Modelos descriptivos basados en aprendizaje supervisado para el tratamiento de grandes volúmenes de datos y flujos continuos de datos

Doctorando: Ángel Miguel García Vico
Directores: Pedro González García y Cristóbal José Carmona del Jesus
Grupo de Investigación "Sistemas Inteligentes y Minería de Datos" (SiMiDat)
Departamento de Informática
Universidad de Jaén, Jaén, España
agvico@ujaen.es

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

I. Introducción

La minería de datos se ha dividido fundamentalmente en dos enfoques: predictivo, cuyo objetivo es la predicción del valor de una variable de interés en nuevas instancias no vistas anteriormente, utilizando para ello aprendizaje supervisado; y descriptivo con el objetivo de encontrar y definir relaciones interesantes en los datos utilizando para ello aprendizaje no supervisado. No obstante, a lo largo de la literatura se han ido desarrollando técnicas que se encuentran a medio camino entre ambos enfoques, agrupadas en el marco denominado "descubrimiento de reglas descriptivas basadas en aprendizaje supervisado" (SDRD) [1], [2] cuyo propósito es la descripción de conocimiento relevante sobre la variable de interés en un conjunto de datos. Dentro del SDRD, las técnicas más destacadas son el descubrimiento de subgrupos (SD) [3], [4], la minería de patrones emergentes (EPM) [5], [6] y la minería de conjuntos de contraste (CSM) [7].

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. En concreto, el objetivo de SD se define como la extracción de reglas cuya distribución estadística sea inusual respecto a una clase de interés. En CSM el objetivo principal es la búsqueda de reglas que definen conjuntos de una población con amplias diferencias de soporte entre grupos del conjuntos de datos. Por último, EPM tiene como objetivo la extracción de reglas respecto a una variable objetivo cuyo soporte sea muy alto en la clase de interés o muy bajo o nulo para el resto, con el objetivo de buscar tendencias emergentes a lo largo del tiempo.

Actualmente vivimos en la era de la información. El desarrollo en las tecnologías de la información y la comunicación ha permitido un aumento exponencial en la cantidad dispositivos generadores de datos, debido principalmente al abaratamiento de los sistemas de almacenamiento y sensores generadores de datos [8]. Toda esta cantidad de datos contiene conocimiento muy relevante para las empresas para poder mejorar sus servicios [9]. Esto ha propiciado en los últimos años el desarrollo de técnicas de extracción de conocimiento en estos enormes volúmenes de información heterogénea, comúnmente conocido como *Big Data*. El *Big Data* no solo se ve influenciado por el volumen de los datos, sino también por su expansión en otras dimensiones como la variedad y la velocidad [10]. Para hacer frente a este tipo de problemas, se han diseñado diferentes herramientas [11], entre la que destaca el sistema de procesamiento *MapReduce*, así como su implementación de código abierto *Hadoop* [12] o *Spark* [13] para algoritmos iterativos, los cuales son un sistema de computación distribuida basado en dos funciones principales: *Map* y *Reduce* que deben ser diseñadas por los usuarios. A lo largo de la literatura científica, se han desarrollado un amplio conjunto de técnicas para el tratamiento de Big Data desde una gran variedad de enfoques.

El desarrollo de dispositivos de generación y transmisión de datos de forma continua como redes de sensores, *smartphones*, dispositivos *wearables*, sistemas de vigilancia, aplicaciones web, banca online, proveedores de energía, etc. [14], han incrementado enormemente la cantidad de flujos continuos de datos existentes. En estos ámbitos, los datos muy antiguos son completamente irrelevantes e incluso contraproducentes en un análisis. En este caso, un análisis continuo de la información conforme los datos lleguen al sistema es más interesante que un análisis tradicional con un gran conjunto de datos históricos. Un flujo de datos se define como un conjunto de instancias potencialmente infinito que llega al sistema a lo largo del tiempo a una velocidad que puede ser variable [15]. Para la extracción de conocimiento en flujos de datos hay que tener en cuenta varios factores que hacen que su extracción sea un desafío en comparación con la minería de datos tradicional. En concreto, se necesita de la actualización continua del modelo de aprendizaje, así como de estrategias para desechar información antigua debido a que no se puede almacenar todo el flujo en memoria [16], [15]. Además, muchos sensores y

fuentes de datos poseen una tasa de refresco muy elevada (del orden de Khz) que implican además un aprendizaje lo más rápido posible [17].

