Análisis de Diferentes Tipos de Reglas en Sistemas Difusos Evolutivos para Minería de Patrones Emergentes

A. M. García-Vico[1], C. J. Carmona[2,3], M. J. del Jesus[1]

[1] Departamento de Informática. Universidad de Jaén
23071, Jaén, Spain
{agvico, mjjesus}@ujaen.es

[2] Área de Lenguajes y Sistemas del Departamento de Ingeniería Civil. Universidad de Burgos
09006, Burgos, Spain
cjcarmona@ubu.es

[3] Leicester School of Pharmacy. De Montfort University
LE1 9BH, Leicester, Reino Unido.

Abstract

La minería de patrones emergentes es una técnica agrupada dentro de los sistemas descriptivos para aprendizaje supervisado cuyo objetivo es la descripción de tendencias emergentes en el tiempo y la caracterización de diferentes clases o grupos de variables. Entre las diferentes metodologías utilizadas para resolver este problema, los sistemas difusos evolutivos han demostrado ser un enfoque muy prometedor para la tarea. No obstante, la representación del conocimiento en estos algoritmos puede ser modificada con el fin de obtener mejores resultados descriptivos. En este trabajo se presenta una comparativa de las diferentes representaciones del conocimiento más utilizadas a lo largo de la literatura, con el objetivo de determinar aquella que obtiene los mejores resultados descriptivos. Los resultados del estudio experimental muestran el poder descriptivo de estas representaciones, resaltando especialmente la capacidad descriptiva de las reglas en forma normal disyuntiva.

1 Introducción

La minería de patrones emergentes (EPM) [7, 8] es una tarea de la minería de datos encuadrada dentro del marco denominado "descubrimiento de reglas descriptivas mediante aprendizaje supervisado" (SDRD, por sus siglas en inglés) [22]. El objetivo de esta tarea es la búsqueda de patrones discriminativos cuyo soporte se incrementa de manera significativa de una clase, o conjunto de datos, a otro. Los fines para los que esta tarea fue diseñada son la caracterización de diferentes clases o grupos de variables y la detección de tendencias emergentes en datos marcados temporalmente.

A lo largo de la literatura, estos patrones han sido ampliamente utilizados como clasificadores debido a sus capacidades discriminatorias. No obstante, este tipo de patrones puede describir las relaciones existentes en los datos de manera interpretable. Por lo tanto, EPM está a medio camino entre la minería de datos descriptiva y predictiva, ya que describe relaciones existentes en los datos mediante una variable objetivo, típicamente utilizada en clasificación. Debido a estas características, la tarea ha tenido éxito en campos como la química [24, 30], bioinformática [13, 29, 31] o medicina [3, 33], entre otros [25].

Dentro del marco SDRD, la interpretabilidad de los resultados es un factor clave. En este aspecto, los algoritmos evolutivos, y en concreto, los sistemas difusos evolutivos [18] son una propuesta prometedora para la extracción de conocimiento preciso e interpretable. Estos métodos hacen uso de la lógica difusa [34–36] ampliándose a través de un proceso de aprendizaje basado en un algoritmo evolutivo [9]. Por un lado, la lógica difusa permite un mejor tratamiento de la incertidumbre y mejora la interpretación de los resultados [20]. Por otro lado, los algoritmos evolutivos permiten una búsqueda eficiente en un espacio complejo como el de los patrones emergentes y permite flexibilizar en gran medida el proceso de aprendizaje. Por estas razones, este tipo de sistema ha sido ampliamente usado en la literatura.

Habitualmente, en EPM se ha utilizado una representación canónica del conocimiento, basada en conjunciones de pares atributo-valor. Sin embargo, existen otro tipo de representación basada en forma normal disyuntiva (DNF) que es interesante de cara a la obtención de conocimiento de calidad y fácilmente interpretable. En este trabajo se presenta un estudio de las diferentes características descriptivas que

poseen las reglas de tipo canónico frente a las reglas tipo DNF en los sistemas evolutivos difusos para EPM. Esto permitirá el desarrollo de nuevos métodos evolutivos centrados únicamente en la representación del conocimiento cuyos resultados sean más descriptivos. Para ello, este trabajo se estructura de la siguiente manera: En la sección 2 se introduce brevemente el concepto de EPM, así como sus principales características. A continuación, en la sección 3, se presentan las principales medidas de calidad en EPM desde el punto de vista descriptivo. Después, la sección 4 presenta el estudio experimental llevado a cabo. Finalmente, la sección 5 presenta las conclusiones de este trabajo.

