Universidad de Jaén

Escuela Politécnica Superior de Jaén

Modelos descriptivos basados en aprendizaje supervisado para el tratamiento de big data y flujos continuos de datos.

Memoria presentada por

Ángel Miguel García Vico

Para optar al Grado de Doctor en Informática
con mención internacional

Fdo.: *Ángel Miguel García Vico.*

Tesis Doctoral parcialmente subvencionada por el Ministerio de Ciencia, Innovación y Universidades a través de la convocatoria *Ayudas para Contratos Predoctorales 2016* (referencia BES-2016-077738), y parcialmente subvencionada por el Ministerio de Economía y Empresa bajo el proyecto con referencia TIN2015-68454-R, Fondos FEDER.

D. Pedro González García y **D. Cristóbal José Carmona del Jesus**, Profesores Titulares de Universidad del Departamento de Informática de la Universidad de Jaén

INFORMAN

Que la memoria titulada "*Modelos descriptivos basados en aprendizaje supervisado para el tratamiento de big data y flujos continuos de datos*" ha sido realizada por **D. Ángel Miguel García Vico** bajo su supervisión dentro del Programa de Doctorado en Tecnologías de la Información y la Comunicación para optar al grado de doctor.

Para su evaluación, esta memoria se presenta como un conjunto de trabajos publicados, acogiéndose y ajustándose a lo establecido en el punto 2 del artículo 25 del *Reglamento de los Estudios de Doctorado de la Universidad de Jaén*, aprobado en febrero de 2012 y modificado en febrero de 2019.

Jaén, a 26 de febrero de 2020

Fdo. Pedro González García
Tutor y Director de la Tesis

Fdo. Cristóbal José Carmona del Jesus
Director de la Tesis

A mis padres, por su enorme esfuerzo y sacrificio.
Os quiero.

Agradecimientos

"Bienvenido a la montaña rusa de la investigación". Estas fueron las primeras palabras de Cristóbal cuando empecé a investigar. No le faltaba razón. Tras estos años, el camino no ha sido fácil y ha estado lleno de obstáculos: jornadas de trabajo que se alargan hasta el infinito, experimentaciones fallidas, rechazo de artículos, etc. Sin embargo, tras haber llegado hasta aquí, y con cierta perspectiva, tengo que decir que ha merecido la pena. Por esta razón, me gustaría dedicarles unas líneas a las personas que me han apoyado y me han permitido llegar hasta este punto, pues sin ellos hubiera sido imposible.

En primer lugar, a mis padres. A pesar de no tener ningún tipo de estudios y de que encuentren muy complejo el mundo académico, me han enseñado una filosofía de vida y una forma de ser que no se aprende en la escuela. Ese esfuerzo que habéis realizado para que dé siempre lo máximo de mí ha permitido que hoy sea quien soy y esté donde estoy. Hoy, os podéis sentir orgullosos del fruto que habéis ido cultivando durante tanto tiempo.

En segundo lugar a mis directores de tesis. Sin Cristóbal y sin Pedro este trabajo no hubiera sido posible. Sus comentarios, ideas y apoyo han permitido que se desarrolle un trabajo de excepcional calidad. Todo ello acompañado de la capacidad necesaria para hacerme crecer como investigador, la cual me ha permitido afrontar con cada vez mayor capacidad la difícil tarea de investigar.

También agradecer al grupo de investigación SiMiDat por hacerme sentir un compañero más de trabajo y por el todo el apoyo aportado. Me gustaría agradecer en especial a María José. Sin ella no hubiera empezado a investigar y no hubiera estado aquí. Por esto, y por todo el apoyo y oportunidades que me has dado desde mis inicios te estoy enormemente agradecido.

Gracias a Milton García-Borroto y a Diana Martín por el apoyo proporcionado a la hora de realizar la revisión de la literatura sobre EPM. Sin vosotros el trabajo realizado hubiera sido extremadamente complejo. También agradecer a David Elizondo por la oportunidad de realizar una estancia de investigación en DeMontfort University, Leicester, de donde ha salido un gran trabajo, junto a una gran experiencia vital. *Thanks to Huseyin Seker for the invitation to the Northumbria University in order to perform a research stay. It was a great experience at Newcastle, where also an amazing work has been done.*

A todos y cada uno de mis compañeros de los seminarios 154, 102 y 103 por hacer más amenas las mañanas y alguna que otra tarde. Por las discusiones, risas, cafés y muchas cosas más. Sois impresionantes y os deseo todo lo mejor.

Finalmente, y no por ello menos importante, a mi hermano Sergio y mis grandes amigos: Beatriz, Jesús Tudela, Juan, Paqui, Ana Belén, Paco, Cristina, Cristian, Chema, Almudena, Vanessa e Irene. Gracias a todos por el apoyo y ánimo en los peores momentos y por todas las alegrías que hemos pasado y que quedan por pasar. Sois inigualables.

A todos vosotros.
Muchas gracias.

Resumen

Actualmente la relevancia del análisis de datos está creciendo exponencialmente, así como el volumen y velocidad de generación de los mismos, creando la necesidad de resolver problemas cada vez más complejos. Dentro de la Ciencia de datos existe un conjunto de técnicas a medio camino entre la predicción y la descripción cuyo objetivo es la explicación de los datos con respecto a una variable de interés, utilizando patrones o reglas para ello. A este grupo de tareas se le denomina descubrimiento de reglas descriptivas basadas en aprendizaje supervisado.

Dentro de este campo, las tareas más relevantes son el descubrimiento de subgrupos y la minería de patrones emergentes, cuyos objetivos son la descripción de subconjuntos interesantes de la población y la explicación de las características diferenciadoras entre clases, respectivamente. La principal característica de estas tareas en problemas complejos es su capacidad de simplificarlos y resumirlos de manera fiable y altamente comprensible para los expertos. Los problemas complejos que actualmente generan mayor interés están relacionados con el análisis *big data* y de flujos continuos de datos, entre otros. En este sentido, el tratamiento *big data* ya ha sido analizado por los investigadores desde diversas perspectivas para la tarea de descubrimiento de subgrupos. Sin embargo, para minería de patrones emergentes este análisis aún no ha sido realizado. También se destaca que, en ambas tareas, existe una falta de análisis y desarrollo de métodos enfocados a la extracción de conocimiento en flujos continuos de datos y otros tipos de problemas complejos.

En esta tesis se analizan en profundidad las tareas de descubrimiento de subgrupos y minería de patrones emergentes enfocadas a la resolución de problemas complejos, como *big data* y flujos continuos de datos, entre otros. Se proponen diferentes métodos y herramientas que permiten la extracción de conocimiento

descriptivo en este tipo de entornos. Además, se destacan diferentes problemas abiertos en este área. En particular, para descubrimiento de subgrupos se presenta un análisis de la influencia de ruido en los datos en los principales sistemas difusos evolutivos desarrollados; un paquete software para la plataforma R con los principales algoritmos basados en sistemas difusos evolutivos; y un análisis del comportamiento de los principales enfoques a problemas multi-instancia, un problema complejo en auge, mediante la realización de adaptaciones de los mismos. Con respecto a la minería de patrones emergentes, se presenta una revisión de los principales enfoques desarrollados en la tarea desde el punto de vista descriptivo y tres propuestas basadas en sistemas difusos evolutivos: una enfocada a mejorar la calidad del conocimiento extraído desde el punto de vista descriptivo; otra enfocada a realizar esta extracción en el ámbito *big data* y un último método enfocado al contexto de la minería de flujo de datos.

Los resultados obtenidos muestran que los métodos propuestos permiten obtener conocimiento de calidad capaz de ayudar a la toma de decisiones por parte de los expertos en problemas complejos.

Palabras Clave: Inteligencia artificial; Metaheurísticas; Minería de patrones emergentes; Descubrimiento de subgrupos; Big data; Minería de flujos de datos; Lógica difusa.

Abstract

Currently, the relevance of data analysis is growing exponentially, as well as its volume and generation speed. This creates the necessity of solving increasingly complex problems. Within data science there is a set of techniques halfway between prediction and description whose objective is to explain the data with respect to a variable of interest, by means of patterns or rules. This group of tasks is called supervised descriptive rule discovery.

Within this field the most relevant tasks are subgroup discovery and emerging pattern mining. Their objectives are the description of interesting subsets of the population and the explanation of the differentiating characteristics between classes, respectively. The main characteristic of these tasks in complex problems is their ability to simplify and summarize them in a reliable and highly understandable way for experts. The complex problems that currently generate most interest are related to the analysis of big data and data streams, among others. In this way, the treatment of big data has already been analysed by researchers from various perspectives of subgroup discovery. However, this analysis has not yet been performed on emerging pattern mining. It is also highlighted that in both tasks there is a lack of analysis and development of methods focused on knowledge extraction in data streams and other types of complex problems.

In this thesis the subgroup discovery and emerging pattern mining tasks for the resolution of complex problems, such as big data and data stream mining, among others, are analysed in depth. Different methods and tools are proposed in order to extract descriptive knowledge from these types of environments. In addition, different open problems in this area are highlighted. In particular, for subgroup discovery an analysis of the influence of data noise on the main evolutionary fuzzy systems developed is presented; a software package for the R platform with the main algorithms based on evolutionary fuzzy systems is pro-

posed; and an initial analysis of the behaviour of the main approaches adapted to multi-instance problems, a complex problem on the rise, is shown. With respect to emerging pattern mining, a review of the main approaches developed in the task from a descriptive point of view is presented, together with three developments based on evolutionary fuzzy systems: one focused on improving the quality of the extracted knowledge from a descriptive point of view; another focused on performing this extraction in the big data domain and a last method focused on the context of data stream mining.

The results obtained show that the proposed methods allow the extraction of quality knowledge capable of helping the decision-making process in complex problems.

Keywords: Artificial Intelligence; Metaheuristics; Emerging pattern mining; Subgroup discovery; Big data; Data stream mining; Fuzzy logic.

Índice general

Índice de figuras

Lista de Acrónimos

CRISP-DM *Cross-Industry Standard Process for Data Mining*

CSM *Contrast Set Mining*

DNF *Disjunctive Normal Form*

EPM *Emerging Pattern Mining*

FPR *False Positive Rate*

GR *Growth Rate*

IoT *Internet of Things*

IRL *Iterative Rule Learning*

JCR *Journal Citation Reports*

MIL *Multiple-Instance Learning*

RDD *Resilient Distributed Dataset*

SD *Subgroup Discovery*

SDRD *Supervised Descriptive Rule Discovery*

TPR *True Positive Rate*

WRAcc *Weighted Relative Accuracy*

Introducción

La invención de la máquina de vapor fue el comienzo de la primera revolución industrial, impulsando el desarrollo económico. Años más tarde, la cadena de montaje, la electricidad y el teléfono mejoraron el desarrollo de bienes y servicios en la segunda revolución industrial. En los años 50, aproximadamente, la invención de los semiconductores, que permitieron la transición de procesos analógicos a digitales, junto a la creación de Internet, marcaron el inicio de la tercera revolución industrial, la cual ha transformado el mundo que hoy conocemos. Toda revolución se caracteriza por influir en la vida y el comportamiento de los individuos y organizaciones. A día de hoy se estipula que la sociedad está inmersa en la cuarta revolución industrial, denominada la Era de la información. Tal y como se muestra en la Figura 1, entre otros avances científicos y tecnológicos, en esta nueva revolución industrial destaca el desarrollo exponencial de las Tecnologías de la información y la comunicación, que permiten la transferencia y almacenamiento de varios *exabytes*[1] de datos diariamente.