II. HIPÓTESIS DE PARTIDA

Desde el punto de vista de SDRD, la extracción de conocimiento relativa a la descripción de los datos y, en particular, de la extracción de conocimiento relativa al fenómeno subyacente que los produce es de vital importancia en diferentes ámbitos profesionales como por ejemplo en medicina o bioinformática. En concreto, en los últimos años, SD ha recibido una especial atención dentro de la comunidad científica [18], [19], [20] con propuestas basadas en enfoques clásicos [21], [22], [23] y en sistemas difusos evolutivos [24], [25], [26], [27]. Por su parte, en EPM se han desarrollado un amplio número de algoritmos [28], [29], [30], [31]. Sin embargo, a pesar de la relevancia de la tarea dentro de SDRD, la mayoría de ellos no han tenido en cuenta el balance interpretabilidad/calidad de los resultados obtenidos, ya que han sido utilizados únicamente para clasificación.

Tal y como se ha comentado anteriormente, este tipo de extracción de conocimiento es aún más interesante en entornos Big Data pues la gran cantidad de información existente y la heterogeneidad de las fuentes permite que estos problemas muy complejos sean más fáciles de comprender por parte de los expertos. Sin embargo, al inicio de este trabajo de investigación, solo se había presentado en la literatura especializada un algoritmo para SD enfocado en la extracción de conocimiento en entornos Big Data: el algoritmo MEFASD-BD [32] y no existía ningún método para EPM en este ámbito.

Por otro lado, este conocimiento es muy importante en minería de flujo de datos para una rápida determinación de las causas que producen el flujo y así poder actuar en consecuencia. Sin embargo, aún no se han planteado dentro de la comunidad investigadora propuestas que aprovechen el potencial que tiene SDRD, y en concreto EPM, para la extracción de tendencias emergentes.

Partiendo de estos antecedentes, surge la necesidad de desarrollar técnicas de EPM enfocadas en los objetivos de SDRD, apoyadas a su vez por las propuestas presentadas para SD, para la extracción no solo de patrones altamente discriminativos, sino que además sean fácilmente interpretables, precisos y generales, y con una distribución estadísticamente inusual con respecto a la clase analizada. Asimismo, se hace indispensable que a su vez sean fácilmente escalables para poder ser utilizadas en ámbitos Big Data y/o utilizables dentro de ámbitos de minería de flujos de datos con el fin de obtener conocimiento fácilmente comprensible por los expertos tanto en conjuntos de datos pequeños, como en conjuntos de datos de mayor tamaño y en flujos continuos de datos.

Tras todos estos antecedentes, se establecen las siguientes hipótesis de partida:

- Las técnicas encontradas en la literatura para EPM están claramente enfocadas a predicción, por lo que se pierden muchas de sus capacidades descriptivas. Asimismo, no son lo suficientemente escalables para abordar problemas de gran dimensionalidad y muchas de ellas por su diseño son incapaces de afrontar la tarea de minería de flujo de datos.
- Con un diseño adecuado, se pueden obtener métodos de extracción de patrones emergentes que permitan la extracción de conocimiento discriminativo, interpretable, preciso y general sobre una distribución de ejemplos estadísticamente inusual respecto a una clase objetivo.
- Con un diseño escalable, se puede mantener un tiempo de procesamiento estable frente al aumento del tamaño del conjunto de datos o a la llegada continua de datos al sistema.
- La extracción de este tipo de conocimiento en bases de datos con alta dimensionalidad o en minería de flujo de datos permitirá cubrir las necesidades de conocimiento que demandan los expertos en este tipo de entornos.
- Los problemas que no eran abordables o impracticables en el pasado, debido a un gran volumen de datos o a una llegada continua de los mismos podrán ser procesados gracias a las nuevas técnicas distribuidas y de minería de flujo de datos.