2 Minería de patrones emergentes

En esta sección se detallan brevemente las principales características de la EPM. En primer lugar, se presenta la definición y los objetivos principales de la tarea. A continuación se presentan los tipos de patrones emergentes más destacados para este trabajo, así como sus posibles representaciones. Finalmente, se muestran las características de los sistemas difusos evolutivos existentes en EPM.

2.1 Definición

La EPM fue definida por Dong y Li [7, 8] como:

> "Sea un patrón X cualquiera, y sea $\rho > 1$ un valor de umbral, X se denominará como emergente si y solo si su índice de crecimiento entre dos conjuntos de datos (D_1 y D_2) es mayor que ρ."

Siguiendo con la definición, el índice de crecimiento (GR) del patrón X de D_1 a D_2 se define como:

$$GR(X) = \begin{cases} 0, & \text{Si } Sop_1(X) = Sop_2(X) = 0, \\ \infty, & \text{Si } Sop_1(X) = 0 \wedge Sop_2(X) \neq 0, \\ \frac{Sop_2(x)}{Sop_1(x)}, & \text{en otro caso} \end{cases} \quad (1)$$

donde $Sop_i(X)$ es el soporte del patrón X en el conjunto de datos i.

Según esta definición, la EPM es capaz de abordar los siguiente objetivos:

- La detección de diferencias características entre clases.
- La detección de tendencias emergentes entre conjuntos de datos marcados temporalmente.
- La detección de diferencias entre múltiples variables.

Habitualmente la representación de estos patrones es llevada a cabo mediante reglas de la forma:

$$R : Cond \rightarrow Clase$$

donde $Cond$ es un conjunto de conjunciones de pares atributo-valor, denominado representación canónica, o bien un conjunto de pares atributo-valor en forma normal disyuntiva para patrones disyuntivos.

2.2 Tipos de patrones emergentes

Debido a la definición anterior, el problema de la extracción de todos los posibles patrones emergentes ha sido clasificado como un problema NP-Duro [32]. Esto se debe a que patrones más específicos, es decir, patrones con más variables, pueden obtener un GR mayor que aquellos patrones más generales, es decir, con menos variables. Por lo tanto, no es posible la aplicación de estrategias como Apriori [1] para la extracción de todos los patrones emergentes en tiempo polinomial. Es por esto que, a lo largo de la literatura, se han propuesto diferentes tipos de patrones emergentes que permiten la reducción del espacio de búsqueda, obteniendo únicamente aquellos patrones que son interesantes para el experto.

De entre los diferentes tipos de patrones emergentes existentes, los relevantes para este trabajo son:

- *Fuzzy emerging patterns* (FEPs). Son patrones emergentes que emplean lógica difusa para representar variables de tipo numérico. Este tipo de patrones mejoran la interpretabilidad de los resultados [20] y tiene una mayor flexibilidad ya que se cubren las instancias con un cierto de grado pertenencia [15].
- *Disjunctive Emerging Patterns* (DEPs). Este tipo de patrón se representa mediante forma normal disyuntiva, permitiendo la introducción de conectores disyuntivos entre items que pertenecen a la misma variable [26]. Estos items pueden ser representados mediante lógica difusa si así se requiere.

Los diferentes tipos de patrones emergentes existentes se detallan en [16].