Al igual que el detonante de la primera revolución industrial fue la máquina de vapor, de la segunda la producción en masa por la electricidad, y de la tercera el desarrollo de Internet, en esta cuarta revolución destaca, entre otras áreas, la Inteligencia artificial alimentada con esta enorme cantidad de datos. La Ciencia de datos es el ámbito de conocimiento que engloba las habilidades asociadas al procesamiento de datos [1]. Es un ámbito multidisciplinar, representado en la Figura 2, en el que participan campos como las Ciencias de la computación, la Estadística y Matemáticas y el dominio específico del problema para extraer significado de los datos. Para la construcción de modelos son necesarias varias técnicas de cada una de las disciplinas anteriores, por ejemplo: algoritmia, modelado de incertidumbre, almacenamiento, reducción de datos, optimización, etc.

[1] 1 exabyte = 10^{18} bytes

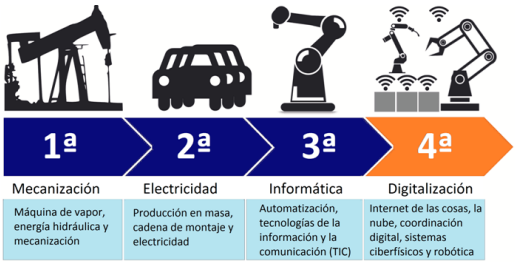


Figura 1: Principales hitos de las revoluciones industriales a lo largo de la historia. *Fuente: economipedia.com.*

[2]. La Ciencia de datos participa en todos los niveles y decisiones que se llevan a cabo en una organización: desde un nivel operacional, en donde únicamente se gestionan datos, hasta un nivel estratégico en donde se producen decisiones a largo plazo con relación a los datos analizados, mediante el empleo de técnicas enmarcadas dentro de la Minería de datos [3].

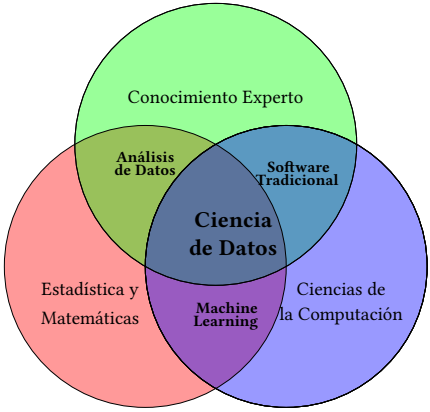


Figura 2: Disciplinas que componen la Ciencia de datos. *Fuente: elaboración propia.*

El proceso de Minería de datos se puede definir como el proceso no trivial de extracción de patrones y relaciones implícitas en los datos. Este conocimiento debe ser válido, novedoso, útil y comprensible para el experto para que el proce-

so sea exitoso. La aplicación industrial de este proceso se conoce, en inglés, como *Cross-Industry Standard Process for Data Mining* (CRISP-DM) y se divide en varias fases interactivas e iterativas aplicadas en un orden concreto, representadas en la Figura 3, que se detallan a continuación:

1. Comprensión del negocio. En esta etapa se entienden los objetivos de negocio y los requerimientos del proyecto. Asimismo, se define el problema de minería de datos a abordar.

2. Comprensión de los datos. En esta etapa se recopila el conjunto de datos inicial, en donde el experto en Ciencia de datos realiza un análisis exploratorio del problema, para poder abordarlo eficazmente.

3. Preparación de datos. El objetivo de esta fase es eliminar inconsistencias o ruido, reducir o aumentar el tamaño de la muestra, así como la aplicación de cualquier método que modifique la forma de los datos para homogeneizar y maximizar la ganancia de conocimiento en las fases posteriores del proceso.

4. Modelado. Este período del proceso de Ciencia de datos es el encargado de la extracción del conocimiento en sí. Se lleva a cabo mediante la creación de un modelo a partir de los datos recopilados y preprocesados anteriormente. Un modelo de conocimiento es una representación de patrones de comportamiento o relaciones entre los datos que se pueden usar, por ejemplo, para predecir, entender mejor los datos o explicar situaciones pasadas, entre otros.

5. Evaluación. Se realiza la presentación del conocimiento extraído al experto de una manera comprensible. Además, se evalúa la validez del mismo. En el caso de que no se satisfagan los requerimientos de calidad o los objetivos establecidos, estos resultados servirán de retroalimentación en una nueva iteración del proceso.

6. Despliegue. La fase final del proceso es la implantación de los modelos resultantes en la práctica. Además, se configura el proceso de Ciencia de datos para que se aplique de manera repetida o continua.

De todas las fases del proceso de Minería de datos, el modelado es la que mayor interés ha recibido por parte de la comunidad científica, debido a que esta fase es la encargada de la extracción de valor a partir de los datos. En concreto, su

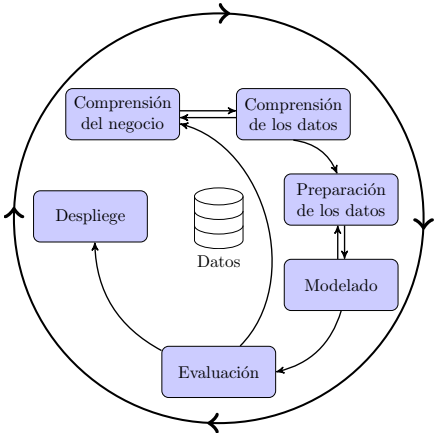


Figura 3: Fases del proceso CRISP-DM. *Fuente: elaboración propia a partir de https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining.*

.

cometido es la extracción de un modelo de conocimiento que represente patrones o relaciones relevantes en problemas definidos vagamente, sin una solución formal o cuyo resultado exacto conlleva un tiempo de cómputo excesivo.

Dependiendo de los objetivos establecidos en el problema a abordar, los paradigmas seguidos en Minería de datos siguen dos vertientes claramente diferenciadas:

- Predicción. Su propósito es anticipar el comportamiento de futuras entradas al sistema. Los algoritmos de Minería de datos dentro de este enfoque están basados en aprendizaje supervisado. Esto implica que los datos contienen una variable de interés que guía el proceso de extracción de patrones. Aquí se encuentran, entre otras tareas:

    - Clasificación [4]. El objetivo es predecir un valor de una variable de interés discreta. Estos valores, denominados clases o etiquetas, tienen un dominio conocido, como por ejemplo la existencia o no de una enfermedad en una serie de pacientes.
    - Regresión [4]. Esta tarea pretende predecir el valor de una variable de interés con un dominio continuo, como por ejemplo la cantidad de precipitación en un área determinada.

    - Análisis de series temporales [5]. Determina el comportamiento de una variable continua a lo largo del tiempo, como por ejemplo el precio de las acciones de bolsa.

- Descripción. La finalidad de este paradigma es la búsqueda de relaciones interesantes entre los datos que describan el fenómeno subyacente. Los algoritmos de Minería de datos dentro de este esquema están basados en aprendizaje no supervisado. Por lo tanto, no necesitan de una variable de interés. En este enfoque se encuentran tareas como, por ejemplo:

    - Agrupamiento [6]. Consiste en la realización de grupos de instancias de datos similares entre sí y diferentes a las del resto de grupos realizados.
    - Asociación [7]. Descubre relaciones de interés entre las diferentes variables de un conjunto de datos.
    - Detección de anomalías [8]. Identifica aquellas instancias que son inusuales o que difieren significativamente de la mayoría, pudiendo ser potenciales instancias resultado de contaminación, error o fraude.

A pesar de la clara diferenciación entre estos enfoques en función de los objetivos establecidos, no todas las necesidades de conocimiento se satisfacen con esta clasificación. A día de hoy, existe un conjunto de técnicas que se encuentran a medio camino entre el enfoque predictivo y descriptivo englobadas dentro del marco del descubrimiento de reglas descriptivas basadas en aprendizaje supervisado (en inglés, *Supervised Descriptive Rule Discovery* (SDRD)) [9]. El objetivo de este marco de tareas es la descripción de relaciones interesantes para los expertos supeditadas a una variable objetivo o clase. Por lo tanto, la finalidad de estas técnicas no se centra en la predicción del comportamiento de una nueva instancia en relación a una variable de interés, sino en la descripción de aquellas características o factores que provocan dicho comportamiento.

Un ejemplo ilustrativo de aplicación de un modelo SDRD podría ser la determinación de las circunstancias en las que un paciente puede sufrir un tipo concreto de cáncer, de modo que los investigadores sanitarios puedan acotar su investigación. En este ejemplo el objetivo no es predecir la enfermedad, sino describir aquellos factores de riesgo que pueden producirla.

Actualmente, las técnicas de Minería de datos encuadradas dentro de SDRD más destacadas son las siguientes:

- Minería de conjuntos de contraste (en inglés, *Contrast Set Mining* (CSM)) [10]. Tiene como finalidad la búsqueda de patrones cuyas distribuciones difieran significativamente entre diferentes grupos de objetos. Es importante destacar que estos grupos son exclusivos entre sí, por lo que una instancia únicamente puede pertenecer a uno de ellos. En concreto, un patrón será determinado como de contraste si la diferencia de soporte del patrón entre los diferentes grupos es mayor a un valor umbral establecido. Debido a esta característica, su aplicación es interesante en investigaciones de ciencias sociales, pues es muy habitual este tipo de análisis, aunque actualmente se está aplicando también a otros campos como analítica de negocio [11] o medicina [12], entre otros [13].

- Descubrimiento de subgrupos (en inglés, *Subgroup Discovery* (SD)) [14], [15]. El objetivo de esta tarea es la descripción de subconjuntos de instancias cuya distribución estadística difiere significativamente respecto al conjunto total de la población, dada una variable de interés. Como esta tarea se centra en la búsqueda de relaciones con características interesantes, no es necesario que estas sean completas, sino que suele ser suficiente con relaciones parciales, que se describen en forma de patrones individuales. En los últimos años, esta tarea ha recibido especial interés por parte de la comunidad científica debido a su éxito en diferentes ámbitos [16]. Una revisión completa de la tarea enfocada en algoritmos evolutivos se presenta en [17]. En [18] puede encontrarse una revisión de métodos exhaustivos. Finalmente, en [19] se presenta una revisión basada en una evaluación empírica de estos modelos.

- Minería de patrones emergentes (en inglés, *Emerging Pattern Mining* (EPM)) [20], [21]. La búsqueda se basa en la obtención de aquellos patrones que sean frecuentes para una clase y poco frecuentes para el resto. Los objetivos que se pueden alcanzar con este enfoque son la búsqueda de tendencias emergentes a lo largo del tiempo y la búsqueda de características discriminativas entre las diferentes clases de un problema. Debido a esta propiedad, EPM no solo puede utilizarse con carácter descriptivo, sino que también puede usarse de manera predictiva. De hecho, a pesar del claro enfoque descriptivo de la tarea, el empleo de patrones emergentes en la literatura ha estado principalmente enfocado a tareas de clasificación [22].

Para alcanzar estos requerimientos, además de ser precisos, es deseable que los modelos SDRD obtengan conocimiento que se pueda representar de un modo en el que se facilite su comprensión por parte de un ser humano, de modo que:

1. Los diseñadores y desarrolladores puedan analizar el modelo generado y entender su estructura.

2. Los usuarios finales sean capaces de utilizar el modelo como un sistema de ayuda a la decisión, explicando el fenómeno que subyace en los datos para así tomar mejores decisiones y permitiendo justificarlas más fácilmente.