III. OBJETIVOS

En base a las hipótesis de partida iniciales, se plantea como objetivo principal de esta tesis el desarrollo de algoritmos de SDRD enfocados a la extracción de conocimiento fácilmente comprensible por el experto dentro de problemas de procesamiento de datos de gran magnitud (*Big Data*) o en problemas de minería de flujos de datos. Este objetivo principal se desglosa en los siguientes subobjetivos:

1) Estudio e identificación de los principales enfoques utilizados en SDRD para problemas complejos.
2) Desarrollo de nuevas propuestas algorítmicas para la extracción de modelos SDRD, principalmente basados en patrones emergentes, capaces de extraer conocimiento altamente representativo del conjunto de datos analizado y fácilmente comprensibles por parte de los expertos. Este objetivo se subdivide en:
   a) Desarrollo de propuestas multiobjetivo para la extracción de patrones emergentes.
   b) Análisis y desarrollo de métodos de filtrado y postprocesamiento.
3) Adaptación de las propuestas anteriores para la extracción de este conocimiento en entornos Big Data bajo el paradigma MapReduce, que se divide a su vez en:
   a) Estudio de los principales enfoques y estrategias utilizadas en sistemas difusos evolutivos para Big Data.
   b) Desarollo de propuestas multiobjetivo para extracción de patrones emergentes en entornos Big Data.
   c) Análisis del equilibrio entre el tiempo y calidad de los resultados obtenidos.
4) Adaptación de las propuestas para la extracción de conocimiento altamente descriptivo en entornos de

minería de flujos de datos, que a su vez se divide en los siguientes subobjetivos:

a) Estudio de los principales enfoques utilizados en minería de flujo de datos.
b) Diseño de propuestas multiobjetivo para la extracción de patrones emergentes en flujos continuos de datos.
c) Estudio de métodos de visualización de los patrones emergentes extraídos a lo largo del tiempo.

5) Aplicación de las propuestas desarrolladas a datos reales con el objetivo de transferir los resultados de investigación al sector productivo y la sociedad en general. En concreto, se pretende aplicar los métodos desarrollados en los siguientes campos:
   a) Energías renovables.
   b) Gestión de flotas.

IV. Metodología

El desarrollo de esta tesis implica una metodología de trabajo teórico-práctica pues es necesario por un lado el desarrollo de nuevas metodologías de manera teórica y por otro lado la implementación de las mismas para confirmar su validez. Para la consecución de los objetivos 1,2,3 y 4 así como sus subobjetivos se seguirá el método científico tradicional, el cual se describe a continuación:

1) Formulación de hipótesis. Se plantean las hipótesis iniciales de los objetivos a llevar a cabo. En este punto, se diseñarán nuevas propuestas algorítmicas para la extracción de modelos SDRD altamente representativos del conjunto de datos a analizar y fácilmente interpretables por el experto en entornos Big Data y en flujos continuos de datos.
2) Recogida de observaciones. Se obtendrán los resultados como consecuencia de la aplicación de los algoritmos desarrollados. En concreto, se utilizarán bases de datos ampliamente conocidas por la comunidad científica del repositorio *UCI Knowledge Discovery in Databases* [33] mediante la utilización de validación cruzada estratificada en conjuntos de datos de gran tamaño. Para flujos continuos de datos, estos se analizarán utilizando la herramienta de análisis de flujos de datos MOA [34], la cual posee generadores de flujos de datos ampliamente conocidos por la comunidad.
3) Contraste de hipótesis. Se compararán los resultados obtenidos por los algoritmos desarrollados con el objetivo de analizar su calidad con otras propuestas dentro de la temática. En concreto, para las comparaciones con los algoritmos del estado del arte se trabajará con un *framework* disponible públicamente en el GitHub del grupo de investigación SiMiDat[1] que contiene los algoritmos más relevantes de EPM hasta la fecha. Asimismo, todos los resultados obtenidos serán validados mediante test estadísticos no paramétricos [35].
4) Demostración o refutación de la hipótesis. La hipótesis se acepta, o se rechaza y se modifica, en función de los resultados obtenidos en las pruebas realizadas.
5) Tesis. Extracción, redacción y aceptación de las conclusiones obtenidas durante el proceso.

Para finalizar, el objetivo 5 pretende analizar todos los modelos desarrollados para la extracción de patrones emergentes en distintos problemas de estudio en casos reales. La idea es la extracción de modelos de reglas interpretables bajo aprendizaje supervisado con modelos difusos evolutivos que puedan proporcionar a los expertos de las temáticas a analizar conocimiento de interés.