2.3 Sistemas difusos evolutivos en EPM

Para la correcta extracción de los diferentes tipos de patrones que se han propuesto a lo largo de la literatura, se han creado diferentes estrategias algorítmicas para la extracción eficiente de los mismos. Estas estrategias se pueden clasificar en algoritmos basados en límites, algoritmos basados en representación mediante árboles, algoritmos basados en árboles de decisión y sistemas evolutivos difusos [16]. Esta última categoría está teniendo actualmente especial interés, pues sus resultados descriptivos superan al resto de paradigmas con una precisión similar. Los sistemas difusos evolutivos son una hibridación de los sistemas difusos aumentados mediante un proceso de aprendizaje basado en un algoritmo evolutivo [18]. Los algoritmos evolutivos [9] son métodos basados en la evolución natural cuyo objetivo es resolver problemas de optimización en espacios de búsqueda complejos. Estos algoritmos tienen una buena relación entre la calidad de los resultados obtenidos y el tiempo transcurrido para la obtención de los mismos. También permiten flexibilizar en gran medida el proceso de búsqueda mediante la modificación de los operadores genéticos así como la representación de los resultados. También es muy importante destacar que son capaces de optimizar varios objetivos simultáneamente gracias a enfoques multiobjetivo [6].

Por otro lado, los sistemas difusos están basados en la lógica difusa [34–36], los cuales permiten tratar la imprecisión así como representar las variables de tipo numérico de una forma más interpretable [20]. Esto es posible gracias al uso de etiquetas lingüísticas (LLs) donde un conjunto difuso representa una única etiqueta de la variable. Estas pueden ser definidas por los expertos o mediante particiones uniformes en caso de que no se disponga de conocimiento experto.

La EPM puede ser vista como un proceso de búsqueda en un amplio y complejo espacio de búsqueda. Por lo tanto, la aplicación de los sistemas difusos evolutivos son un enfoque interesante para solucionar el problema de la extracción de patrones emergentes de manera eficiente.

Dentro de esta reciente categoría se encuentra únicamente el algoritmo EvAEP [17]. Este método está basado en un algoritmo evolutivo mono-objetivo, que emplea un esquema de codificación "Cromosoma = Regla" donde un individuo de la población representa una regla potencial. Asimismo, el consecuente de la regla no se representa, por lo que es necesaria la ejecución del método tantas veces como clases tenga el problema. El algoritmo permitía únicamente la codificación de los resultados para la obtención de reglas de tipo canónico, pues ha sido la representación que más se ha usado en la literatura. Esta se representa como en la Figura 1.

Genotipo

x_1	x_2	x_3	x_4
2	0	3	0

$\Downarrow$

Fenotipo

$$SI\ (x_1 = LL_1^2) \wedge (x_3 = LL_3^1)\ ENTONCES\ (x_{Obj} = Clase)$$

Figure 1: Representación de una regla canónica en EvAEP.

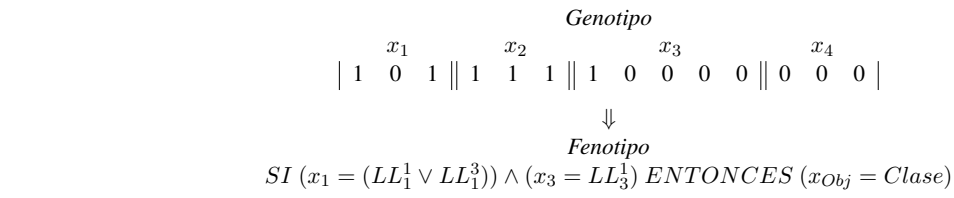

Figure 2: Representación de una regla DNF en EvAEP.

Para reglas canónicas, la representación utilizada es entera, con tamaño igual al número de variables del problema, donde cada valor indica o bien la LL o el valor discreto de la variable, donde el valor cero indica la ausencia de la variable en el resultado. En la Figura 2 se presenta la nueva codificación propuesta para representar reglas DNF que ha sido añadida al algoritmo EvAEP. La representación DNF también puede permitir la obtención de resultados interesantes desde el punto de vista descriptivo. La posibilidad de incluir diferentes valores mediante conectores disyuntivos puede permitir obtener reglas más compactas, generales e interpretables para el experto, por lo que su estudio es interesante. Para poder representar este tipo de regla, se ha utilizado una codificación binaria de tamaño igual al número de posibles valores para todas las variables. Con este tipo de representación, el valor uno indica la presencia de un valor de la variable en la regla y cero su ausencia. Es importante destacar que, en esta representación, una variable no participará totalmente si todos sus valores se encuentran con el valor cero o con el valor uno.