Una manera simple de combinar ambos aspectos es mediante la utilización de conjuntos de patrones y/o reglas, las cuales se emplean en SDRD. No obstante, el principal hándicap es la extracción eficiente de dichos patrones. Para la mayoría de problemas actuales, la aplicación de estos métodos es inviable debido a las altas dimensiones que poseen los datos de entrada. En este contexto un procesamiento distribuido se vuelve obligatorio. Recientemente se han desarrollado diversas plataformas, herramientas y paradigmas para el procesamiento de datos masivos, facilitando al usuario final el desarrollo de métodos en entornos distribuidos. La principal ventaja de estos sistemas es la gestión transparente al usuario de los recursos y mecanismos necesarios para el correcto funcionamiento de un procedimiento dentro de un entorno distribuido.

De hecho, actualmente el análisis de datos masivos, conocido como *big data* [23], se aplica en prácticamente cualquier ámbito: educación [24], ciudades inteligentes [25], [26], seguridad [27], entre otros [28]-[30]. Esto se debe principalmente a la proliferación de herramientas capaces de manejar estos datos masivos de manera eficaz y eficiente. Entre estos, uno de los paradigmas de computación más populares es MapReduce [31]. Este se basa en un marco de programación funcional basado en el paradigma "divide y vencerás" para el procesamiento distribuido de *big data*.

Los modelos SDRD son especialmente interesantes en el ámbito *big data* ya que permiten obtener una descripción simplificada, fiable y altamente comprensible de los fenómenos que subyacen sobre la inmensa cantidad de información que se está procesando. No obstante, los desarrollos actuales para la extracción de modelos SDRD sufren con el crecimiento en la dimensión de los datos ya que la mayoría no están diseñados para un procesamiento distribuido. En consecuencia, su rendimiento decae significativamente o es imposible su aplicación. Por lo tanto, el desarrollo de métodos de extracción de modelos SDRD con un diseño distribuido es actualmente un reto importante en la comunidad científica.

Pero no sólo se requiere procesar grandes conjuntos de datos estáticos. Debido al avance en los últimos años en la tecnología, protocolos y aplicaciones empleados en el denominado Internet de las Cosas (en inglés, *Internet of Things* (IoT)) [32], entre otros, también es necesario procesar conjuntos dinámicos e infinitos que llegan constantemente al sistema a lo largo del tiempo. A este tipo de conjunto de datos dinámico se le conoce como flujo de datos (en inglés, *data stream*) [33]. Debido a su naturaleza, el procesamiento de flujos de datos impone una serie de restricciones añadidas a la minería de datos tradicional. Entre otras, destaca la necesidad de actualizar constantemente el modelo de aprendizaje conforme los datos llegan al sistema, así como importantes restricciones de memoria y tiempo de procesamiento de cada instancia.

En el ámbito de la minería de flujo de datos, los modelos SDRD son muy interesantes para el experto debido a sus características. Entre otras aplicaciones, estos modelos de aprendizaje pueden ser utilizados para proporcionar un resumen del estado actual del flujo de datos, describiendo aquellos aspectos más interesantes para el experto. También son útiles para la identificación de los elementos que han propiciado un cambio en los datos, o pueden utilizarse para la descripción de tendencias emergentes en los datos. Sin embargo, la mayoría de métodos SDRD desarrollados no tienen en cuenta todas las restricciones presentes en la minería de flujo de datos. No obstante, algunos de estos procedimientos se basan en métodos incrementales para administrar flujos estáticos y finitos, como por ejemplo en [34]-[36]. Pero la principal dificultad que se encuentra a la hora de realizar una adaptación directa de las técnicas de extracción de modelos SDRD se basa en el empleo de medidas de calidad y esquemas de aprendizaje que asumen la invariabilidad de las propiedades de los datos a lo largo del tiempo, así como la disponibilidad de la totalidad de los datos desde el inicio. Por lo tanto, se hace indispensable el desarrollo de modelos SDRD capaces de actualizarse a lo largo del tiempo, teniendo en cuenta la naturaleza cambiante de los flujos de datos.

Objetivos

Con estas premisas, se establece como hipótesis de partida que el uso de modelos SDRD para la extracción de conocimiento en problemas complejos ayuda a cubrir satisfactoriamente las necesidades de conocimiento que demandan los expertos en este tipo de entornos. Para ello, el objetivo principal asociado será

el análisis y desarrollo de diferentes modelos SDRD para entornos *big data* y flujos continuos de datos. Este objetivo principal se subdivide en los siguientes objetivos específicos:

- Estudio e identificación de los principales enfoques utilizados en SDRD para problemas complejos.
- Desarrollo de nuevas propuestas algorítmicas para la extracción de modelos SDRD capaces de obtener conocimiento altamente representativo del conjunto de datos analizado y fácilmente comprensibles por parte de los expertos.
- Análisis, diseño y desarrollo de propuestas para la extracción de conocimiento en entornos *big data* bajo el paradigma MapReduce.
- Análisis, diseño y desarrollo de propuestas para la extracción de conocimiento altamente descriptivo en entornos de flujos de datos.
- Aplicación de las propuestas a datos reales con el objetivo de transferir los resultados de investigación al sector productivo y a la sociedad en general.

Estructura de la memoria

En virtud con lo establecido en el artículo 25, punto 2 de la normativa vigente para los Estudios de Doctorado de la Universidad de Jaén, correspondiente al programa establecido en el RD. 99/2011, esta tesis doctoral se presenta como un compendio de artículos publicados por el doctorando. Por consiguiente, y tras el capítulo introductorio, esta memoria se estructura de la siguiente manera:

- En el Capítulo 1 se introducen los detalles de los principales conceptos que se abordan en esta tesis doctoral como SDRD, computación flexible, en donde se hace especial hincapié en los sistemas difusos evolutivos y sus componentes, y los problemas complejos dentro de la Ciencia de datos.
- El Capítulo 2 presenta una descripción general de las ideas propuestas en los trabajos realizados por el doctorando, destacando las características principales de las mismas, junto a una discusión de los resultados obtenidos.
- El Capítulo 3 expone las principales conclusiones extraídas, los trabajos futuros que se derivan del trabajo realizado en esta tesis, así como las publicaciones asociadas con el trabajo principal de investigación.

- Finalmente, en el Capítulo 4 se presentan las publicaciones indexadas en el *Journal Citation Reports* (JCR), fruto del trabajo realizado en esta tesis, las cuales conforman el núcleo de la investigación realizada.

1 Conceptos teóricos

En este capítulo se describen en detalle los conceptos principales que se utilizan a lo largo de esta memoria. En concreto, en el Apartado 1.1 se presentan los conceptos principales del marco de tareas SDRD. Después, en el Apartado 1.2 presenta el paradigma de la computación flexible, centrándose en la lógica difusa, los algoritmos evolutivos y la hibridación de ambos: los sistemas difusos evolutivos. A continuación, el Apartado 1.3 presenta los detalles de problemas complejos como aprendizaje multi-instancia en el Apartado 1.3.1, *big data* en el Apartado 1.3.2 y las características más destacadas de los flujos de datos en el Apartado 1.3.3.

1.1. Extracción de reglas descriptivas mediante aprendizaje supervisado

Tradicionalmente, la minería de datos ha tenido dos paradigmas de aprendizaje claramente diferenciados: supervisado y no supervisado. En general, los algoritmos basados en aprendizaje supervisado tienen un enfoque más predictivo, mientras que los enfoques no supervisados tienden a un enfoque más descriptivo de los datos de entrada. Sin embargo, estos modelos no son suficientes para cubrir todas las necesidades de conocimiento que se demandan en la actualidad.

SDRD [9] es un área de investigación que aglutina a todas las técnicas de minería de datos capaces de obtener conocimiento descriptivo relacionado con una variable de interés para el usuario, mediante el uso de patrones o reglas. El objetivo es la descripción de las características que producen el valor de la variable objetivo. Por ejemplo, en el estudio de un tipo de cáncer, SDRD no pretende predecir si un nuevo paciente puede tener la enfermedad, sino describir aquellos factores más relevantes que lo producen.

Una definición clásica de patrón se puede encontrar en [37] como: sea $I = \{i_1, i_2, ..., i_n\}$ un conjunto de selectores. Un selector se define mediante una tripleta (v, r, S), donde v es una variable del problema; r es un operador relacional, como por ejemplo $=, \neq, \in, \notin, >, <, \geqslant, \leqslant$, entre otros; y S es un valor o conjunto de valores pertenecientes al dominio de la variable v. Uno de estos selectores será la variable objetivo del problema, denominada v_c. Esta definición de patrón, establecida anteriormente al desarrollo de SDRD, sigue manteniéndose vigente en este ámbito. Por tanto, un patrón P de un modelo SDRD se representa como:

$$P : C \rightarrow v_c \tag{1.1}$$

donde v_c es el valor de la variable objetivo y $C \subseteq I$. Los selectores presentes en C suelen representarse unidos mediante conjunciones o en forma normal disyuntiva (en inglés, *Disjunctive Normal Form* (DNF)) para una representación más compacta.

Como se ha comentado anteriormente, la finalidad de SDRD se centra en la comprensión del fenómeno subyacente que determina el valor de la variable de interés. De este objetivo se puede deducir que no es necesaria la extracción de relaciones completas, sino que con relaciones parciales que sean relevantes o interesantes es suficiente. De hecho, estas relaciones deben ser analizadas y tratadas de manera individual por parte del experto, por lo que es interesante que el solapamiento entre los patrones extraídos sea el menor posible para evitar redundancias.

De entre las técnicas enmarcadas dentro de SDRD (véase Página 5), esta tesis se centra en SD y en EPM debido a su diversidad de aplicaciones reales. Por lo tanto, a continuación se presentan:

1. Las características que deben tener los modelos SDRD, haciendo especial hincapié en las medidas de calidad utilizadas. La gran cantidad de medidas establecidas a lo largo de la literatura para determinar la calidad de

los patrones desde diferentes puntos de vista hace que sea un factor fundamental a la hora de guiar el proceso de búsqueda. Por lo tanto, se hace obligatorio la determinación de aquellas que resultan más interesantes.

2. Se presenta una introducción del concepto de SD, en donde se presentarán los principales objetivos, enfoques y herramientas utilizadas a lo largo de la literatura.

3. Finalmente, el tercer concepto que se introduce en esta memoria es EPM, analizando el problema y los enfoques más empleados para la extracción de conocimiento.

1.1.1. Medidas de calidad en el marco de descubrimiento de reglas descriptivas basadas en aprendizaje supervisado

Es deseable que el conocimiento extraído en SDRD muestre un buen balance entre generalidad, precisión, interés e interpretabilidad de los modelos, de modo que estos sirvan como herramienta de apoyo a las decisiones finales del usuario. Este enfoque supone que, además de las particularidades de cada una de las tareas que se engloban dentro del SDRD, se deben de optimizar varios objetivos de manera simultánea. Además, dichos objetivos a optimizar son contrapuestos entre sí [38], por lo que el problema de la extracción de patrones descriptivos mediante aprendizaje supervisado se torna en un problema multiobjetivo.

Todos estos objetivos deben ser definidos para determinar la calidad de los patrones obtenidos por los modelos SDRD y para guiar los procesos de búsqueda de forma adecuada. Sin embargo, no existe un claro consenso sobre qué medida de calidad es la más adecuada para la determinación de los diferentes objetivos [17]. Por esta razón, en la literatura se han definido una gran cantidad de métricas para determinar la calidad del conocimiento extraído [39], [40]. Todas estas medidas pueden ser determinadas a partir de los valores de una tabla de contingencias, en donde se muestra el número de ejemplos correcta e incorrectamente cubiertos y no cubiertos para cada uno de los patrones extraídos. Este tipo de tabla se presenta en la Tabla 1.1.

En esta tabla se define tp como el número de instancias correctamente cubiertas, es decir, que satisfacen tanto el antecedente como el consecuente del patrón. El valor de fp es el número de instancias que cumplen el antecedente, pero no el consecuente del patrón. El valor de fn es el número de instancias que

Tabla 1.1: Tabla de contingencia de un patrón.

