidos en el intervalo de 0.01 a 0.06. Para cada conjunto de datos y umbral de mínima frecuencia se utilizó validación cruzada con 10 repeticiones, en las que el 80% de las instancias fueron utilizadas para el entrenamiento y el 20% restante para clasificar.

Resultados y discusión

En esta sección se muestra el desempeño de los algoritmos de minado de PSF con respecto a los algoritmos de minado de PF del enfoque tradicional, tanto en la etapa de minado, como en la de clasificación.

Minado de PSF y PF

La eficiencia de los algoritmos de minado se muestra en la Tabla III, la cual muestra la cantidad de patrones obtenidos por cada algoritmo una vez que el proceso de minado termina.

Tabla III. Desempeño de los algoritmos de minado de PSF con respecto a los algoritmos de minado de PF tradicionales teniendo en cuenta el número de PF minados.

Nombre	Umbral	STreeDC-Miner	A priori
Glass Identifition	0.1 0.2	20326 5297	11 11
Iris	0.1 0.2	207 41	6 3
Teaching Assistant Evaluation	0.1 0.2	221 35	30 15
Wine	0.1 0.2	1559 285	3 3
Heart Disease	0.1 0.2	8434 882	1010 247
Liver Disorders	0.1 0.2	2764 703	12 3
Auto MPG	0.1 0.2	1622 250	12 6
Metadata	0.1 0.2	307308 89130	21 15
Balance Scale	0.1 0.2	42 10	42 10
Indian Liver Patient	0.1 0.2	103625 41623	2 22
Breast Cancer Wisconsin	0.1 0.2	1311 277	1311 277
Credit Approval	0.1 0.2	2678886 852004	2340 562

Pima Indians Diabetes	0.1 0.2	5181 1094	16 6
Vehicle Silhouettes	0.1 0.2	34244 3756	16 5
AutoUniv au6	0.1 0.2	1622 250	12 6
Contraceptive Method Choice	0.1 0.2	475 121	423 121
Car	0.1 0.2	86 31	86 31
Abalone	0.1 0.2	187624 8724	7 3

Como puede ser visto en la Tabla III, la cantidad de Patrones Frecuentes minados, para los diferentes valores de soporte probados, en el 76% de los casos es mayor cuando son utilizadas funciones de semejanzas diferentes de la igualdad en el cálculo de la frecuencia.

También se puede apreciar que cuando se utiliza la igualdad estricta como criterio de comparación se pierden en algunos conjuntos de datos más del 90% de los PSF, información que pudiera ser de utilidad para el problema de estudio a resolver. Esta pérdida de conocimiento por los algoritmos de minado de PF del enfoque tradicional se debe a que valores de un mismo atributo con semánticas casi idénticas no son tomados en cuenta al realizar el cálculo de la frecuencia y al contar sólo los valores estrictamente iguales, los valores de frecuencia obtenidos no sobrepasan los umbrales de frecuencias establecidos.

Como se mencionó con anterioridad, la calidad del conjunto de patrones minados por ambos enfoques es medida teniendo en cuenta la precisión obtenida por un clasificador que en su funcionamiento utiliza los patrones minados para clasificar nuevas instancias. La Tabla IV muestra la precisión obtenida por el clasificador al utilizar los PSF y los PF minados en los conjuntos de datos Car, Contractive y Census para los diferentes valores de frecuencia testeados.

Tabla IV. Precisión obtenida por el clasificador al utilizar los PSF y los PF del enfoque tradicional.

Frecuencia	CD: Car		CD: Contractve		CD: Car	
	PSF	PF	PSF	PF	PSF	PF
0.01	**80.49**	80.44	**45.69**	41.75	**74.66**	73.06
0.04	**76.29**	75.60	40.20	**40.65**	**76.00**	72.39
0.08	69.75	**70.28**	37.46	**37.66**	**76.00**	71.26
0.12	**65.50**	65.43	**36.17**	33.76	**72.66**	70.80
0.16	**56.49**	55.65	**35.64**	29.65	**73.33**	70.93
Promedio	**69.70**	69.48	**39.03**	36.69	**74.53**	71.68

Como se puede apreciar, para la mayoría de los umbrales de mínima frecuencia, basado en la precisión obtenida por el clasificador y teniendo en cuenta que definimos que la calidad del conjunto de patrones es proporcional a la precisión obtenida por el clasificador, la calidad de los PSF siempre es mayor y en el peor de los casos igual a la calidad de los PF minados. Esto evidencia la utilidad de los patrones que se pierden al emplear el enfoque tradicional de minado de Patrones Frecuentes.