EvAEP emplea un esquema de aprendizaje de reglas iterativo en donde se extrae únicamente la mejor solución del algoritmo evolutivo, y el conjunto de reglas final se obtiene mediante sucesivas iteraciones del algoritmo evolutivo hasta que se cumple cierta condición de parada. Esta condición de parada es o bien la regla extraída no es patrón emergente, o no cubre ejemplos que no han sido cubiertos por reglas extraídas anteriormente. Con esta condición de parada se asegura la diversidad de los resultados obtenidos, a fin de cubrir la mayor superficie del espacio de búsqueda posible. Los operadores que emplea este algoritmo son una selección por torneo [27] de tamaño dos, un operador de cruce en dos puntos [19] y un operador de mutación sesgada [4].

3 Principales medidas de calidad

La medida de calidad más importante de la EPM es el GR, ya que es la medida con la se define la tarea. No obstante, es necesario analizar los patrones emergentes desde otros puntos de vista, como por ejemplo medir la generalidad, el interés y la precisión de las reglas obtenidas. Estas tres cualidades de una regla son fundamentales dentro del marco SDRD para la obtención de reglas altamente descriptivas.

La EPM fue concebida para el análisis de problemas entre dos clases o conjuntos de datos. Sin embargo, la tarea puede ser fácilmente extendida a problemas multiclase mediante estrategias como *One vs All* (OVA) [11]. En OVA, se considera como clase positiva a la clase representada en la regla y como clase negativa al resto de clases. Con esto, se puede representar una matriz de confusión de cada regla con la que se pueden calcular de manera sencilla las medidas de calidad deseadas.

Table 1: Matriz de confusión de una regla.

	Clase predicha		
Clase real	Positiva	Negativa	
Positiva	$p = tp$	$\overline{p} = fn$	$p + \overline{p} = P$
Negativa	$n = fp$	$\overline{n} = tn$	$n + \overline{n} = N$
	$p + n$	$\overline{p} + \overline{n}$	$P + N = T$

En la Tabla 1 se puede observar dicha matriz de confusión, donde: p representa el número de ejemplo

correctamente cubiertos, $\bar{p}$ como el número de ejemplos de la clase no cubiertos, n como el número de ejemplos cubiertos incorrectamente, $\bar{n}$ como el número de ejemplos no cubiertos que no pertenecen a la clase positiva, $\bar{p}+\bar{n}$ es el número de ejemplos no cubiertos por la regla, $p+n$ es el número de ejemplos cubiertos por la regla, $P = p+\bar{p}$ es el número de ejemplos de la clase positiva, $N = n+\bar{n}$ es el número de ejemplos de la clase negativa y $T = P+N$ indica el número total de ejemplos.

En un estudio comparativo llevado a cabo en [14] se determinaron grupos de medidas de calidad de SDRD en función de sus objetivos, basado en correlaciones de Pearson. Siguiendo el resultado de este trabajo, las medidas descriptivas de calidad a estudiar serán las siguientes:

- Growth Rate (GR). Definida en la Ecuación 1, mide el poder discriminativo de una regla.
- Confianza (Conf). Se define como el ratio de la capacidad predictiva de la regla para la clase positiva [10].

$$Conf(R) = \frac{p}{p+n} \tag{2}$$

- Atipicidad (Atip). Esta medida híbrida muestra el balance existente entre generalidad y ganancia de precisión de la regla [23].

$$Atip(R) = \frac{p+n}{P+N}\left(\frac{p}{p+n} - \frac{P}{P+N}\right) \tag{3}$$

  Para esta medida, el dominio tiene una dependencia directa con el porcentaje de la clase a medir, por lo tanto, para realizar comparaciones es necesario normalizar esta medida. Esta normalización se ha llevado a cabo de la siguiente manera [5]:

$$Atip\_Norm(R) = \frac{Atip(R) - \left(\frac{P}{T}\left(0 - \frac{P}{T}\right)\right)}{\left(\frac{P}{T}\left(1 - \frac{P}{T}\right)\right) - \left(\frac{P}{T}\left(0 - \frac{P}{T}\right)\right)} \tag{4}$$