	Clase	No Clase
Cubierto	tp	fp
No cubierto	fn	tn

no cumplen el antecedente del patrón, pero satisfacen el consecuente. Finalmente, tn se refiere al número de instancias que no cumplen ni con el antecedente, ni con el consecuente del patrón.

Entre los esfuerzos llevados a cabo para establecer un consenso en las métricas a utilizar, destaca el análisis llevado a cabo en [41] en donde se establece una conexión directa entre las técnicas más relevantes de SDRD y la medida de atipicidad o *Weighted Relative Accuracy* (WRAcc). Dicha medida se define en la Ecuación (1.2):

$$WRAcc(P) = \frac{tp + fp}{tp + fp + fn + tn}\left(\frac{tp}{tp + fp} - \frac{tp + fn}{tp + fn + fp + tn}\right) \quad (1.2)$$

Esta métrica presenta complejidades a la hora de realizar comparativas entre patrones de diferentes clases, pues posee una dependencia directa con el porcentaje de instancias que pertenecen a la misma. Por tanto, se hace necesaria una normalización de la medida, la cual se define en la Ecuación (1.3):

$$WRAcc_{Norm}(P) = \frac{WRAcc(P) - (1 - \%Pos)\,(0 - \%Pos)}{\%Pos\,(1 - \%Pos) - (1 - \%Pos)\,(0 - \%Pos)} \quad (1.3)$$

donde $\%Pos = \frac{tp+fn}{tp+fp+tn+fn}$ es el porcentaje de instancias pertenecientes a la clase objetivo del problema. Tras esto, la métrica pasa de un dominio variable que contiene valores negativos al dominio [0,1], permitiendo establecer comparativas en estudios, de manera que:

- P es irrelevante si $WRAcc_{Norm}(P) \leqslant 0{,}5$.
- P es de interés si $WRAcc_{Norm}(P) > 0{,}5$ y además, cuanto mayor sea este valor, mayor será su ganancia de precisión.

Un análisis en profundidad y pormenorizado de las características de la medida WRAcc en las tareas de SDRD se encuentra en [41].

Asimismo, existen diferentes estudios en el contexto de SDRD donde se analizan una gran cantidad medidas de calidad [40], [42]. Atendiendo a los resultados presentados en estos estudios, las métricas que pueden utilizarse para definir los diferentes objetivos son las siguientes:

- Precisión. Es el objetivo más importante pues los patrones deben definir de manera fiable las relaciones descritas. Por ello, se establecen varias métricas:
  - Confianza. Determina la precisión del patrón con respecto a las instancias que ha cubierto. Se calcula como se describe en la Ecuación (1.4) [43].

$$\mathrm{Conf}(P) = \frac{tp}{tp + fp} \tag{1.4}$$

  - Índice de crecimiento (en inglés, *Growth Rate* (GR)). Determina la capacidad discriminatoria del patrón al calcular el ratio de soporte entre las diferentes clases. Se calcula como se describe en la Ecuación (1.5).

$$GR(P) = \begin{cases} 0, & \text{Si } fp\,(tp + fn) = tp\,(fp + tn) = 0, \\ \infty, & \text{Si } fp\,(tp + fn) = 0 \wedge tp\,(fp + tn) \neq 0, \\ \frac{tp(fp+tn)}{fp(tp+fn)}, & \text{en otro caso} \end{cases} \tag{1.5}$$

  - Tasa de falsos positivos (en inglés, *False Positive Rate* (FPR)). Calcula el porcentaje de instancias incorrectamente cubiertas con respecto al total que no pertenecen a la clase objetivo. De este modo, nos permite determinar la magnitud de los errores de precisión del patrón. Es importante destacar que esta métrica debe ser minimizada. Se calcula como se presenta en la Ecuación (1.6) [44].

$$FPR(P) = \frac{fp}{fp + tn} \tag{1.6}$$

- Interés. Pretende extraer conocimiento que sea novedoso, sorprendente o útil para el experto. Para ello, las métricas más utilizadas para determinar el interés de un patrón son:

- WRAcc, presentado en la Ecuación (1.2). Esta medida busca un balance entre el número de instancias cubiertas por un patrón y la ganancia de precisión respecto al porcentaje de la clase. Por lo tanto, cuanto más alto sea este valor, más interesante será el patrón, pues se obtiene una mayor precisión influyendo sobre más instancias del problema.
- Índice de Jaccard (Jacc). Esta medida se utiliza para determinar la similitud entre dos conjuntos de elementos. En concreto, estos son el formado por las instancias que pertenecen a la clase y por las instancias cubiertas por el patrón, respectivamente. Esta medida ayuda a encontrar patrones interesantes con un buen balance entre generalidad y fiabilidad. Se calcula como se presenta en la Ecuación (1.7) [45].

$$Jacc(R) = \frac{tp}{tp + fp + fn} \tag{1.7}$$

- Generalidad. Se busca que la cobertura de un patrón respecto a la totalidad de instancias de un problema sea máxima. Por lo general, una de las métricas más empleadas para determinar este objetivo es la tasa de verdaderos positivos (en inglés, *True Positive Rate* (TPR)). Esta medida cuantifica el porcentaje de instancias correctamente cubiertas por el patrón con respecto a la totalidad de ejemplos pertenecientes a la clase del patrón. Esta medida se define según se presenta en la Ecuación (1.8) [14].

$$TPR(P) = \frac{tp}{tp + fn} \tag{1.8}$$

- Interpretabilidad. En este objetivo se pretende minimizar la complejidad del modelo de patrones extraído. Se pretende así que sea más fácil y rápido analizar y entender el conocimiento extraído para mejorar la toma de decisiones. En un modelo SDRD, las métricas de complejidad más utilizadas son el número de patrones extraídos y el número medio de selectores o variables que forman parte del antecedente de un patrón.

Se puede observar que el empleo de una u otra medida de calidad va a depender del contexto de aplicación particular. No obstante, en líneas generales, es interesante remarcar la capacidad de la medida WRAcc no solo para encontrar patrones interesantes de una manera robusta, sino para relacionar las diferentes

técnicas de SDRD entre sí [41]. Asimismo, se destaca el índice de Jaccard por su gran capacidad de obtener patrones interesantes, especialmente en problemas con desbalanceo de clases [39], muy comunes en *big data* o flujos continuos de datos. Por lo tanto, dichas métricas pueden ser empleadas como base para guiar procesos de búsqueda de patrones. También se remarca el empleo de otras medidas específicas para cada uno de los objetivos de SDRD. El empleo de estas métricas en la etapa de búsqueda puede ayudar a crear sinergias con las descritas anteriormente. Por otro lado, su empleo en la evaluación de resultados puede permitir obtener conclusiones desde diferentes puntos de vista. Por ejemplo, empleando TPR se puede determinar el alcance global de la cobertura de un patrón. Unido a lo anterior, puede ser interesante el empleo a su vez de medidas como la confianza y el FPR, o similares, que permitan determinar la precisión del patrón a nivel local, es decir, respecto al número de ejemplos cubiertos, y a nivel global, es decir, respecto al total de datos.

1.1.2. Descubrimiento de subgrupos

SD [14], [15] se define como la búsqueda de subconjuntos de la población cuya distribución estadística con respecto a una variable objetivo difiera significativamente respecto al conjunto total de la población. Esta desviación hace que los subgrupos generados sean interesantes para el experto. En particular, la tarea lo que pretende es la búsqueda de aquellos subgrupos más interesantes. Estos, entre otras características, son los que influyen en el mayor número de instancias posible con la distribución estadística más inusual.

En los últimos años, SD ha sido ampliamente analizado por la comunidad científica [16]-[18] en donde han surgido propuestas basadas en algoritmos clásicos como AprioriSD [46], SD-Map [47] o MergeSD [48]; basadas en programación genética como CGBA-SD [49]; y basadas en sistemas difusos evolutivos como SDIGA [50], MESDIF [51], NMEEF-SD [52], FuGePSD [53]. También hay que destacar enfoques para el análisis *big data* basados en algoritmos exactos como AprioriK-SD-OE o PFP-SD-OE [54], y basados en sistemas difusos evolutivos como MEFASD-BD [55]. Es importante destacar que, hasta donde se conoce, no se han realizado implementaciones para SD enfocadas a la extracción de conocimiento en flujos de datos.

Uno de los aspectos más relevantes a la hora de solucionar un problema mediante SD, es la influencia de factores externos, como ruido o presencia de datos perdidos, que pueden afectar negativamente en la calidad del conocimiento extraído. En este sentido, en la literatura se ha analizado el comportamiento de

los algoritmos con respecto a valores perdidos [56]. Asimismo, factores internos como la redundancia en los patrones extraídos provoca un decremento de la calidad del conocimiento extraído. En este aspecto, los investigadores de la literatura sí han analizado el problema y se han propuesto diversas soluciones al mismo [57], [58].

Finalmente, es importante destacar que actualmente se encuentran disponibles para la comunidad científica muchos paquetes desarrollados para distintas plataformas de software abierto. Por ejemplo, existe un paquete con algoritmos clásicos de SD desarrollado para la herramienta R denominado *rsubgroup*[1]. También están disponibles paquetes de algoritmos integrados en las herramientas KEEL[2] [59] y Orange[3] [60].

1.1.3. Minería de patrones emergentes

EPM [20], [61] tiene como objetivo la búsqueda de patrones que describen un alto número de instancias de una clase determinada, minimizando la inclusión de elementos que no pertenecen a dicha clase. Por lo tanto, el patrón tendrá un soporte muy alto para una clase y muy bajo para el resto. Así, la tarea intenta buscar patrones con un gran poder discriminatorio.

En concreto, la minería de patrones emergentes se define como la búsqueda de aquellos patrones P cuyo GR entre dos conjuntos de datos, o dos clases, sea mayor a un valor umbral preestablecido $\rho > 1$. Este GR se calcula siguiendo la fórmula presentada en Ecuación (1.9) [20].

$$GR(P) = \begin{cases} 0, & \text{Si } fp\,(tp + fn) = tp\,(fp + tn) = 0, \\ \infty, & \text{Si } fp\,(tp + fn) = 0 \wedge tp\,(fp + tn) \neq 0, \\ \frac{tp\,(fp+tn)}{fp\,(tp+fn)}, & \text{en otro caso} \end{cases} \tag{1.9}$$

donde tp, fp, tn y fn son los valores de la tabla de contingencias presentada en la Tabla 1.1.

Con estas características, los principales objetivos abordados por EPM son la búsqueda de patrones que describen características discriminatorias entre clases de un conjunto de datos, la búsqueda de tendencias emergentes a lo largo del tiempo o la diferenciación a través de un grupo de variables.

[1] https://cran.r-project.org/web/packages/rsubgroup/index.html
[2] http://keel.es
[3] http://kt.ijs.si/petra_kralj/SubgroupDiscovery

A lo largo de la literatura se han ido desarrollando una amplia variedad de paradigmas y algoritmos de extracción de patrones emergentes, entre los que destacan el algoritmo DeEPs [62] como un método de extracción basado en límites; el algoritmo StrongJEP [35] o Tree-based JEP [63], como enfoques basados en árboles, entre otros. No obstante, a pesar de las capacidades descriptivas de estos patrones, los investigadores han centrado sus esfuerzos en el aspecto predictivo de los mismos [22]. Es decir, de los factores que son deseables en un modelo SDRD (generalidad, precisión, interés e interpretabilidad), únicamente se ha analizado en profundidad la precisión, ignorando prácticamente el resto de cualidades. También ha habido grandes esfuerzos en el área para paliar la influencia del ruido en los datos [35], [64]-[66].