V. Plan de trabajo

Para poder desarrollar la tesis dentro del período máximo de tres años del programa de doctorado actual, se establece un plan de trabajo en donde se especifica una estimación temporal para la consecución de cada unos de los objetivos propuestos en la sección anterior, así como los diferentes hitos que marcarán el cumplimiento de los mismos. En particular, para cada uno de los subobjetivos propuestos anteriormente se define la siguiente planificación temporal:

- Objetivo 1: Estudio e identificación de los principales enfoques utilizados en SDRD para problemas complejos.
  - Programación: M1-M6.
  - Descripción: Realización de un estudio del estado del arte sobre los algoritmos SDRD, en especial aquellos que usen sistemas difusos evolutivos.
  - Resultados esperados: Informe donde se resuman los enfoques más relevantes de la literatura así como posibles vías de desarrollo de nuevas propuestas para la extracción de modelos SDRD.
- Objetivo 2.a: Desarrollo de propuestas multiobjetivo para la extracción de patrones emergentes.
  - Programación: M6-M15.
  - Descripción: Desarrollo de nuevas propuestas multiobjetivo basadas en sistemas difusos evolutivos para la extracción de patrones emergentes descriptivos.
  - Resultados esperados: Métodos para la extracción de patrones emergentes capaces de mejorar los resultados obtenidos respecto a descripción y carácter diferenciador por los algoritmos presentes en la literatura.
- Objetivo 2.b: Análisis y desarrollo de métodos de filtrado y post-procesamiento.
  - Programación: M10-M15.
  - Descripción: Análisis y estudio de diferentes mecanismos que puedan incorporarse tanto en el proceso evolutivo como en una etapa final para optimizar diversos factores de los métodos evolutivos.
  - Resultados esperados: Obtención de nuevos mecanismos que mejoren los resultados de los algoritmos ya desarrollados.
- Objetivo 3.a: Estudio de los principales enfoques y estrategias utilizadas en sistemas evolutivos para extracción de patrones emergentes en entornos Big Data.

[1] https:/github.com/simidat

  - Programación: M12-M18.
  - Descripción: Estudio de la bibliografía especializada en Big Data para el análisis de las estrategias para abordarlo eficientemente.
  - Resultados esperados: Obtención de un informe donde se describen las estrategias más relevantes y vías a desarrollar en SDRD.
- Objetivo 3.b: Desarrollo de propuestas multiobjetivo para extracción de patrones emergentes en entornos Big Data.
  - Programación: M15-M24.
  - Descripción: Adaptación de los modelos desarrollados para entornos Big Data.
  - Resultados esperados: Desarrollo de propuestas evolutivas de extracción de patrones emergentes en entornos Big Data capaces de extraer conocimiento en este ámbito.
- Objetivo 3.c: Análisis del equilibrio entre el tiempo y la calidad de los resultados obtenidos.
  - Programación: M20-M24.
  - Descripción: Estudio de la escalabilidad de los métodos implementados para Big Data.
  - Resultados esperados: Obtención de conclusiones que permitan mejorar las implementaciones desarrolladas anteriormente.
- Objetivo 4.a: Estudio de los principales enfoques utilizados en minería de flujo de datos.
  - Programación: M24-M30.
  - Descripción: Estudio de la bibliografía especializada en minería de flujo de datos para el análisis de las estrategias llevadas a cabo para la extracción de conocimiento en este ámbito.
  - Resultados esperados: Obtención de un informe donde se describen las estrategias más relevantes y vías a desarrollar en SDRD.
- Objetivo 4.b: Diseño de propuestas multiobjetivo para la extracción de patrones emergentes en flujos continuos de datos.
  - Programación: M24-M33.
  - Descripción: Adaptación de las propuestas desarrolladas anteriormente a minería de flujo de datos.
  - Resultados esperados: Desarrollo de propuestas evolutivas de extracción de patrones emergentes en entornos de minería de flujos de datos.
- Objetivo 4.c: Estudio de métodos de visualización de los patrones emergentes a lo largo del tiempo.
  - Programación: M28-M33.
  - Descripción: Creación de propuestas para una visualización simple de los patrones emergentes extraídos en entornos de minería de flujo de datos.
  - Resultados esperados: Métodos de visualización que permitan mejorar el análisis a los expertos.
- Objetivo 5.a: Análisis a datos sobre energías renovables.
  - Programación: M30-M36.
  - Descripción: Análisis de datos reales sobre sistemas fotovoltaicos de concentración de la Universidad de Jaén que registran datos en tiempo real.
  - Resultados esperados: Descubrimiento de posibles anomalías en los equipamientos mediante patrones emergentes en flujos continuos de datos.
- Objetivo 5.b: Análisis a datos de gestión que permitan optimizar el funcionamiento del sistema.
  - Programación: M30-M36.
  - Descripción: Análisis de datos reales de la red de taxis de la ciudad de Nueva York.
  - Resultados esperados: Devolver patrones emergentes a lo largo del tiempo sobre el comportamiento de los usuarios que utilizan el taxi.

