FEPDS: Una propuesta para la extracción de patrones emergentes difusos en flujos continuos de datos

A. M. Garcia-Vico
Instituto Andaluz Interuniversitario en Data Science and Computational Intelligence (DaSCI)
Universidad de Granada, Granada, España
agvico@decsai.ugr.es
H. Seker
Faculty of Computing, Engineering and the Build Environment, *Birmingham City University*, Birminghan, Reino Unido
huseyin.seker2@bcu.ac.uk
C. J. Carmona, P. González, M. J. del Jesus
Instituto Andaluz Interuniversitario en Data Science and Computational Intelligence (DaSCI)
Universidad de Jaén, Jaén, España
{ccarmona|pglez|mjjesus}@ujaen.es

***Resumen*—Este trabajo es un resumen del artículo publicado por los autores en la revista *IEEE Transactions on Fuzzy Systems* [1] en el que se presenta un modelo evolutivo multi-objetivo para la extracción de patrones emergentes difusos en flujos continuos de datos con gran capacidad descriptiva. Asimismo, la propuesta es aplicada en un caso de determinación en tiempo real de perfiles de los usuarios de taxis de la ciudad de Nueva York.**

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

I. Resumen

La minería de flujos de datos consiste en la extracción de conocimiento en una secuencia ordenada y potencialmente infinita de instancias que llegan al sistema a lo largo del tiempo a una velocidad variable [2]. Debido al incremento exponencial de dispositivos inteligentes e interconectados entre sí, este tipo de técnicas está cobrando una mayor relevancia actualmente. Sin embargo, las características de este tipo de datos suponen un reto para la extracción de conocimiento. Esto se debe principalmente a las estrictas restricciones de memoria, tiempo de cómputo y adaptación continua del modelo a la naturaleza cambiante de los datos [3].

Dentro de este ámbito, una de las aplicaciones de interés es describir o monitorizar el comportamiento de los datos respecto a una variable de interés para el experto. De este modo, se puede realizar una toma de decisiones con conocimiento del estado actual del flujo. Entre otras tareas, la minería de patrones emergentes (EPM) [4], [5] es útil para este propósito. EPM se define como la búsqueda de patrones que, dados dos conjuntos de datos, 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 altamente discriminativos.

A pesar del interés que puede suscitar el conocimiento extraído por EPM, no se han encontrado métodos de este tipo para minería de flujos de datos. Aunque muchos de los métodos en la literatura emplean un esquema incremental basado en árboles [6], éstos se rigen por unos supuestos que son incompatibles con la minería de flujos de datos. Por ejemplo, asumen la disponibilidad total de los datos y no contemplan mecanismos de actualización y olvido de datos obsoletos a lo largo del tiempo. Además, la mayoría están enfocados a maximizar la precisión, ignorando por completo las bondades descriptivas de estos patrones.

En EPM es necesario que los patrones extraídos tengan, además de una gran capacidad discriminativa, gran capacidad descriptiva. Esto es especialmente relevante en minería de flujo de datos, ya que una posible aplicación es la monitorización del estado del sistema. 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 [7], 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 [8], [6]. Sin embargo, el computo de estas métricas también requiere de un conjunto finito de datos para calcularas, lo que dificulta su aplicación en ámbitos de flujos continuos de datos.

Nuestra aportación a la literatura en este trabajo es un algoritmo evolutivo multi-objetivo capaz de extraer y adaptar de manera continua el conocimiento extraído en flujos de datos. El método, llamado FEPDS (*Fuzzy Emerging Patterns*

in Data Streams) tiene como principales características el procesamiento de los datos del flujo mediante bloques de tamaño fijo, un proceso de adaptación del aprendizaje basado en un enfoque ciego y un sistema evolutivo difuso multi-objetivo como método de aprendizaje con un sistema de recompensa que permite olvidar patrones antiguos. El empleo de un esquema de procesamiento por bloques de datos de tamaño fijo nos permite el cómputo directo de las métricas habituales en EPM. Por otro lado, el esquema de adaptación ciego, que ejecuta el método evolutivo por cada bloque nuevo, permite una mejor adaptación a cambios de tipo gradual [9]. Finalmente, para mejorar la adaptación, se incluye una ventana deslizante de tamaño fijo que almacena el conocimiento extraído en los m bloques de datos previos. Esta ventana será usada en el proceso evolutivo duranta la aplicación del sistema de recompensa propuesto.