Recientemente, la comunidad científica ha estado más interesada en otros aspectos de la tarea y se han presentado propuestas como el algoritmo FEPM [67], el cual hace uso de lógica difusa para mejorar la interpretabilidad; o el algoritmo EvAEP [68], basado en un sistema difuso evolutivo que intenta obtener un balance entre todos los aspectos de los modelos SDRD. Sin embargo, hasta donde se sabe, no se ha presentado ninguna propuesta para extraer patrones emergentes en problemas complejos como *big data* o flujos continuos de datos.

1.2. Inteligencia computacional

La Inteligencia computacional [69] es un amplio abanico de áreas que se incluyen dentro del campo de la Inteligencia artificial centradas en el diseño de sistemas inteligentes capaces de resolver problemas complejos. Estos se encuentran habitualmente inspirados en fenómenos de la naturaleza o en el razonamiento lingüístico humano, ya que poseen una alta tolerancia hacia la imprecisión e incertidumbre, lo que les confiere una gran capacidad de adaptación a entornos cambiantes de manera robusta. Entre otras, las áreas más destacadas que se incluyen dentro de la Inteligencia computacional son los sistemas difusos [70], [71], los modelos probabilísticos [72], los algoritmos de optimización bioinspirados como la computación evolutiva [73] y las redes neuronales artificiales [74].

Esta tesis doctoral se centra en el desarrollo de sistemas basados en reglas difusas, los algoritmos evolutivos y la combinación de ambas técnicas en los sistemas difusos evolutivos orientados a SDRD, debido al éxito que han tenido estas técnicas en tareas como SD y a su alta interpretabilidad. Dichas técnicas se describen con detalle en las siguientes secciones.

1.2.1. Sistemas basados en reglas difusas

La lógica difusa tiene como finalidad modelar el conocimiento impreciso y cuantitativo, así como la posibilidad de manejar la incertidumbre, con el fin de adaptar este conocimiento de una forma más cercana al razonamiento humano. Fue introducida por Zadeh [70], [71] y se fundamenta en el concepto de conjunto difuso. Un conjunto difuso es una generalización de los conjuntos clásicos, en los que un elemento x cualquiera únicamente puede estar o no estar en un conjunto S. Por lo tanto, la función de pertenencia $\mu(x)$ se define como:

$$\mu(x) = \begin{cases} 0, & \text{Si } x \notin S, \\ 1, & \text{Si } x \in S \end{cases} \tag{1.10}$$

Como se puede observar, la función de pertenencia en conjuntos clásicos únicamente posee dos valores. En contraposición, una función de pertenencia sobre un conjunto difuso puede adquirir cualquier valor dentro del intervalo $[0, 1]$. Esto permite que un elemento x pueda pertenecer a varios conjuntos difusos en mayor o menor medida. Asimismo, se permiten representar límites difusos, pero con un significado mucho más preciso en aplicaciones reales. En [75] se puede encontrar una descripción detallada sobre la teoría de conjuntos difusos.

No obstante, uno de los mayores potenciales de los conjuntos difusos es la creación de variables lingüísticas a partir de variables de tipo numérico. Para ello se establece un conjunto de valores lingüísticos y se define el significado de cada uno de ellos a partir de un conjunto difuso, utilizando solapamiento entre ellos [71]. De este modo, una variable numérica se puede tratar mediante términos lingüísticos, como *Bajo, Medio* y *Alto*, los cuales tienen cierto solapamiento entre sí. Cada uno de estos términos es especificado mediante un conjunto difuso, el cual tendrá una función de pertenencia asociada. De este modo, se puede saber de manera precisa el grado de pertenencia de cualquier valor de la variable a cada uno de los términos lingüísticos especificados. En la Figura 1.1 se muestra un ejemplo de una partición difusa para una variable numérica considerando cinco etiquetas lingüísticas: *Muy bajo, Bajo, Medio, Alto* y *Muy alto*, donde *Bajo, Medio* y *Alto* tienen una función de pertenencia triangular y el resto una función de pertenencia trapezoidal.

Esta expresividad que poseen los conjuntos difusos permite la simplificación de reglas y sistemas que utilizan lógica difusa. Los sistemas basados en reglas que emplean lógica difusa se denominan sistemas basados en reglas difusas [76], [77]. Estos son una extensión de los sistemas de reglas clásicos, en donde se

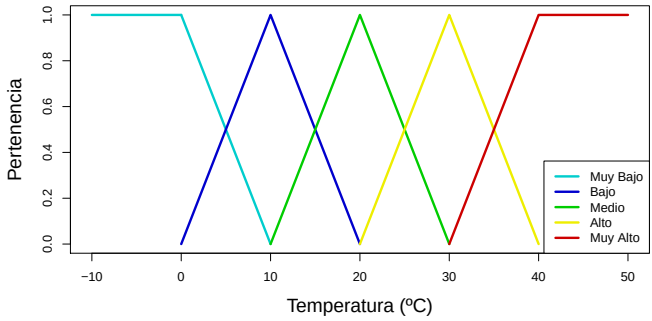


Figura 1.1: Ejemplo de definición de la variable *Temperatura* con cinco etiquetas lingüísticas. *Fuente: elaboración propia.*

emplean una base de reglas "SI-ENTONCES" donde tanto en la parte del antecedente, como en el consecuente, se considera el empleo de lógica difusa. Por esta razón, además del conjunto de reglas, estos sistemas deben almacenar información relativa a la definición de los diferentes términos lingüísticos establecidos para cada variable. Los sistemas basados en reglas difusas han demostrado su habilidad para enfrentarse a diferentes problemas como control, clasificación, minería de datos, etc. Los trabajos pioneros de aplicación de estos sistemas a estos problemas se pueden observar en [78]-[82].

Dentro de los sistemas basados en reglas se pueden emplear diferentes esquemas para su representación. Las más habituales son la representación canónica y la representación DNF. En la primera, el antecedente de una regla es creado mediante pares variable-valor unidos por conjunciones. Por lo que una variable podrá tener únicamente un valor. Por el contrario, la representación DNF permite una mayor flexibilidad permitiendo que una variable pueda tener varios valores al mismo tiempo, unidos mediante disyunciones, mientras que la unión entre las diferentes variables del antecedente se realiza mediante conjunciones. Un ejemplo de reglas canónicas y DNF se muestra en las Ecuaciones (1.11) y (1.12), respectivamente.

$$\mathrm{P}_{\mathrm{can}} : \mathrm{SI}\ v_1 = \mathit{Bajo} \wedge v_3 = \mathit{Medio}\ \ \mathrm{ENTONCES}\ \ v_{\mathrm{c}} \tag{1.11}$$

$$P_{dnf} : \text{SI } v_1 = (Bajo \vee Medio) \wedge v_3 = Medio \text{ ENTONCES } v_c \quad (1.12)$$

Para aplicar una regla difusa en una instancia concreta, es necesario que el grado de compatibilidad con el antecedente de la regla sea mayor a un umbral α preestablecido. Esto implica que la instancia se encuentra dentro de la zona del espacio que describe la regla. Dicho grado de compatibilidad se calcula según se indica en la Ecuación (1.13).

$$APC(e, P) = T\left(TC\left(\mu_1^1(e_1), ..., \mu_n^1(e_k)\right), ..., TC\left(\mu_1^k(e_1), ..., \mu_n^k(e_k)\right)\right) \quad (1.13)$$

donde:

- e_k indica el valor de la variable v_k en la instancia e.
- μ_n^k indica la función de pertenencia del término lingüístico n en la variable v_k.
- TC indica la aplicación de una t-conorma difusa, es decir, una operación *O* difusa. Habitualmente se emplea la t-conorma máximo.
- T indica la aplicación de una t-norma difusa, es decir, la aplicación de la operación *Y* difusa. Habitualmente se emplea una t-norma mínimo.

1.2.2. Algoritmos evolutivos

La computación evolutiva [83], [84] se basa en la utilización de algoritmos estocásticos inspirados en la evolución natural de las especies, o en otros procesos naturales, para problemas de optimización y búsqueda. En esta tarea se engloban numerosos paradigmas como los algoritmos genéticos [84], la programación genética [85], [86] o las estrategias evolutivas [87], entre otras.

El esquema general de funcionamiento de este tipo de algoritmos, representado en la Figura 1.2, se podría resumir de la siguiente manera [88]:

1. Se parte de una población inicial de individuos o cromosomas, los cuales codifican un conjunto de soluciones completas o parciales al problema dado.

2. Estos cromosomas irán evolucionando a lo largo del tiempo a través de la aplicación de los operadores genéticos de selección, cruce y mutación para generar una nueva población de descendientes. Mediante este proceso, se produce en la población una competición y variación controlada entre los individuos.

3. Una vez creada la población de descendientes, esta sustituirá a la población de padres mediante un operador de remplazo.

4. Tras esto, el proceso evolutivo volverá a empezar de nuevo. Este ciclo continua hasta que se satisfaga cierta condición de parada, como por ejemplo el número de ciclos evolutivos llevados a cabo, devolviendo la mejor o mejores soluciones encontradas al experto.

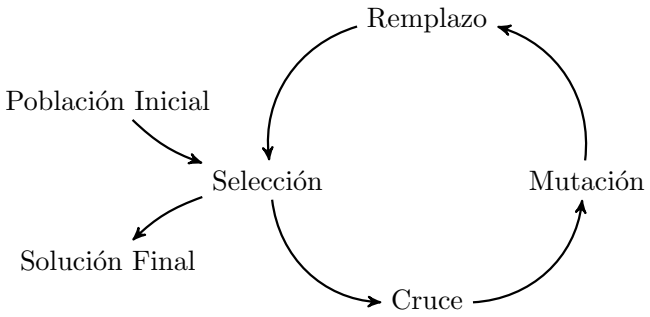


Figura 1.2: Esquema general de funcionamiento de un algoritmo evolutivo. *Fuente: elaboración propia.*

Para poder resolver un problema utilizando computación evolutiva, es necesario definir una serie de elementos, los cuales se describen a continuación:

- Representación de la solución en los individuos. Es el factor más importante pues va a influir en cómo el resto de mecanismos van a ser aplicados. Existen diferentes tipos de representación, entre los que destacan la codificación binaria, real o basada en orden, entre otras.

- Mecanismo de selección. Se encarga de seleccionar aquellos individuos candidatos a reproducirse para crear la población de descendientes. El uso de un operador u otro determinará la presión en la competición entre individuos, así como la convergencia del método [89].

- Operadores genéticos. Son los encargados de generar la población de descendientes modificando y compartiendo la información almacenada entre los individuos de la población seleccionados previamente. Estos operadores son clave en el proceso de búsqueda pues permiten obtener un balance adecuado entre la capacidad de exploración y explotación del espacio de búsqueda [90].

- Modelo de población o remplazo. Define cómo evoluciona la población de soluciones candidatas a lo largo del tiempo. Entre otros, los modelos más destacados son el modelo generacional y el estacionario [91].

1.2.3. Sistemas difusos evolutivos

Los sistemas basados en reglas difusas han demostrado ser una herramienta útil en un amplio número de problemas para representar el conocimiento de una manera más cercana al razonamiento humano, junto a una mejora de la robustez frente a la incertidumbre [92]-[94]. Sin embargo, los componentes de estos sistemas se pueden mejorar mediante un proceso evolutivo de aprendizaje u optimización de modo que su comportamiento se adapte a un contexto en particular [95]. La combinación entre lógica difusa y algoritmos evolutivos para la inducción de conocimiento se conoce en la literatura como sistemas difusos evolutivos y han sido exitosos en múltiples áreas de aplicación [95]-[100].