- Tasa de falsos positivos (FPR). Mide el porcentaje de ejemplos incorrectamente cubiertos respecto al total de ejemplos de la clase negativa. Esta medida debe ser minimizada [12].

$$FPr(R) = \frac{n}{N} \tag{5}$$

- Tasa de verdaderos positivos (TPR). Mide el porcentaje de ejemplos correctamente cubiertos respecto al número total de ejemplos de la clase positiva [21].

$$TPr(R) = \frac{p}{P} \tag{6}$$

- Número de reglas. Esta medida se utiliza en el conjunto total de reglas y mide la cardinalidad de dicho conjunto.
- Número de variables. Esta medida también es aplicada al conjunto de reglas final. Mide el valor medio de variables que se obtienen en las reglas.

4 Estudio experimental

El objetivo de este trabajo es la determinación de la mejor representación del conocimiento para EPM en sistemas difusos evolutivos. Para llevar a cabo este objetivo se ha realizado una amplia experimentación sobre una batería de bases de datos. En esta sección se definirán todos los pasos realizados: en primer lugar, se muestras las características de los conjuntos de datos utilizados. A continuación, se muestran los parámetros utilizados para el algoritmo evolutivo. Por último, se muestran los resultados del estudio y un detallado análisis de los mismos.

4.1 Conjuntos de datos utilizados

Las diferentes representaciones del conocimiento han sido comparadas respecto a 67 conjuntos de datos del repositorio KEEL [2], cuyas características se muestran en la Tabla 2.

Estos conjuntos de datos son un compilación de problemas muy conocidos y trabajados en la literatura. La comparación se ha llevado a cabo mediante un procedimiento de 5-validación cruzada estratificada y óptimamente balanceada [28].

Table 2: Conjuntos de datos usados en el estudio, se incluye también el número de variables, ejemplos y clases de cada uno.

Nombre	# Var.	# Ej.	# Clases	Nombre	# Variables	# Ejemplos	# Clases
Abalone	8	4174	28	Movement_libras	90	360	15
Appendicitis	7	106	2	Mushroom	22	5644	2
Australian	14	690	2	Newthyroid	5	215	3
Automobile	25	150	6	Nursery	8	12690	5
Balance	4	625	3	Page-blocks	10	5472	5
Bands	19	365	2	Penbased	16	10992	10
Breast	9	277	2	Phoneme	5	5404	2
Bupa	6	345	2	Pima	8	768	2
Car	6	1728	4	Post-operative	8	87	3
Chess	36	3196	2	Ring	20	7400	2
Cleveland	13	297	5	Saheart	9	462	2
Coil2000	85	9822	2	Satimage	36	6435	7
Contraceptive	9	1473	3	Segment	19	2310	7
Crx	15	653	2	Shuttle	9	58000	7
Dermatology	34	358	6	Sonar	60	208	2
Ecoli	7	336	8	Spambase	57	4597	2
Flare	11	1066	6	Spectfheart	44	267	2
German	20	1000	2	Splice	60	3190	3
Glass	9	214	7	Tae	5	151	3
Haberman	3	306	2	Texture	40	5500	11
Hayes-roth	4	160	3	Thyroid	21	7200	3
Heart	13	270	2	Tic-tac-toe	9	948	2
Hepatitis	19	80	2	Titanic	3	2201	2
Housevotes	16	232	2	Twonorm	20	7400	2
Ionosphere	33	351	2	Vehicle	18	846	4
Iris	4	150	3	Vowel	13	990	11
Kr-vs-k	6	28056	17	Wdbc	30	569	2
Led7digit	7	500	10	Wine	13	178	3
Letter	16	20000	26	Winequality-red	11	1599	11
Lymphography	18	148	4	Winequality-white	11	4898	11
Magic	10	19020	2	Wisconsin	9	683	2
Mammographic	5	830	2	Yeast	8	1484	10
Marketing	13	6876	9	Zoo	16	101	7
Monk-2	6	432	2				

4.2 Parámetros utilizados

Para esta comparación se ha utilizado el único algoritmo evolutivo para EPM existente hasta la fecha, EvAEP, sobre los conjuntos de datos. La configuración utilizada en este algoritmo para ambas representaciones del conocimiento es la recomendada por lo autores [17].

