MOEA-EFEP: Un algoritmo evolutivo multi-objetivo para la extracción de patrones emergentes difusos

A.M. Garcia-Vico, C.J. Carmona, P. González, M.J. del Jesus
Departamento de Informática, Instituto Andaluz Interuniversitario en Data Science and Computational Intelligence
Universidad de Jaén, Jaén, España
{agvico|ccarmona|pglez|mjjesus}@ujaen.es

***Resumen*—Este trabajo es un resumen del publicado por los autores en la revista *IEEE Transactions on Fuzzy Systems* [1] en el que se presenta un nuevo modelo evolutivo multi-objetivo para la extracción de patrones emergentes difusos con un gran balance entre la capacidad descriptiva de los mismos y su fiabilidad.**

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

I. Resumen

Tradicionalmente, en minería de datos se distinguen dos enfoques claramente diferenciados: un enfoque predictivo, cuyo objetivo es la obtención de un modelo para la predicción del valor de una variable de interés en nuevas instancias no vistas hasta el momento, utilizando para ello aprendizaje supervisado; y un enfoque descriptivo con el objetivo de crear un modelo que describa relaciones interesantes en los datos utilizando para ello habitualmente aprendizaje no supervisado. Sin embargo, a lo largo de la literatura se han ido desarrollando una serie de técnicas que se encuentran a medio camino entre ambos enfoques, agrupadas bajo el nombre de "descubrimiento de reglas descriptivas basadas en aprendizaje supervisado" (SDRD) [2], [3] cuyo propósito es la obtención de un modelo descriptivo con de conocimiento relevante sobre una variable de interés en un conjunto de datos.

El principal objetivo de las técnicas SDRD no es la extracción de un modelo con el fin de clasificar nuevas instancias, sino la obtención de un modelo que permita describir de una manera simple y fácilmente comprensible el fenómeno subyacente en los datos por parte de los expertos. Por tanto, en este grupo de tareas se agrupan todas aquellas técnicas de minería de datos que utilizan un modelo de reglas y aprendizaje supervisado para obtener conocimiento descriptivo sobre los datos, como por ejemplo el descubrimiento de subgrupos [4], [5], la minería de conjuntos de contraste [6] o la minería de patrones emergentes [7], [8], entre otras.

La minería de patrones emergentes (EPM) se define como la búsqueda de todos los patrones que, dados dos conjuntos de datos D_1 y D_2, tengan un índice de crecimiento (GR) mayor a un umbral $\rho > 1$. Este índice de crecimiento se define como la siguiente función:

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

donde $Sop_i(X)$ indica el soporte del patrón X en el conjunto de datos i. EPM tiene como objetivos principales la descripción de características discriminativas entre clases, la descripción de fenómenos emergentes en datos temporales, donde el primero de estos objetivos ha sido el más desarrollado a lo largo de la literatura.

El principal problema de la extracción de patrones emergentes reside en su propia definición, ya que definen un espacio de búsqueda que no es convexo debido a que el GR es una proporción de soportes [9]. Esto permite que aquellos patrones con soportes más altos puedan tener un GR menor que aquellos patrones con soportes menores. Por esta razón a lo largo de la literatura se han desarrollado diferentes tipos de patrones emergentes, como por ejemplo los patrones emergentes *Jumping* [10] o los patrones emergentes χ^2 [11]. Asimismo, se han desarrollado diferentes técnicas de extracción como las técnicas basadas en límites o en árboles con el objetivo de extraer de manera eficiente aquellos de mayor calidad y poder discriminativo. Sin embargo, la gran mayoría de estos métodos han sido completamente enfocados a clasificación ignorando por completo las capacidades descriptivas de los mismos. No obstante, en los últimos años se han desarrollado técnicas basadas en algoritmos evolutivos que han permitido la extracción de patrones emergentes con un buen balance entre las capacidades descriptivas y la fiabilidad de las mismas [8].

Nuestra aportación a la literatura es un algoritmo genético multi-objetivo que permite la extracción de patrones emergentes difusos llamado MOEA-EFEP (*Multi-Objective Evolutionary Algorithm for the Extraction of Fuzzy Emerging Patterns*). MOEA-EFEP utiliza lógica difusa para el tratamiento de variables de tipo numérico con el fin de evitar pérdidas de información y mejorar la interpretabilidad de los resultados. Asimismo, MOEA-EFEP está basado en el algoritmo multi-objetivo NSGA-II [12], el cual ha sido modificado para que el proceso de búsqueda esté orientado a la extracción de patrones

emergentes de gran calidad, es decir, un conjunto de patrones simple, con gran capacidad de generalización y fiable.