Los sistemas difusos evolutivos se pueden dividir en dos tipos según el componente del sistema basado en reglas difusas que el algoritmo evolutivo está optimizando [98]: algoritmos que ajustan la base de datos de términos lingüísticos, o el sistema de inferencia; y algoritmos que aprenden la base de conocimiento, la base de términos lingüísticos o todos los componentes.

Como se ha comentado anteriormente, el principal elemento de los algoritmos evolutivos es la representación empleada para los cromosomas. En este aspecto, las codificaciones más habituales empleadas en los sistemas difusos evolutivos para aprendizaje de reglas son:

- El enfoque "Cromosoma = Conjunto de reglas", también conocido como enfoque *Pittsburgh*, en donde un individuo representa un conjunto de reglas y se devuelve el mejor de ellos [101].
- El enfoque "Cromosoma = Regla", en donde un individuo representa una única regla y el resultado final se presenta como un conjunto de diferentes individuos seleccionados según cierto criterio. Dentro de este enfoque, podemos encontrar tres propuestas diferentes para presentar este conjunto de reglas:
  - El enfoque Michigan [102]. Un individuo representa una única regla y se devuelve toda la población. Estos sistemas se emplean habitualmente en clasificación.
  - El enfoque *Iterative Rule Learning* (IRL) [103]. En este enfoque se devuelve únicamente el mejor individuo encontrado a lo largo del proceso evolutivo. Este se ejecuta de manera iterativa hasta cumplir cierta condición de parada, formando así el conjunto de reglas final.
  - El enfoque "cooperativo-competitivo" [104]. Se extrae un subconjunto de la población, de modo que los individuos por un lado compiten entre sí para no ser eliminados, pero por otro cooperan para mejorar la calidad del subconjunto de reglas final.

1.3. Problemas complejos en Ciencia de datos

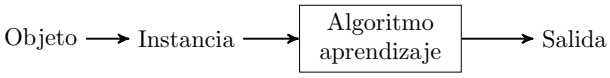
Debido a la gran cantidad de información generada a día de hoy, existe un interés elevado por parte de la comunidad hacia el diseño y desarrollo de métodos capaces de abordar esta cantidad de información en un tiempo razonable. No obstante, esto es todo un reto ya que los axiomas en los que funcionan los enfoques tradicionales ya no se sustentan en estos ámbitos. Por lo tanto, muchos problemas son inabordables desde una perspectiva tradicional.

En esta tesis se abordarán estos problemas desde el punto de vista descriptivo y por ello se profundiza en sus características en las siguientes secciones.

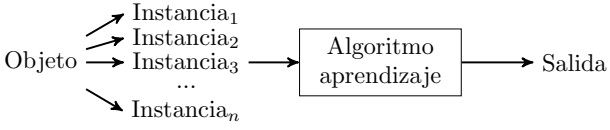
1.3.1. Aprendizaje multi-instancia

En la minería de datos tradicional, cada objeto o instancia se describe inequívocamente por medio de un vector de características que se asocia, a su vez, con un valor de salida o variable objetivo. En el aprendizaje multi-instancia (en

inglés, *Multiple-Instance Learning* (MIL)) [105], las estructuras de datos son más complejas. En este caso, cada valor de salida se asocia de manera ambigua con un conjunto indeterminado de vectores de características que están relacionadas con dicho valor objetivo. Un ejemplo ilustrativo se muestra en la Figura 1.3. En MIL, cada una de las instancias que comparten el valor de salida se les denomina bolsa. Es importante destacar que, cada una de las instancias individuales que conforman la bolsa tiene un valor asociado de la variable. Sin embargo, ese valor es desconocido y únicamente se conoce el valor de salida asociado con la bolsa. De hecho, a veces no todas las instancias incluidas en un bolsa son relevantes, puede haber algunas que no tengan ninguna información relevante sobre el valor de salida o que están más relacionadas con otras clases, incluyendo ruido en la información.



(a) Aprendizaje con una única instancia.



(b) Aprendizaje multi-instancia.

Figura 1.3: Diferencia entre un aprendizaje clásico con una única instancia y un problema MIL *Fuente: elaboración propia a partir de [106].*

Un ejemplo donde un problema MIL se puede representar de manera natural puede ser la clasificación de imágenes. En este caso, se extraen de cada imagen diferentes zonas o regiones del espacio, pero únicamente se etiqueta la imagen de manera global. Para clasificar imágenes en donde haya una playa, se pueden extraer diferentes regiones de una imagen que representen, por ejemplo, el agua, la arena u otros elementos que pueden no corresponderse con una playa. Sin embargo, en su conjunto todas estas características extraídas en la imagen si representaría una playa.

A lo largo de la literatura se han realizado varias implementaciones de algoritmos para resolver problemas MIL. En [107] se presenta una revisión de los paradigmas más destacados que se han utilizado en MIL para extraer conocimiento de este tipo de información:

- Espacio de instancia. La información discriminadora se encuentra en el ámbito de la instancia. Esto implica que únicamente se tienen en cuenta las características que presentan las instancias. Por lo tanto, la clasificación se realiza en función de una puntuación entre instancias. A este proceso se le denomina asunción colectiva [108], o basada en umbral. Esta se basa en la clasificación de una bolsa como positiva (en problemas binarios) si y solo si el número de instancias clasificadas como positivas es mayor a dicho umbral. En caso contrario, la bolsa se clasificará como negativa. Si este valor umbral es igual a uno, se le denomina asunción estándar del problema MIL.

- Espacio de bolsa. En este caso la información discriminadora recae en las características propias de la bolsa, por lo que estas son tratadas como una entidad completa. En este enfoque, el proceso de aprendizaje discrimina entre bolsas. Habitualmente se definen funciones de distancia entre bolsas para ser utilizas en clasificadores como K-NN o SVM.

- Espacio embebido. También basado en información relativa a la bolsa, este enfoque mapea información global de la bolsa a un único vector de características que la resume. Por lo tanto, cualquier algoritmo de clasificación clásico pude ser empleado.

A pesar de que el desarrollo de algoritmos para MIL tiene un claro componente predictivo, también posee especial interés en un análisis descriptivo de los datos. Por ejemplo, dentro de un problema clásico de análisis de cesta de la compra, múltiples compras realizadas por un mismo cliente pueden estar agrupadas en la misma bolsa. En este tipo de análisis descriptivo, el enfoque MIL no puede llevarse a cabo mediante los enfoques del espacio de bolsa y espacio embebido ya que cada instancias es relevante de manera aislada y se debería de analizar una a una para cada bolsa. Sin embargo, hasta donde alcanza nuestro conocimiento, no se ha realizado ningún desarrollo de algoritmos SDRD para este tipo de problema.

1.3.2. Big data

Según un informe de la empresa Cisco[4], para el año 2021 se espera una generación de 847 *zettabytes*[5] de datos. De hecho, se espera que la cantidad de datos generados sea dos órdenes de magnitud mayor que la cantidad de datos almacenados en ese mismo año. Por lo tanto, una gran cantidad de información tiene un carácter efímero. La extracción de conocimiento en estos datos masivos se ha convertido en uno de los problemas más importantes debido a que los algoritmos de minería de datos actuales son incapaces de abordar la magnitud de estos datos. Por esta razón, se necesitan tecnologías que sean capaces de recoger, mantener, transmitir y procesar grandes volúmenes de datos en un tiempo razonable.

A esta ingente cantidad de datos se le conoce actualmente como *big data*. Laney [109] definía el concepto como un conjunto de datos de gran volumen, que llega a gran velocidad a los sistemas de cómputo desde una gran variedad de fuentes. A esta definición, conocida como el modelo de las tres uves del *big data*, se le han ido añadiendo elementos posteriormente como veracidad y valor de los datos, entre otros. En general, cualquier problema de análisis de datos que sea lo suficientemente grande y complejo como para poder ser procesado por una única máquina en un tiempo razonable puede ser definido como un problema *big data*.

Para intentar mitigar los efectos de los problemas *big data*, han surgido múltiples herramientas y algoritmos que hacen uso de la computación distribuida para abordar esta ingente cantidad de datos. Como resultado de la evolución de estas tecnologías, ha surgido todo un ecosistema de plataformas y técnicas para afrontar los diferentes escenarios que se pueden dar en la industria y en la ciencia [23]. El objetivo último de todas estas tecnologías es acercar la computación distribuida en grandes centros de computación al usuario estándar, haciendo totalmente transparente los aspectos técnicos de estos entornos.

Uno de los paradigmas de computación distribuida que más éxito ha cosechado en el análisis *big data* es el paradigma MapReduce [31], [110], así como su contraparte de código abierto, Hadoop [111]. La principal ventaja de este sistema es que el empleo de computación distribuida es totalmente transparente al usuario. De hecho, provee de manera automática mecanismos de tolerancia a fallos, creación de particiones de datos, así como asignación de recursos computacionales a diferentes trabajos, entre otros.

[4] https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.html

[5] 1 *zettabyte* = 10^{21} *bytes*

MapReduce se basa en el paradigma divide y vencerás, llevado al ámbito de la computación distribuida. Para el diseño efectivo de un algoritmo bajo este paradigma, el desarrollador deberá dividir el procesamiento de datos en dos etapas principales: *map*, y *reduce*. En la primera, los datos será divididos y procesados de manera distribuida, mientras que en la segunda etapa se agregarán los resultados obtenidos en la fase anterior.

Concretamente, uno de los pilares fundamentales del paradigma es el empleo de pares clave-valor, utilizados internamente durante todo el flujo de trabajo. El funcionamiento básico del paradigma MapReduce, representado gráficamente en la Figura 1.4, se define a continuación:

1. En la fase *map*, los datos se dividen en diferentes particiones, las cuales se envían a los diferentes nodos de cómputo, optimizando la localidad de los datos. Cada una de estas particiones se identifica con una tupla (k_i, v_i) siendo k_i la clave y v_i los datos de la partición. A continuación, los nodos procesan de manera distribuida los datos de acuerdo a la función *map* programada por el usuario. Al finalizar, se generará una nueva tupla (k_i, v_i') que contendrá resultados parciales.

2. Tras esto, existe una fase intermedia, denominada *shuffle*, en donde se agrupan y se ordenan los resultado parciales obtenidos anteriormente por clave, si fuera necesario. Este proceso es totalmente transparente al usuario pero debe ser tenido en cuenta en el proceso de diseño debido a su alto impacto en el rendimiento.

3. Finalmente, en la fase *reduce* se agregan los resultados parciales producidos por el procedimiento *map*, utilizando la función *reduce* programada por el usuario. Este procedimiento se ejecuta para cada clave k_i. El resultado final de la función es una nueva tupla (k_i, v_i'') que contendrá el resultado final para cada clave.