Debido a que los sistemas difusos evolutivos son métodos no deterministas se ha ejecutado el algoritmo tres veces con diferentes semillas. Por lo tanto, se tienen 15 ejecuciones del algoritmo en cada

conjunto de datos.

4.3 Análisis de resultados

En esta sección se muestran los resultados obtenidos por las diferentes representaciones del conocimiento utilizadas con respecto a los diferentes conjuntos de datos, a fin de determinar la mejor de ellas.

Como únicamente se comparan dos casos, es decir, reglas de tipo canónico contra reglas DNF, se utilizará el test de Wilcoxon para determinar estadísticamente que representación del conocimiento es mejor para cada medida de calidad estudiada. El nivel de significancia utilizado en el estudio es de $\alpha = 0.05$. Debido al gran tamaño de los resultados obtenidos, estos se encuentran disponibles en la página web `http://simidat.ujaen.es/papers/MAEB2017` a fin de que cualquier persona interesada sea capaz de consultarlos.

En la Tabla 3 se muestra el p-valor obtenido por el test de Wilcoxon para cada medida de calidad. También se muestra la mediana de los resultados obtenidos por las diferentes representaciones del conocimiento en cada medida de calidad, pues es necesario para la realización de la comparativa.

Table 3: Resultados del estudio estadístico para la comparitiva de reglas canónicas frente reglas DNF. Se muestra la mediana de los resultados obtenidos para cada representación y medida, además del p-valor asociado al test de Wilcoxon.

	Atip	Conf	GR	TPr	FPr	Nº Reglas	Nº Variables
CAN	0.5433	0.6340	0.7397	0.3952	**0.1482**	9.7333	4.3629
DNF	**0.5611**	**0.6552**	**0.7777**	**0.5141**	0.1828	**8.1333**	**3.1744**
p-valor	**0.0006**	**0.0350**	**0.0131**	**0.0000**	**0.0000**	**0.0099**	**0.0000**

En esta tabla, la mejor mediana para cada medida de calidad es resaltada. Asimismo, el p-valor resaltado en negrita indica que el valor obtenido está por debajo del valor de significancia, indicando por tanto resultados significativos.

Tal y como se puede observar, los resultados de la tabla ofrecen resultados significativos en todas las medidas de calidad estudiadas. De estos resultados obtenidos, se destaca:

- Atipicidad. La representación en forma normal disyuntiva obtiene el mejor resultado de manera significativa, superando a la representación canónica en casi dos puntos. Este resultado indica que las reglas de este tipo aportan al experto una información más relevante que las reglas de tipo canónico.
- Confianza. En esta medida sucede un comportamiento similar al anterior. En este aspecto las reglas DNF obtienen una confianza más elevada que en reglas de tipo canónico. Esto se debe principalmente a la flexibilidad de este tipo de reglas, pues permiten ajustarse mucho mejor al conjunto de datos y, por tanto, ser más precisas.
- GR. En esta medida se mide el porcentaje de patrones emergentes que siguen siendo emergentes en los datos de test. Esto se debe a la imposibilidad de promediar los resultados de esta medida debido a su dominio $[0,\infty]$. Tal y como se puede observar, las reglas en formato DNF obtienen un conjunto de reglas con un porcentaje de patrones más elevado que las reglas de tipo canónico.
- TPr. Para esta medida el resultado demuestra una mayor generalidad de las reglas de tipo DNF respecto a las canónicas. Esto implica que las reglas obtenidas son capaces de cubrir una mayor cantidad de ejemplos de la clase considerada positiva. Las reglas DNF son capaces de generalizar mejor gracias a la capacidad de seleccionar diferentes valores o LLs para cada variable, aumentando en gran medida su cobertura.
- FPr. En esta medida es mejor de manera significativa las reglas de tipo canónico. Esto es debido a que estas reglas no tienen tanta capacidad de generalización como las reglas DNF. Una generalidad mayor implica un riesgo de error más elevado que se ve claramente reflejado en este resultado.