Esta planificación temporal se puede observar también en la Figura 1 donde se presenta un cronograma a fin de facilitar la comprensión de la duración de los diferentes objetivos de la tesis a lo largo del tiempo.

VI. RELEVANCIA

A día de hoy existen tecnologías de aprendizaje automático como por ejemplo el aprendizaje profundo (Deep Learning), entre otros, que permiten la extracción de modelos muy precisos en una gran variedad de problemas. Sin embargo, la gran desventaja de estos modelos es que es prácticamente imposible extraer una conclusión clara de las razones por las que se ha realizado dicha predicción. En este sentido, en ámbitos como la medicina o algunos campos de la industria, economía, entre otros, es de vital importancia la extracción del tipo de conocimiento obtenido mediante técnicas SDRD. En este sentido, la aportación original de esta tesis es la obtención de este tipo de conocimiento simple, fiable y fácilmente interpretable por el experto en conjuntos de datos donde hasta la fecha era imprácticable su extracción como en Big Data o en flujos de datos.

A nivel científico, fruto del trabajo desarrollado hasta el momento para esta tesis, se han publicado varias publicaciones en revistas internacionales, entre los que se destacan:

1) A. M. García-Vico, F. Charte, P. González, C. J. Carmona, and M. J. del Jesus, "Subgroup discovery with evolutionary fuzzy systems in r: The sdefsr package," *The R Journal*, vol. 8, no. 2, pp. 307–323, 2016 (Ranking: 52/124 (Q2) en *Statistics & probability*, JCR 2016)
   - En este trabajo se presenta el paquete SDEFSR para el software estadístico R. Dicho paquete contiene los principales algoritmos de descubrimiento de subgrupos basados en sistemas difusos evolutivos.
2) J. Luengo, A. M. García-Vico, M. D. Pérez-Godoy, and C. J. Carmona, "The influence of noise on the evolutionary fuzzy systems for subgroup discovery," *Soft Computing*, vol. 20, pp. 4313–4330, 2016 (Ranking: 33/105 (Q2) en *Computer Science Interdisciplanary Applications*, JCR 2016)
   - En este trabajo se presenta un estudio sobre la influencia del ruido en los algoritmos de descubrim-

	Primer año												Segundo año												Tercer año											
Objetivo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Objetivo 1: Estudio e identificación de los principales enfoques utilizados en	x	x	x	x	x	x																														
Objetivo 2.a: Desarrollo de propuestas multiobjetivo para la extracción de						x	x	x	x	x	x	x	x	x	x																					
Objetivo 2.b: Análisis y desarrollo de métodos de filtrado y post-										x	x	x	x	x	x																					
Objetivo 3.a: Estudio de los principales enfoques y estrategias utilizadas en												x	x	x	x	x	x	x																		
Objetivo 3.b: Desarrollo de propuestas multiobjetivos para extracción de															x	x	x	x	x	x	x	x	x	x												
Objetivo 3.c: Análisis del equilibrio entre el tiempo y la calidad de los																				x	x	x	x	x												
Objetivo 4.a: Estudio de los principales enfoques utilizados en minería de flujo																								x	x	x	x	x	x	x						
Objetivo 4.b: Diseño de propuestas multiobjetivo para la extracción de																								x	x	x	x	x	x	x	x	x	x			
Objetivo 4.c: Estudio de métodos de visualización de los patrones																												x	x	x	x	x	x			
Objetivo 5.a: Análisis a datos energías renovables.																														x	x	x	x	x	x	x
Objetivo 5.b: Análisis a datos de gestión.																														x	x	x	x	x	x	x

Fig. 1. Cronograma de la tesis.

iento de subgrupos basados en sistemas difusos evolutivos.