Sin embargo, uno de los principales inconvenientes de la herramienta Hadoop es su bajo rendimiento en ciertas condiciones, principalmente en entornos iterativos, debido a la sobrecarga de re-ejecutar todo un trabajo completo cuando no es necesario [112]. Por esta razón, se han realizado esfuerzos para el desarrollo de nuevas herramientas basadas en MapReduce que solucionen los problemas que presenta Hadoop. Entre otras, una de las herramientas más populares que soluciona este problema es Apache Spark [113], el cual presenta una mejora significativa de rendimiento con respecto a Hadoop, principalmente en procesos

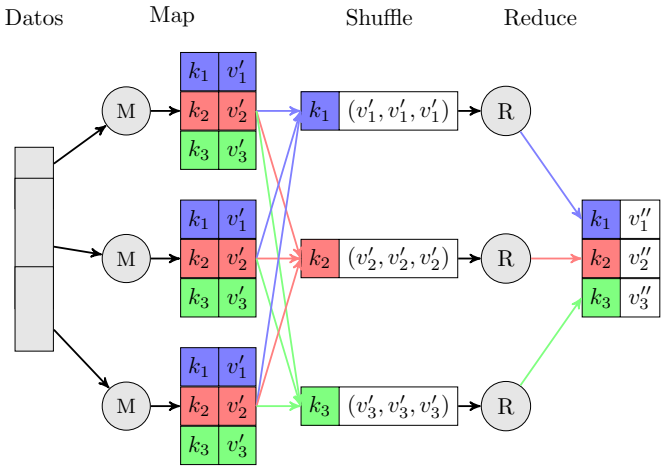


Figura 1.4: Esquema de trabajo del paradigma MapReduce. *Fuente: elaboración propia.*

iterativos, gracias a una abstracción de datos denominada *Resilient Distributed Dataset* (RDD). Los RDD se caracterizan por poseer un conjunto de operaciones denominadas transformaciones y acciones, que se basan en un uso intensivo de memoria principal. Asimismo, Spark posee una librería donde los principales algoritmos de minería de datos han sido adaptados para el procesamiento distribuido, denominada MLlib [114], junto a un submódulo para el procesamiento por lotes de flujos de datos, denominado Apache Spark Streaming.

En esta tesis doctoral los métodos de extracción de patrones descriptivos se basan en algoritmos evolutivos, los cuales tienen un importante componente iterativo. Por esta razón, el desarrollo de estos métodos complejos se desarrollarán para la herramienta Apache Spark.

1.3.3. Minería de flujo de datos

La cantidad de datos que se generan y no se almacenan es de dos órdenes de magnitud superior a la cantidad de datos que finalmente se almacena, de acuerdo al informe de Cisco mencionado en el apartado anterior. Esto supone que la mayoría de datos generados tiene un carácter efímero y que su interés es rele-

vante hasta unos instantes después de su generación, para ser posteriormente desechados. Esta información, generada principalmente por la alta conectividad entre dispositivos y personas, y el IoT, contiene un conocimiento que es interesante para la mejora a corto y largo plazo de la calidad de los procesos, servicios o mejora del bienestar de las personas. A todos estos datos efímeros y generados constantemente se les denomina flujos de datos. Analizar estos flujos de datos es uno de los retos más relevantes de la minería de datos actual, no sólo por sus características, que se detallarán más adelante, sino por la necesidad de crear sistemas que integren los datos provenientes de diferentes fuentes para realizar un análisis eficiente de los mismos.

Un flujo de datos se define como una secuencia potencialmente infinita de instancias, que llegan al sistema a lo largo del tiempo, a una velocidad que puede ser variable [115]. Estas características suponen una gran diferencia respecto a la minería de datos tradicional. Las más relevantes son [33], [116]:

- No es posible almacenar la totalidad del flujo de datos al ser potencialmente infinito. Por lo que únicamente se pueden almacenar de manera temporal, limitando así la cantidad de veces que puede ser procesado un dato. Normalmente se limita a una única visualización y, tras procesarlo, se descarta.

- Las instancias no están disponibles para el sistema en su totalidad de antemano, sino que estas llegan al sistema a lo largo del tiempo de manera secuencial o mediante bloques o conjuntos de datos ordenados. De hecho, habitualmente este orden implica cierta dependencia entre un dato y el siguiente.

- Se asume que la velocidad de llegada de datos es muy alta con respecto a la capacidad de procesamiento del sistema. Por lo tanto, los métodos deben dar respuesta inmediata con el objetivo de evitar demoras y comprometer la estabilidad del sistema.

- Las características estadísticas de los datos que llegan al sistema pueden evolucionar a lo largo del tiempo, haciendo que los modelos aprendidos no sean válidos. A este proceso se le denomina en la literatura como cambio de concepto [117].

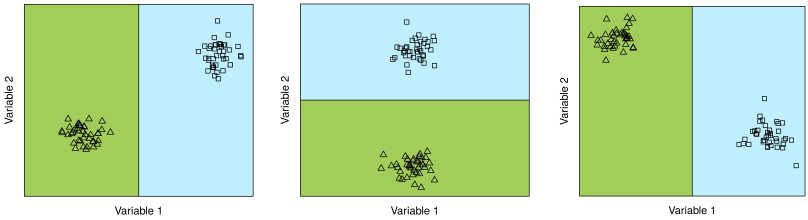
Con estas características, una mayoría amplia de algoritmos tradicionales de minería de datos no son válidos para el tratamiento de flujos de datos. Esto se debe a que asumen la disponibilidad de la totalidad de los datos, tienen un tiempo de cómputo excesivo o no asumen la variabilidad en la distribución estadística de los datos a lo largo del tiempo, también denominado como cambio de concepto.

El cambio de concepto es uno de los factores más analizados en la literatura especializada en minería de flujo de datos. Este fenómeno se define como cambios en la distribución que subyace en los datos a lo largo del tiempo. La presencia de este tipo de cambio supone que las propiedades aprendidas por un algoritmo de minería de datos se degradan cuando este se produce. En muchos casos, el cambio es lo suficientemente grande como para que la aplicación del modelo no tenga sentido pues su calidad se ha degradado significativamente. Por lo tanto, aplicar técnicas de detección y manejo de cambio de concepto es fundamental para un correcto análisis de los flujos de datos.

Hay varios aspectos inherentes al cambio de concepto que son necesarios tener en cuenta cuando se analiza su naturaleza [117]:

- Tipo de cambio. En este aspecto, se pueden dar dos tipos de cambio de concepto principales, representados gráficamente en la Figura 1.5:

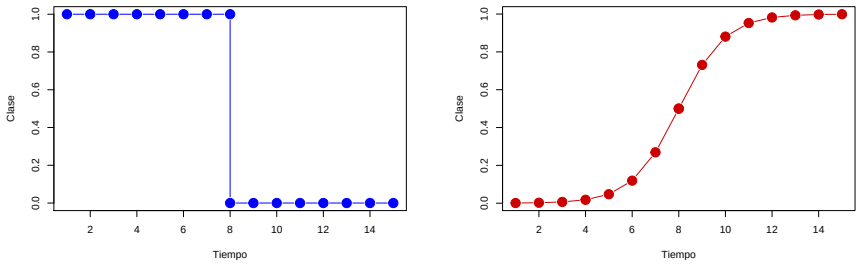
  - Cambio de concepto real. Ocurre cuando se produce un cambio en la probabilidad a posteriori de la clase del patrón X, es decir $P(v_c|X)$. Esta definición, llevada al ámbito SDRD y a la nomenclatura utilizada en la Tabla 1.1, implica que $P(v_c|X) = \frac{P(v_c X)}{P(X)} = \frac{tp}{tp+fp} = Conf(X)$. Esto afecta por tanto a la frontera de decisión de los modelos pues ha habido una modificación en la distribución que subyace en los datos lo suficientemente grande como para que ya no sean válidas. Por tanto, la confianza (Ecuación (1.4)) de un patrón se ve alterada en este tipo de cambio de concepto.
  - Cambio de concepto virtual. Se produce cuando cambia la probabilidad condicional de la clase $P(X|v_c)$, sin afectar a la probabilidad a posteriori $P(v_c|X)$. Esto implica que la distribución de los datos ha cambiado, pero la frontera de decisión no, por lo que el modelo seguiría siendo válido. Siguiendo la nomenclatura de la Tabla 1.1, $P(X|v_c) = \frac{P(v_c X)}{P(v_c)} = \frac{tp}{tp+fn} = TPR(X)$. No obstante, este tipo de cambio de concepto es interesante en tareas descriptivas como SDRD pues se puede describir un suceso que ha modificado el comportamiento de los datos.



(a) Datos originales. (b) Cambio real. (c) Cambio virtual.

Figura 1.5: Tipos principales de cambio de concepto en función de su influencia en las fronteras de decisión del modelo. *Fuente: elaboración propia a partir de [118].*

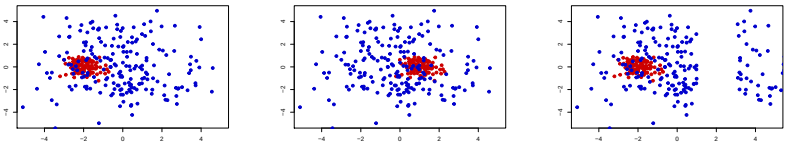
- Velocidad. En este punto se distingue entre cambios súbitos, es decir, que ocurren abruptamente de una instancia o bloque al siguiente o aquellos que se producen de manera gradual con mayor o menor velocidad a lo largo de una serie de instancias. Al primer tipo de cambio se lo conoce en la literatura como cambio abrupto, mientras que el segundo es conocido como cambio gradual. Un ejemplo ilustrativo de ambos tipos de cambio se muestra en la Figura 1.6.



(a) Cambio abrupto. (b) Cambio gradual.

Figura 1.6: Velocidad del cambio de concepto. *Fuente: elaboración propia a partir de [118].*

- Severidad. Un cambio de concepto puede afectar a una zona reducida del espacio de búsqueda, denominado cambio local, o puede afectar a todo el espacio de búsqueda, siendo así un cambio global. Un ejemplo ilustrativo de ambos tipos de cambio se muestra en la Figura 1.7.



(a) Datos originales. (b) Cambio Global. (c) Cambio Local.

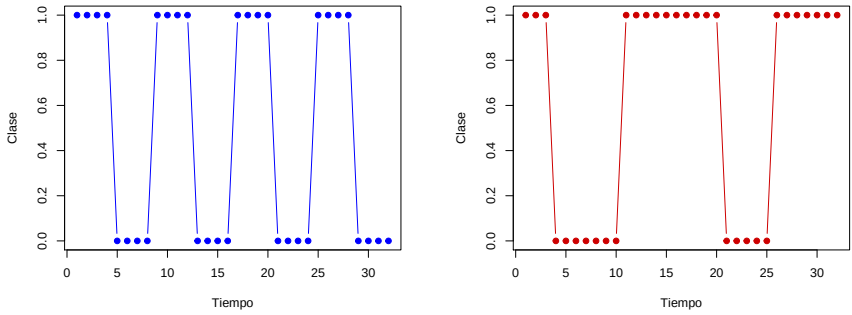
Figura 1.7: Severidad del cambio de concepto. *Fuente: elaboración propia a partir de [117].*

- Recurrencia. Un concepto puede ser cíclico si a intervalos regulares se produce el cambio, como por ejemplo aquellas aplicaciones sujetas a variaciones estacionales, o bien pueden ser cambios acíclos, cuando se desconoce la repetición del concepto. Un ejemplo ilustrativo de ambos tipos de cambio se muestra en la Figura 1.8.

Para gestionar el cambio de concepto, a lo largo de la literatura se han empleado tres enfoques diferentes [118]:

- Entrenar el modelo desde cero cada vez que llegue una nueva instancia o bloque de datos. Este enfoque tiene un alto coste computacional y es inviable en la mayoría de escenarios [119]-[121].

- Entrenar el modelo desde cero cuando se detecte un cambio de concepto. Para ello, se utilizan detectores de cambio de concepto que pueden ser dependientes o no del modelo de aprendizaje. Estos sistemas mandan una señal de alerta e informan sobre la severidad del cambio al algoritmo de aprendizaje. Este punto de vista permite aligerar la carga computacional siempre y cuando el entorno no sea altamente cambiante [122]-[125].



(a) Cambio cíclico.

(b) Cambio acíclico.