- Nº de Reglas. En esta caso las reglas DNF obtienen un conjunto de reglas más reducido que las reglas canónicas. En este aspecto, la representación utilizada por las reglas DNF tiende a que el conjunto de reglas sea reducido, pues una regla DNF puede ser representada mediante varias reglas canónicas.

- Nº de variables. Para esta medida, el número de variables medio que participan en cada regla para cada conjunto de reglas es menor para DNF. Esto se debe igualmente a la propia representación de reglas DNF, pues permiten que una regla contenga más valores de una variable en concreto. Esto fomenta que se formen reglas que cubran más ejemplos y que, por tanto, impiden la creación de reglas más específicas con el objetivo de cubrir los ejemplos restantes.

En general, los resultados obtenidos muestran una clara ventaja de las reglas DNF respecto a las reglas canónicas. Estas obtienen un conjunto de reglas más reducido y con un menor número de variables en el antecedente, lo que facilita la comprensión del conocimiento extraído por parte de los expertos. Asimismo, las reglas DNF obtienen reglas más interesantes, más generales y más precisas que las reglas de tipo canónico. Por lo tanto, las reglas en formato DNF son una representación del conocimiento que debe ser adoptada en el desarrollo de nuevas estrategias evolutivas para EPM.

5 Conclusiones

En este trabajo se ha presentado una comparativa de diferentes representaciones del conocimiento que pueden ser adoptadas en sistemas difusos evolutivos para EPM respecto a las características descriptivas señaladas en el marco SDRD. En particular, se ha presentado una comparativa entre reglas de tipo canónico (conjunciones de pares atributo-valor) frente a reglas en forma normal disyuntiva. La comparativa se ha llevado a cabo utilizando el único algoritmo evolutivo desarrollado para EPM hasta la fecha, el algoritmo EvAEP. El resultado de este estudio muestra una clara preferencia por las reglas de tipo DNF, las cuales obtienen un conjunto de reglas más reducido, con menor número de variables y que, en general, son mas interesantes, precisas y generales que las reglas de tipo canónico. Estas características indican que el desarrollo de futuras estrategias evolutivas debe estar orientado a una representación del conocimiento en formato DNF, a fin de obtener una descripción del conjunto de datos más relevante y fácilmente comprensible por los expertos.

Agradecimientos

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

References

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and data mining*, pages 307–328. AAAI Press, 1996.

[2] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17(2-3):255–287, 2011.

[3] P. W. Angriyasa, Z. Rustam, and W. Sadewo. Non-invasive intracranial pressure classification using strong jumping emerging patterns. In *Proc. of the 2011 International Conference on Advanced Computer Science and Information System (ICACSIS)*, pages 377–380. IEEE, 2011.

[4] C. J. Carmona, P. González, M. J. del Jesus, and F. Herrera. NMEEF-SD: Non-dominated Multiobjective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery. *IEEE Transactions on Fuzzy Systems*, 18(5):958–970, 2010.

[5] 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*, 298:180–197, 2015.

[6] K. Deb. *Multi-Objective Optimization using Evolutionary Algorithms*. John Willey & Sons, 2001.

[7] 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*, pages 43–52. ACM Press, 1999.

[8] G. Z. Dong and J. Y. Li. Mining border descriptions of emerging patterns from dataset pairs. *Knowledge and Information Systems*, 8(2):178–202, 2005.

[9] A. E. Eiben and J. E. Smith. *Introduction to evolutionary computation*. Springer, 2003.

[10] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: an overview. In *Advances in knowledge discovery and data mining*, pages 1–34. AAAI/MIT Press, 1996.

[11] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, 44(8):1761 – 1776, 2011.

[12] D. Gamberger and N. Lavrac. Expert-Guided Subgroup Discovery: Methodology and Application. *Journal Artificial Intelligence Research*, 17:501–527, 2002.

[13] T. Gambin and K. Walczak. A new classification method using array comparative genome hybridization data, based on the concept of limited jumping emerging patterns. *BMC bioinformatics*, 10(1):1, 2009.

[14] M. García-Borroto, O. Loyola-Gonzalez, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa. *Comparing Quality Measures for Contrast Pattern Classifiers*, pages 311–318. Springer Berlin Heidelberg, 2013.