El método de aprendizaje es un sistema difuso evolutivo basado en el algoritmo NSGA-II [10]. Se emplea un esquema "cromosoma=regla" en el que un individuo representa a un potencial patrón. Siguiendo la metodología de aprendizaje adaptativa, la población de individuos se inicia aleatoriamente pero incluyendo los patrones extraídos en el bloque anterior. Tras esto, el proceso evolutivo se lleva a cabo a partir del empleo de los operadores clásicos de cruce en dos puntos y mutación sesgada. Finalmente, se aplica un operador de reinicio de la población actual que también actualiza una población élite cuando el proceso se estanca. Se asume estancamiento cuando no se cubren nuevos ejemplos durante un 25 % de las iteraciones totales.

Este proceso de reinicio se basa en el procedimiento de competición de tokens [11], modificado con un esquema de recompensas. Este sistema refuerza positivamente aquellos patrones que han sido extraídos recientemente, pues la probabilidad de que el área del espacio de búsqueda cercana a estos sea prometedora es alta. Sin embargo, conforme el tiempo avanza, esta probabilidad decae debido a la naturaleza cambiante de los datos. Por tanto, se debe dar menos recompensa a los patrones extraídos con mayor antigüedad. Esto se consigue con la fórmula presentada en la Ecuación 1.

$$Div_t(fEP_i) = WRAcc_t(P_i) + \sum_{j=t-m}^{t} SW(P_i, j) \cdot 2^{-(t-j)} \cdot WRAcc_j(P_i) \quad (1)$$

donde $WRAcc_t$ es el valor de WRAcc en el instante t y SW es una función que devuelve uno si el patrón P_i fue extraído en el instante j o cero en caso contrario. Nótese que para poder usar SW es necesario determinar una ventana deslizante que almacene los patrones extraídos en los m bloques de datos anteriores.

La validez del método propuesto se analiza a través de un exhaustivo estudio experimental formado por 94 flujos de datos artificiales con diferentes características de cambio, en donde se analizan la calidad de los patrones extraídos, su capacidad de adaptación al cambio de concepto y su tiempo de cómputo y escalabilidad. Los resultados demuestran que el método propuesto es capaz de procesar bloques de hasta 15000 instancias en dos segundos. La calidad media de los patrones extraídos es buena para realizar un análisis rápido y de calidad. Finalmente, se demuestra que el método presenta una capacidad rápida de adaptación del conocimiento a los cambios en los datos.

Tras determinar las bondades del método sobre datos artificiales, se aplica el algoritmo propuesto en un caso real. El objetivo de este estudio es doble: por un lado se pretende obtener los perfiles de usuario en tiempo real de los taxis de la ciudad de Nueva York. Por otro lado, con el conocimiento extraído y almacenado a lo largo del tiempo, se pretenden extraer perfiles de usuario más generales. Los resultados de este estudio muestran una gran calidad en la extracción de conocimiento en tiempo real aplicable a niveles operativos, mientras que también puede usarse para realizar análisis de calidad a largo plazo en niveles niveles estratégicos.

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, C. J. Carmona, P. González, and M. J. del Jesus, "Fepds: A proposal for the extraction of fuzzy emerging patterns in data streams," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 12, pp. 3193–3203, 2020.

[2] J. Gama, *Knowledge discovery from data streams*. CRC Press, 2010.

[3] I. Khamassi and M. Sayed Mouchaweh, "Drift detection and monitoring in non-stationary environments," in *2014 IEEE Conference on Evolving and Adaptive Intelligent Systems, EAIS 2014, Linz, Austria, June 2-4, 2014*, 2014, pp. 1–6.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] I. Khamassi, M. Sayed Mouchaweh, M. Hammami, and K. Ghédira, "Discussion and review on evolving data streams and concept drift adapting," *Evolving Systems*, vol. 9, no. 1, pp. 1–23, 2018.

[10] 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.

[11] 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.