# E2PAMEA: un algoritmo evolutivo para la extracción eficiente de patrones emergentes difusos en entornos big data

A.M. Garcia-Vico
*Instituto Andaluz Interuniversitario en Data Science and Computational Intelligence (DaSCI)*
*Universidad de Granada*, Granada, España
agvico@decsai.ugr.es
D. Elizondo
*School of Computer Science and Informatics*
*DeMonfort University*, Leicester, Reino Unido
elizondo@dmu.ac.uk
F. Charte, P. González, C. J. Carmona
*Instituto Andaluz Interuniversitario en Data Science and Computational Intelligence (DaSCI)*
*Universidad de Jaén*, Jaén, España
{fcharte|pglez|ccarmona}@ujaen.es

***Resumen*—Este trabajo es un resumen del artículo publicado por los autores en la revista *Neurocomputing* [1] en el que se presenta un modelo evolutivo multi-objetivo adaptativo junto a un sistema de evaluación eficiente para la extracción de patrones emergentes difusos en entornos de datos masivos.**

***Index Terms*—Descubrimiento de reglas descriptivas supervisadas, minería de patrones emergentes, algoritmos evolutivos multi-objetivo, sistemas difusos evolutivos, Big Data.**

## I. Resumen

La cantidad de información que se genera actualmente crece de manera exponencial y habitualmente no puede ser procesada por una única máquina, por lo que la aplicación de técnicas de computación distribuida es obligatoria. A esta cantidad ingente de datos se denomina en la literatura como *big data* [2].

Dentro de este ámbito, la extracción de conocimiento capaz de describir, de manera comprensible, el comportamiento de los datos respecto a una variable de interés para el experto es fundamental. Entre otras tareas, la minería de patrones emergentes (EPM) [3], [4] es útil para este propósito. EPM se define como la búsqueda de patrones que, dados dos conjuntos de datos $D_1$ y $D_2$, tengan un índice de crecimiento (GR) mayor a un umbral $\rho > 1$. Concretamente, consiste en describir comportamientos que se produzcan mayoritariamente en un único conjunto de datos. Esto hace que los patrones extraídos sean muy discriminativos. Para calcular el GR, es necesario determinar qué instancias cumplen las características descritas por un patrón en concreto, para posteriormente poder determinar su calidad. Esto se lleva a cabo a través del cálculo de una matriz de contingencia de cuatro valores ($tp$, $fp$, $fn$, $tn$) que determinan el número de instancias cubiertas/no cubiertas de manera correcta/incorrecta, respectivamente. Con estos cuatro valores, el GR se define como $GR = \frac{tp(fp+tn)}{fp(tp+fn)}$.

No obstante, en EPM es necesario que los patrones extraídos tengan, además de una gran capacidad discriminativa, gran capacidad descriptiva. Para ello, se debe encontrar un balance entre varios objetivos como generalidad, precisión e interés. Estos se determinan a su vez a partir de diferentes métricas, como por ejemplo WRAcc [5], entre otras. Estos objetivos son conflictivos entre sí, de modo que si aumentamos el valor de una, disminuye el valor del resto. Por tanto, las metaheurísticas multi-objetivo son adecuadas para la búsqueda de conocimiento con un buen balance entre ellos. En concreto, los métodos evolutivos han sido especialmente exitosos en la literatura para esta tarea [6], [7].

A pesar de sus buenos resultados, el principal problema de estas técnicas es su escalabilidad. Esto se debe principalmente a que para calcular estas métricas objetivo, se requiere determinar para cada patrón candidato su matriz de contingencia asociada. Para obtenerla se necesita recorrer todo el conjunto de datos completo, lo que en entornos *big data* es muy costoso. El paradigma MapReduce [8] es la estrategia más habitual en la literatura para aliviar este problema. Dentro de este paradigma, se han empleado dos enfoques diferentes: 1) un enfoque local, donde se ejecuta un método no distribuido en cada partición de datos y 2) un enfoque global, donde una parte, o el método completo, se rediseña para su trabajo de manera distribuida. Los enfoques locales son fácilmente escalables, pero el diseño del método de agregación no es trivial y su calidad depende de cómo se han distribuido los datos en cada partición. Por otro lado, los enfoques globales habitualmente permiten extraer el resultado exacto independientemente del particionado empleado. Normalmente, un enfoque global es deseable respecto a un enfoque local. Sin embargo, su diseño suele ser más complejo y su tiempo de ejecución más lento.

Nuestra aportación a la literatura en este trabajo es doble:

por un lado se presenta un método de evaluación distribuido denominado Bit-LUT junto a un método evolutivo multi-objetivo de extracción de patrones emergentes difusos denominado E2PAMEA. El método Bit-LUT se fundamenta en la premisa de que las definiciones de las etiquetas lingüísticas difusas no se modifican durante el proceso evolutivo, por lo que se puede precalcular el grado de pertenencia de cada ejemplo a cada etiqueta para mejorar la eficiencia. En concreto, Bit-LUT almacena para cada par variable-valor $(i, j)$ del problema un vector binario $BS_i^j$ que indica si una instancia $k$ es cubierta o no por dicho par. Este proceso de precálculo se lleva a cabo mediante un proceso MapReduce, en donde cada partición calcula un subconjunto de este vector $BS_i^j$ en la fase *map*, mientras que la fase *reduce* únicamente concatena estos vectores para obtener el resultado final.