[15] M. García-Borroto, J.F. Martínez-Trinidad, and J.A. Carrasco-Ochoa. Fuzzy emerging patterns for classifying hard domains. *Knowledge and Information Systems*, 28(2):473–489, 2011.

[16] A. M. García-Vico, C. J. Carmona, P. González, and M. J. del Jesus. Minería de patrones emergentes: Una oportunidad para la extracción evolutiva de conocimiento. In *Proc. of the XVII Conference of the Spanish Association for Artificial Intelligence (CAEPIA)*, pages 149–159, Salamanca, Spain, 2016.

[17] A. M. García-Vico, J. Montes, J. Aguilera, C. J. Carmona, and M. J. del Jesus. Analysing Concentrating Photovoltaics Technology through the use of Emerging Pattern Mining. In *Proc. of the 11th International Conference on Soft Computing Models in Industrial and Environmental Applications*, pages 1–8. Springer, 2016.

[18] F. Herrera. Genetic fuzzy systems: taxomony, current research trends and prospects. *Evolutionary Intelligence*, 1:27–46, 2008.

[19] J. H. Holland. Adaptation in natural and artificial systems. *University of Michigan Press*, 1975.

[20] E. Hüllermeier. Fuzzy methods in machine learning and data mining: Status and prospects. *Fuzzy Sets and Systems*, 156(3):387–406, 2005.

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

[22] 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*, 10:377–403, 2009.

[23] 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*, volume 1634 of *LNCS*, pages 174–185. Springer, 1999.

[24] A. Lepailleur, G. Poezevara, and R. Bureau. Automated detection of structural alerts (chemical fragments) in (eco) toxicology. *Computational and structural biotechnology journal*, 5(6):1–8, 2013.

[25] G. Li, R. Law, H. Q. Vu, J. Rong, and X. R. Zhao. Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism management*, 46:311–321, 2015.

[26] E. Loekito and J. Bailey. Fast mining of high dimensional expressive contrast patterns using zero-suppressed binary decision diagrams. In *Proc. of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 307–316, 2006.

[27] B. L. Miller and D. E. Goldberg. Genetic Algorithms, Tournament Selection, and the Effects of Noise. *Complex System*, 9:193–212, 1995.

[28] J. G. Moreno-Torres, J. A. Sáez, and F. Herrera. Study on the impact of partition-induced dataset shift on-fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1304–1312, 2012.

[29] M. Piao, H. G. Lee, G. Y. Sohn, G. Pok, and K. H. Ryu. Emerging patterns based methodology for prediction of patients with myocardial ischemia. In *Proc. of the 6th International Conference on Fuzzy Systems and Knowledge Discovery*, pages 174–178. IEEE, 2009.

[30] R. Sherhod, V. J. Gillet, T. Hanser, P. N. Judson, and J. D. Vessey. Toxicological knowledge discovery by mining emerging patterns from toxicity data. *Journal of Chemical Information and Modeling*, 5(S-1):9, 2013.

[31] G. Tzanis, I. Kavakiotis, and I. P. Vlahavas. Polya-iep: A data mining method for the effective prediction of polyadenylation sites. *Expert Systems with Applications*, 38(10):12398–12408, 2011.

[32] L. Wang, H. Zhao, G. Dong, and J. Li. On the complexity of finding emerging patterns. In *Proc. of the 28th Annual International Computer Software and Applications Conference*, volume 2, pages 126–129, 2004.

[33] Y. Yu, K. Yan, X. Zhu, and G. Wang. Detecting of PIU Behaviors Based on Discovered Generators and Emerging Patterns from Computer-Mediated Interaction Events. In *Proc. of the 15th International Conference on Web-Age Information Management*, volume 8485 of *LNCS*, pages 277–293. Elsevier, 2014.

[34] L. A. Zadeh. Fuzzy sets. *Information Control*, 8:338–353, 1965.

[35] L. A. Zadeh. The concept of a linguistic variable and its applications to approximate reasoning. Parts I, II, III. *Information Science*, 8-9:199–249,301–357,43–80, 1975.

[36] L. A. Zadeh. Soft Computing and Fuzzy Logic. *IEEE Software*, 11(6):48–56, 1994.