3) 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 (Ranking: 29/104 (Q2) en *Computer Science Theory & Methods*, JCR 2016)
   - Se presenta una revisión bibliográfica de EPM desde el punto de vista de SDRD, fruto del profundo estudio realizado sobre el problema.
4) 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 (Ranking: 4/133 (Q1) en *Computer Science Artificial Intelligence*, JCR 2016)
   - En este trabajo se presenta un algoritmo basado en un sistema difuso evolutivo para la extracción de patrones emergentes con un gran equilibrio entre la fiabilidad y la facilidad de comprensión del modelo.
5) ——, "A big data approach for extracting fuzzy emerging patterns," *Cognitive Computation*, Submitted
   - Se presenta un algoritmo escalable basado en un sistema difuso evolutivo para la extracción de patrones emergentes de gran calidad en Big Data.

Asimismo, se han presentado las siguientes comunicaciones en congresos internacionales:

1) 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*. Springer, 2016, pp. 1–8
   - En este trabajo se presenta una aplicación del algoritmo de minería de patrones emergentes EvAEP en donde se describir las características más relevantes de placas fotovoltaicas de concentración en función del rendimiento de las mismas.
2) A. M. García-Vico, P. González, C. J. Carmona, and M. J. del Jesus, "Impact of the type of rule in fuzzy emerging pattern mining on a big data approach," in *Proc. of the 2nd International Symposium on Fuzzy and Rough Sets*, 2017, pp. 1–10
   - En este trabajo se presenta un estudio donde se determina la representación del conocimiento en sistemas difusos evolutivos que mejor se adecua a los objetivos de EPM en entornos Big Data.
3) A. M. García-Vico, P. González, M. J. del Jesus, and C. J. Carmona, "A first approach to handle emergining patterns mining on big data problems: The evaefp-spark algorithm," in *IEEE International Conference on Fuzzy Systems*, 2017, pp. 1–6
   - En este trabajo se presenta una versión escalable del algoritmo EvAEP para la extracción de patrones emergentes mediante un sistema difuso evolutivo en entornos Big Data.

Actualmente se está trabajando en el desarrollo de un método de extracción de patrones emergentes en entornos de minería de flujo de datos. Además, el trabajo futuro se centrará en el desarrollo de métodos que mejoren tanto en calidad

como en tiempo de procesamiento a las propuestas presentadas anteriormente. Asimismo, se pretende abrir nuevas líneas de investigación en el campo de SDRD enfocado a regresión.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] A. Fernández, S. Río, V. López, A. Bawakid, M. del Jesus, J. Benítez, and F. Herrera, "Big Data with Cloud Computing: An Insight on the Computing Environment, MapReduce and Programming Frameworks," *WIREs Data Mining and Knowledge Discovery*, vol. 5, no. 4, pp. 380–409, 2014.

[9] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[10] T. Kraska, "Finding the Needle in the Big Data Systems Haystack," *IEEE Internet Computing*, vol. 17, no. 1, pp. 84–86, 2013.

[11] A. Fernández, C. J. Carmona, M. J. del Jesus, and F. Herrera, "A View on Fuzzy Systems for Big Data: Progress and Opportunities," *International Journal of Computational Intelligence Systems*, vol. 9, no. 1, pp. 69–80, 2016.

[12] T. White, *Hadoop, The Definitive Guide*. O'Reilly Media, Inc., 2012.

[13] M. Zaharia, M. Chowdhury, M. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, ser. HotCloud'10, 2010, pp. 10–10.

[14] I. Žliobaitė, M. Pechenizkiy, and J. Gama, "An overview of concept drift applications," in *Big Data Analysis: New Algorithms for a New Society*. Springer, 2016, pp. 91–114.

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

[16] A. Bifet, "Adaptive learning and mining for data streams and frequent patterns," Ph.D. dissertation, Universitat Politècnica de Catalunya, 2009.

[17] D. Han, C. Giraud-Carrier, and S. Li, "Efficient mining of high-speed uncertain data streams," *Applied Intelligence*, vol. 43, no. 4, pp. 773–785, 2015.

[18] F. Herrera, C. J. Carmona, P. González, and M. J. del Jesus, "An overview on Subgroup Discovery: Foundations and Applications," *Knowledge and Information Systems*, vol. 29, no. 3, pp. 495–525, 2011.