Por otro lado, E2PAMEA es un enfoque evolutivo multi-objetivo basado en un enfoque "cromosoma = regla" en el que un individuo de la población representa un potencial patrón. El resultado final es el conjunto formado por la unión de varios individuos. Estos interactúan entre sí mediante el empleo de un enfoque cooperativo-competitivo en donde los individuos compiten entre sí debido al empleo de los operadores genéticos mientras que, por otro lado, colaboran en la búsqueda del mejor conjunto de patrones final gracias al empleo de una población élite. En esta población élite se aplica el operador de competición de tokens [9] para buscar la población élite con el valor de atipicidad medio más elevado. Respecto a los operadores genéticos empleados, se emplean dos operadores de cruce: cruce en dos puntos y HUX; y dos operadores de mutación: cambio aleatorio de un gen y eliminación al azar de una variable completa. Estos operadores se aplican con una probabilidad adaptativa en función del número de individuos generados que han sobrevivido de una generación a la siguiente. De este modo, se evita la necesidad de optimizar manualmente estos parámetros. Para llevar a cabo una evaluación eficiente E2PAMEA hace uso de Bit-LUT para determinar los ejemplos cubiertos por un individuo a través del proceso mostrado en la Ecuación (1).

$$Cov(P_i) = \left(BS_1^1 \mid BS_1^2 \mid \ldots \mid BS_1^j\right) \& \ldots \& \left(BS_i^1 \mid BS_i^2 \mid \ldots \mid BS_i^j\right) \quad (1)$$

donde $\mid$ y $\&$ son las operaciones de bits OR y AND, respectivamente. Esta operación se distribuye a su vez mediante un proceso MapReduce. Para ello, la fase *map* se encarga del cómputo de las operaciones OR para cada variable del problema, mientras que la fase *reduce* se encargaría de aplicar todas las operaciones AND para obtener el resultado final. Para mejorar la eficiencia este cómputo se realiza para todos los individuos de la población.

La validez del método propuesto se estudia a través de un exhaustivo estudio experimental en el que se emplean 6 conjuntos de datos de gran dimensionalidad en el que se busca determinar la calidad del conocimiento extraído y el tiempo de ejecución empleado con respecto al método BD-EFEP [10], uno de lo métodos evolutivos más prometedores del ámbito. Finalmente, se determina la escalabilidad de la propuesta respecto a ambas dimensiones del conjunto de datos.

En el estudio experimental realizado se demuestra que la calidad del conocimiento extraído es más precisa, más interesante y más general en E2PAMEA que en BD-EFEP, debido principalmente al enfoque adaptativo propuesto. Por otro lado, el tiempo de ejecución se reduce en hasta 60 veces respecto a BD-EFEP, mientras que su consumo de memoria se reduce en hasta 10 veces gracias principalmente al empleo de conjuntos de bits y operaciones bits, que son altamente eficientes. Finalmente, como resultado del análisis de escalabilidad empleado se determina que el coste computacional del método propuesto aumenta de manera lineal respecto al número de instancias y variables. En concreto, se han conseguido procesar hasta $10^8$ instancias con 100 pares variable-valor y $10^6$ instancias con hasta 75000 pares variable-valor en aproximadamente 5000 segundos.

## AGRADECIMIENTOS

Este trabajo ha sido subvencionado por la Consejería de Transformación Económica, Industria, Conocimiento y Universidades de la Junta de Andalucía, programa Personal Investigador Doctor, referencia DOC_00235.

## REFERENCIAS

[1] A. García-Vico, F. Charte, P. González, D. Elizondo, and C. J. Carmona, "E2pamea: A fast evolutionary algorithm for extracting fuzzy emerging patterns in big data environments," *Neurocomputing*, vol. 415, pp. 60 – 73, 2020.

[2] M. A. Beyer and D. Laney, "The importance of big data: a definition," *Gartner Research Report*, pp. 1–9, 2012.

[3] G. Dong and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 1999, pp. 43–52.

[4] A. M. García-Vico, C. J. Carmona, D. Martín, M. García-Borroto, and M. J. del Jesus, "An overview of emerging pattern mining in supervised descriptive rule discovery: Taxonomy, empirical study, trends and prospects," *WIREs: Data Mining and Knowledge Discovery*, vol. 8, no. 1, 2018.

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

[6] F. Pulgar-Rubio, A. J. Rivera-Rivas, M. D. Pérez-Godoy, P. González, C. J. Carmona, and M. J. del Jesus, "MEFASD-BD: Multi-Objective Evolutionary Fuzzy Algorithm for Subgroup Discovery in Big Data Environments - A MapReduce Solution," *Knowledge-Based Systems*, vol. 117, pp. 70–78, 2017.

[7] A. M. García-Vico, C. J. Carmona, P. González, and M. J. del Jesus, "MOEA-EFEP: Multi-objective evolutionary algorithm for extracting fuzzy emerging patterns," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 5, pp. 2861 – 2872, 2018.

[8] N. Desai and A. Ganatra, "Incorporating boundary value concept and recency constraint to capture emerging trends in time stamp based sequence dataset," in *Proc. of the 2015 International Conference on Communication, Information and Computing Technology*, 2015, pp. 1–7.

[9] K. S. Leung, Y. Leung, L. So, and K. F. Yam, "Rule Learning in Expert Systems Using Genetic Algorithm: 1, Concepts," in *Proc. of the 2nd International Conference on Fuzzy Logic and Neural Networks*, K. Jizuka, Ed., 1992, pp. 201–204.

[10] A. M. García-Vico, C. J. Carmona, P. González, and M. J. del Jesus, "A big data approach for extracting fuzzy emerging patterns," *Cognitive Computation*, vol. 11, no. 3, pp. 400 – 417, 2019.