En concreto, MOEA-EFEP emplea un enfoque "cromosoma = regla" en el que un individuo de la población representa una regla potencial y el resultado final es el conjunto formado por la unión de varios individuos. En este punto, es importante remarcar que MOEA-EFEP representa tanto el antecedente como el consecuente de la regla por lo que puede extraer patrones emergentes para todas las clases del problema en una única ejecución. Asimismo, se permite el uso de dos representaciones del conocimiento diferentes en función de las necesidades del problema: reglas canónicas, formadas por conjunciones de pares atributo-valor o bien reglas en forma normal disyuntiva (DNF) donde se permite que un atributo pueda tener más de un posible valor mediante disyunciones.

Los individuos de la población interactúan entre si mediante el uso de un enfoque cooperativo-competitivo. Este enfoque se basa, por un lado, en la competición propia de un algoritmo evolutivo a través de los operadores genéticos. Dichos operadores son, en concreto, un operador de inicialización guiada, que genera individuos con pocas variables y, por tanto, muy generales; un operador de cruce multi-punto; un operador de mutación orientada con capacidad para eliminar o modificar aleatoriamente un gen de un individuo; y el operador de reemplazo de la población basado en la ordenación rápida por frentes de dominancia propia del algoritmo NSGA-II. Asimismo, se utiliza una población élite y un mecanismo de competición adicional basado en el proceso de competición de tokens [13]. En la actualización de esta población élite es cuando los individuos cooperan entre sí con el objetivo de obtener una población élite con la atipicidad [3] media más elevada. Esta población élite se actualiza en el operador de reinicialización, que se activa si la población actual no ha sido capaz de cubrir ejemplos no cubiertos anteriormente por un 5 % del total de las evaluaciones totales con el objetivo de evitar estancamientos en óptimos locales.

Finalmente, una vez termina el proceso evolutivo, se realiza un filtro de post-procesamiento con el fin de eliminar aquellas reglas que no tengan una calidad suficiente para el experto. En concreto, para el algoritmo MOEA-EFEP se proponen tres filtros diferentes: obtener patrones con un valor de confianza superior al 60 %, obtener sólo patrones minimales u obtener sólo patrones maximales.

La validez del método propuesto se estudia a través de un exhaustivo estudio experimental con 50 conjuntos de datos con tres objetivos diferentes: por un lado, se plantea la elección de la mejor representación del conocimiento; por otro, se plantea la elección del mejor filtrado de post-procesamiento y, finalmente, se compara MOEA-EFEP con aquellos métodos más relevantes para EPM en función de su metodología de extracción de patrones tal y como se expone en [8]. Todos los resultados obtenidos han sido avalados mediante el uso de test estadísticos no-paramétricos.

En el estudio experimental realizado se demuestra que la capacidad de las reglas en forma normal disyuntiva para la extracción de patrones emergentes con un mayor equilibrio entre generalidad y fiabilidad de los resultados. Por otro lado, se demostró la calidad de las reglas extraídas mediante el uso del filtro de post-procesamiento basado en valores de confianza mayores al 60 %. Por último, MOEA-EFEP superó de manera significativa al resto de métodos comparados en cuanto a la generalidad de las reglas descriptivas con una fiabilidad similar utilizando para ello un número de reglas mucho menor. Por lo tanto, MOEA-EFEP es un algoritmo que obtiene un modelo de patrones emergentes con el mejor balance generalidad-fiabilidad así como el modelo de reglas más simple.

Agradecimientos

Este trabajo ha sido subvencionado por el Ministerío de Economía y Competitividad bajo el proyecto TIN2015-68454-R y el contrato predoctoral FPI referencia BES-2016-077738 asociado al mismo (Fondos FEDER).

Referencias

[1] 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*, In Press.

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

[3] 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.

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

[5] S. Wrobel, "An Algorithm for Multi-relational Discovery of Subgroups," in *Proc. of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery*, ser. LNAI, vol. 1263. Springer, 1997, pp. 78–87.

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

[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*. ACM Press, 1999, pp. 43–52.

[8] 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.

[9] 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*, vol. 2, 2004, pp. 126–129.

[10] G. Z. Dong, J. Y. Li, and X. Zhang, "Discovering jumping emerging patterns and experiments on real datasets," in *Proc. on International Database Conference Heterogeneous and Internet Databases*, 1999, pp. 155–168.

[11] K. Ramamohanarao and H. Fan, "Patterns Based Classifiers," *World Wide Web*, vol. 10, no. 1, pp. 71–83, 2007.

[12] K. Deb, A. Pratap, S. Agrawal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[13] 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.