Conclusiones

La minería de PSF está atrayendo fuertemente la atención como una solución alternativa en el desarrollo de estrategias descriptivas. Los experimentos realizados en este trabajo validan que cuando se utilizan funciones de semejanza diferentes a las de la igualdad se puede

obtener un cúmulo de conocimiento superior a cuando se utiliza la igualdad entre los valores de los atributos como función de semejanza. Además, cuando los PSF son utilizados en tareas como la clasificación, el clasificador obtiene una mayor precisión que cuando los PF del enfoque tradicional son usados.

Este resultado fue posible gracias a que al utilizar funciones de semejanza y no la igualdad estricta entre subdescripciones como condición para contar las ocurrencias de una de éstas en el conjunto de datos es posible descubrir nuevo conocimiento. Por tal motivo, cuando el enfoque de minado de Patrones Similares Frecuentes es usado, es descubierto todo el universo de Patrones Frecuentes existente y como consecuencia se tiene un mayor conocimiento del conjunto de datos.

En trabajos futuros se propone reducir el número de subdescripciones similares frecuentes minadas. Existen muchas que son similares entre sí, según la función de similitud definida por el usuario. En consecuencia, se presentan al usuario como casos diferentes. Por lo tanto, es necesario minar un subconjunto de subdescripciones similares frecuentes que describa al conjunto de PSF existentes. Para ello se podrían aplicar algoritmos de optimización para minar un subconjunto de soluciones óptimas, en correspondencia con la función de aptitud definida. Además de diseñar nuevas funciones para identificar cuando varias subdescripciones están brindando el mismo conocimiento del conjunto de datos.

Referencias

[1] A. Y. Rodríguez González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, y J. Ruiz Shulcloper, “Mining frequent patterns and association rules using similarities”, *Expert Syst Appl*, vol. 40, núm. 17, pp. 6823–6836, 2013, https://www.academia.edu/56422936/Mining_frequent_patterns_and_association_rules_using_similarities.

[2] A. Y. Rodríguez-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, y J. Ruiz-Shulcloper, “Using non boolean similarity functions for frequent similar pattern mining”, *Canadian Conference on Artificial*

Intelligence, vol. 6085 LNAI, pp. 374–378, 2010, doi: 10.1007/978-3-642-13059-5_50/COVER.

[3] A. Y. Rodríguez-González, F. Lezama, C. A. Iglesias-Alvarez, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, y E. M. de Cote, "Closed frequent similar pattern mining: Reducing the number of frequent similar patterns without information loss", *Expert Systems with Applications*, vol. 96, pp. 271–283, 2018, doi: 10.1016/J.ESWA.2017.12.018.

[4] R. Danger, J. Ruíz-Shulcloper, y R. Berlanga, "Objectminer: A new approach for mining complex objects", *Sixth International Conference on Enterprise Information Systems* (ICEIS), Porto, Portugal, pp. 42–47, 2004.

[5] A. M. Mohammed y A. Bassam, "An Improved Apriori Algorithm for Association Rules", *International Journal on Natural Language Computing*, vol. 3, núm. 1, pp. 21–29, 2014, doi: 10.5121/IJNLC.2014.3103.

[6] A. Y. Rodríguez-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, y J. Ruiz-Shulcloper, "RP-Miner: a relaxed prune algorithm for frequent similar pattern mining", *Knowledge and Information Systems*, vol. 27, núm. 3, pp. 451–471, 2011, doi: 10.1007/s10115-010-0309-9.

[7] A. Y. Rodriguez-Gonzalez, J. F. Martinez-Trinidad, J. A. Carrasco-Ochoa, J. Ruiz-Shulcloper, y M. Alvarado-Mentado, "Frequent similar pattern mining using non-Boolean similarity functions", *Journal of Intelligent y Fuzzy Systems*, vol. 36, núm. 5, pp. 4931–4944, 2019, doi: 10.3233/JIFS-179040.

[8] A. Y. Rodríguez-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, y J. Ruiz-Shulcloper, "Mining frequent similar patterns on Mixed Data", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5197 LNCS, pp. 136–144, 2008, doi: 10.1007/978-3-540-85920-8_17/COVER.