[19] C. J. Carmona, P. González, M. J. del Jesus, and F. Herrera, "Overview on evolutionary subgroup discovery: analysis of the suitability and potential of the search performed by evolutionary algorithms," *WIREs Data Mining and Knowledge Discovery*, vol. 4, no. 2, pp. 87–103, 2014.

[20] M. Atzmueller, "Subgroup discovery," *WIREs Data Mining and Knowledge Discovery*, vol. 5, pp. 35–49, 2015.

[21] M. Atzmueller and F. Puppe, "SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery," in *Proc. of the 17th European Conference on Machine Learning and 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, ser. LNCS, vol. 4213. Springer, 2006, pp. 6–17.

[22] B. Kavsek and N. Lavrac, "APRIORI-SD: Adapting association rule learning to subgroup discovery," *Applied Artificial Intelligence*, vol. 20, pp. 543–583, 2006.

[23] H. Grosskreutz, S. Rueping, and S. Wrobel, "Tight optimistic estimates for fast subgroup discovery," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2008, pp. 440–456.

[24] M. J. del Jesus, P. González, F. Herrera, and M. Mesonero, "Evolutionary Fuzzy Rule Induction Process for Subgroup Discovery: A case study in marketing," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 4, pp. 578–592, 2007.

[25] M. J. del Jesus, P. González, and F. Herrera, *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models*. Springer, 2007, vol. 220, ch. Subgroup Discovery with Linguistic Rules, pp. 411–430.

[26] C. J. Carmona, P. González, M. J. del Jesus, and F. Herrera, "NMEEF-SD: Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 5, pp. 958–970, 2010.

[27] 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*, vol. 298, pp. 180–197, 2015.

[28] J. Bailey, T. Manoukian, and K. Ramamohanarao, "Fast Algorithms for Mining Emerging Patterns," in *Principles of Data Mining and Knowledge Discovery*. Springer, 2002, vol. 2431, pp. 187–208.

[29] J. Y. Li, G. Z. Dong, K. Ramamohanarao, and L. Wong, "DeEPs: A New Instance-Based Lazy Discovery and Classification System," *Machine Learning*, vol. 54, no. 2, pp. 99–124, 2004.

[30] H. Fan and K. Ramamohanarao, "Fast discovery and the generalization of strong jumping emerging patterns for building compact and accurate classifiers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 6, pp. 721–737, 2006.

[31] M. García-Borroto, J. Martínez-Trinidad, and J. Carrasco-Ochoa, "Fuzzy emerging patterns for classifying hard domains." *Knowledge and Information Systems*, vol. 28, no. 2, pp. 473–489, 2011.

[32] 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.

[33] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: `http://archive.ics.uci.edu/ml`

[34] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: massive online analysis," *Journal of Machine Learning Research*, vol. 11, pp. 1601–1604, 2010. [Online]. Available: `https://moa.cms.waikato.ac.nz/`

[35] S. García and F. Herrera, "An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons," *Journal of Machine Learning Research*, vol. 9, pp. 2677–2694, 2008.

[36] A. M. García-Vico, F. Charte, P. González, C. J. Carmona, and M. J. del Jesus, "Subgroup discovery with evolutionary fuzzy systems in r: The sdefsr package," *The R Journal*, vol. 8, no. 2, pp. 307–323, 2016.

[37] J. Luengo, A. M. García-Vico, M. D. Pérez-Godoy, and C. J. Carmona, "The influence of noise on the evolutionary fuzzy systems for subgroup discovery," *Soft Computing*, vol. 20, pp. 4313–4330, 2016.

[38] 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.

[39] ——, "A big data approach for extracting fuzzy emerging patterns," *Cognitive Computation*, Submitted.

[40] 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*. Springer, 2016, pp. 1–8.

[41] A. M. García-Vico, P. González, C. J. Carmona, and M. J. del Jesus, "Impact of the type of rule in fuzzy emerging pattern mining on a big data approach," in *Proc. of the 2nd International Symposium on Fuzzy and Rough Sets*, 2017, pp. 1–10.

[42] A. M. García-Vico, P. González, M. J. del Jesus, and C. J. Carmona, "A first approach to handle emergining patterns mining on big data problems: The evaefp-spark algorithm," in *IEEE International Conference on Fuzzy Systems*, 2017, pp. 1–6.