Figura 1.8: Recurrencia del cambio de concepto. *Fuente: elaboración propia a partir de [118].*

- Usar un método de aprendizaje adaptativo que permite seguir el estado del flujo de datos y adaptarse automáticamente al mismo. En este punto, existen diferentes enfoques que se han utilizado a lo largo de la literatura, entre los que destacan:

  - Ventanas deslizantes. Se define como una o varias memorias que almacenan los datos más recientes, o resúmenes sobre los mismos, que pertenecen al concepto actual. Ha sido uno de los métodos más utilizados a lo largo de la literatura gracias a su versatilidad y bajo consumo de recursos [126]-[131].

  - Métodos *online*. Estos métodos actualizan el modelo instancia a instancia, por lo que son capaces de adaptarse a los cambios tan pronto como ocurren. Es importante destacar que algunos métodos de clasificación como las redes neuronales o Naïve Bayes pueden funcionar de esta manera. Sin embargo, diseñar este tipo de métodos es complejo.

- Uso de *ensembles*. Aglutina diferentes métodos de aprendizaje de modo que pueden adaptarse sencillamente modificando la estructura de los diferentes algoritmos de los que se compone. Esto permite ganancias tanto en calidad como en flexibilidad [132].

Actualmente existe un gran interés por la minería de flujos de datos. Esto se observa en el número de trabajos que han aparecido recientemente, en donde se pueden encontrar trabajos para tareas de reglas de asociación [133], mediante *ensembles* [134]-[136], *deep learning* [137], *big bata* [138], [139], aplicaciones [140], [141], entre otros.

Sin embargo, dentro de la comunidad investigadora aún no se han propuesto métodos para aprovechar el potencial que posee SDRD en general, y en especial EPM, a los flujos continuos de datos. En esta tesis se propone el desarrollo de sistemas difusos evolutivos capaces de extraer patrones emergentes en flujos continuos de datos para explicar el fenómeno que subyace en los mismos. El reto en este aspecto es afrontar los problemas de dimensionalidad, variedad y carácter dinámico de los datos, haciendo especial hincapié en la necesidad de abordar correctamente la evolución de los conceptos.

2
Discusión de los resultados

En este capítulo se resumen las propuestas llevadas a cabo para satisfacer los objetivos planteados en el capítulo introductorio. Para cada uno de los diferentes trabajos presentados en esta memoria, se muestra un breve resumen de las principales ideas propuestas, así como una discusión de los resultados obtenidos. La investigación realizada para esta tesis doctoral y los resultados asociados se agrupan en las siguientes publicaciones en revistas internacionales indexadas en JCR:

- Trabajos relacionados con SD:
  - J. Luengo, A. M. García-Vico, M. D. Pérez-Godoy y C. J. Carmona, «The influence of noise on the evolutionary fuzzy systems for subgroup discovery,» *Soft Computing*, vol. 20, n.º 11, págs. 4313-4330, 2016. DOI: 10.1007/s00500-016-2300-1, IF (JCR 2016): 2.472, Ranking: 33/105 (Computer Science, Interdisciplinary Applications), Cuartil: Q2.
  - A. M. García, F. Charte, P. González, C. J. Carmona y M. J. del Jesús, «Subgroup Discovery with Evolutionary Fuzzy Systems in R: The SDEFSR Package,» *The R Journal*, vol. 8, n.º 2, págs. 307-323, 2016. DOI: 10.32614/RJ-2016-048, IF (JCR 2016): 1.075, Ranking: 55/124 (Statistics & Probability), Cuartil: Q2.

- J. M. Luna, C. J. Carmona, A. M. García-Vico, M. J. del Jesus y S. Ventura, «Subgroup Discovery on Multiple Instance Data,» *International Journal of Computational Intelligence Systems*, vol. 12, n.º 2, págs. 1602-1612, 2019. DOI: 10.2991/ijcis.d.191213.001, IF (JCR 2018): 2.153, Ranking: 55/106 (Computer Science, Interdisciplinary Applications), Cuartil: Q3.

- Trabajos relacionados con EPM:

  - A. M. García-Vico, C. J. Carmona, D. Martín, M. García-Borroto y 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, n.º 1, e1231, 2018. DOI: 10.1002/widm.1231, IF (JCR 2018): 2.541, Ranking: 26/105 (Computer Science, Theory & Methods), Cuartil: Q1.
  - A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «MOEA-EFEP: Multi-Objective Evolutionary Algorithm for Extracting Fuzzy Emerging Patterns,» *IEEE Transactions on Fuzzy Systems*, vol. 26, n.º 5, págs. 2861-2872, 2018. DOI: 10.1109/TFUZZ.2018.2814577, IF (JCR 2018): 8.759, Ranking: 6/134 (Computer Science, Artificial Intelligence), Cuartil: Q1.
  - A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «A Big Data Approach for Extracting Fuzzy Emerging Patterns,» *Cognitive Computation*, vol. 11, n.º 3, págs. 400-417, 2019. DOI: 10.1007/s12559-018-9612-7, IF (JCR 2018): 4.287, Ranking: 25/134 (Computer Science, Artificial Intelligence), Cuartil: Q1.
  - A. M. García-Vico, C. J. Carmona, P. González, H. Seker y M. J. del Jesus, «FEPDS: A proposal for the Extraction of Fuzzy Emerging Patterns in Data Streams,» *IEEE Transactions on Fuzzy Systems*, Submitted (Major revision). DOI: No disponible, IF (JCR 2018): 8.759, Ranking: 6/134 (Computer Science, Artificial Intelligence), Cuartil: Q1.

Este capítulo se organiza de la siguiente manera: En primer lugar, el Apartado 2.1 presenta un estudio sobre la influencia del ruido en los sistemas difusos evolutivos desarrollados para SD. A continuación, el Apartado 2.2 muestra las ideas principales del desarrollo de un paquete para la herramienta R que contiene los principales sistemas difusos evolutivos para SD. En el Apartado 2.3 se realiza un análisis sobre el comportamiento de los principales enfoques para SD

en problemas de aprendizaje multi-instancia. Después, el Apartado 2.4 presenta una revisión bibliográfica sobre la EPM desde el punto de vista descriptivo. Tras esto, se desarrolla un nuevo sistema difuso evolutivo para la extracción de patrones emergentes descriptivos en el Apartado 2.5. A continuación, se propone un nuevo enfoque evolutivo para la extracción de patrones emergentes en entornos *big data* en el Apartado 2.6. Finalmente, esta tesis doctoral se cierra con el desarrollo de un nuevo sistema difuso evolutivo para la extracción de patrones emergentes en flujos continuos de datos en el Apartado 2.7.

2.1. Influencia del ruido en sistemas difusos evolutivos para descubrimiento de subgrupos

En la mayoría de aplicaciones reales existen corrupciones de datos que perjudican el análisis, la interpretación y las decisiones tomadas a partir de los modelos obtenidos de estos datos. Estas alteraciones, conocidas como ruido, son especialmente graves en problemas supervisados ya que se desfigura la relación que existe entre la entrada y la salida del sistema. Por lo tanto, en la literatura existe un gran interés por la detección y tratamiento del ruido en problemas de clasificación y regresión [147]. Como SD es una tarea que utiliza aprendizaje supervisado, el efecto negativo de este ruido puede producirse. Sin embargo, en aplicaciones descriptivas este problema no ha sido estudiado en profundidad. En este trabajo se analiza el efecto del ruido en los algoritmos basados en sistemas difusos evolutivos para SD y cómo se pueden aliviar los efectos del mismo utilizando los enfoques de manejo del ruido más apropiados.

En concreto, uno de los enfoques más populares es el empleo de filtros [148], [149] ya que estos actúan como una fase de preprocesamiento eliminando aquellas instancias identificadas como ruido. Por lo tanto, el algoritmo original no necesita ser modificado. Este hecho, que es especialmente relevante para clasificación o regresión, no está claro en SD pues aún no se ha analizado en profundidad el efecto de la eliminación de instancias con las métricas que se emplean en SDRD.

El estudio experimental se realizó sobre una batería de 37 conjuntos de datos del repositorio de datos de KEEL[1]. A estos datos, se les introdujo ruido sobre la variable objetivo de manera de artificial utilizando el método *uniform* [150]. Siguiendo este proceso, se crearon diferentes niveles de ruido: 0 %, 5 %, 10 %, 15 % y 20 %, para determinar la influencia del mismo en los algoritmos.

Los algoritmos de SD utilizados fueron SDIGA, NMEEFSD y FuGePSD. Es importante destacar que todos ellos son sistemas difusos evolutivos, ya que el objetivo de este trabajo consiste en comprobar la robustez de los sistemas basados en lógica difusa y cómo son capaces de recuperarse en entornos con ruido. Por su parte, los filtros de ruido empleados fueron: *ensemble filter* [148], *cross-validated committees filter* [151] e *iterative-partitioning filter* [149]. En este estudio la calidad de las reglas se midió desde tres perspectivas diferentes: interés, mediante la métrica WRAcc; generalidad, usando TPR; y fiabilidad, utilizando la medida confianza. Además, se analizó la interpretabilidad de los modelos extraídos en función del número de patrones y variables. Los resultados fueron comparados entre sí utilizando pruebas estadísticas.

De acuerdo a los resultados extraídos en la experimentación, se observa que el ruido en este tipo de algoritmos influye de manera negativa en los resultados obtenidos a pesar de la robustez que proporciona la lógica difusa en este tipo de escenarios. No obstante, esta influencia se puede mitigar empleando algoritmos de filtrado de ruido. De acuerdo a los resultados, el *ensemble filter* reporta mejoras en la calidad de los algoritmos de SD en todos los aspectos que se han analizado. Por lo tanto, es interesante que se aplique este tipo de algoritmos como preprocesamiento en aplicaciones reales. En contraposición, se destaca que el algoritmo FuGePSD es capaz de detectar el ruido y aislarlo, de modo que no es necesario la aplicación de ningún tipo de filtro adicional. Esto abre la posibilidad de continuar en la línea de desarrollo de métodos robustos frente a ruido.

En cuanto a la interpretabilidad de los modelos, de este estudio se extrae que la aplicación de filtros de ruido mejora ligeramente la interpretabilidad de los mismos. Se destaca además que la aplicación de estos filtros parece tener una mayor mejora en aquellos algoritmos que tienen propensión a obtener modelos con muchos patrones simples, es decir, con pocas variables. Esto se debe a la eliminación de instancias en las fronteras entre clases, permitiendo así una mejor detección mediante patrones simples.

[1] https://sci2s.ugr.es/keel/datasets.php

Como conclusión final de este trabajo, se destaca la necesidad de seguir investigando en esta línea de trabajo desde el desarrollo de nuevos métodos robustos frente a ruido, así como en la creación de técnicas de preprocesamiento enfocadas al tratamiento de ruido específicas para SD.

El trabajo de investigación asociado a esta parte es:

- J. Luengo, A. M. García-Vico, M. D. Pérez-Godoy y C. J. Carmona, «The influence of noise on the evolutionary fuzzy systems for subgroup discovery,» *Soft Computing*, vol. 20, n.º 11, págs. 4313-4330, 2016.
  DOI: 10.1007/s00500-016-2300-1, IF (JCR 2016): 2.472, Ranking: 33/105 (Computer Science, Interdisciplinary Applications), Cuartil: Q2.

2.2. Descubrimiento de subgrupos con sistemas difusos evolutivos en R: el paquete SDEFSR

A día de hoy, existen para SD varias herramientas disponibles para el empleo de los algoritmos más destacados de la literatura. El software más conocido actualmente que contienen paquetes de algoritmos de SD son KEEL [59], VIKAMINE [152], ORANGE [60] y CORTANA [153].

R[2] es un lenguaje de programación bajo licencia GNU para computación estadística y generación de gráficas que provee una amplia variedad de funcionalidades estadísticas, así como una amplia variedad de algoritmos de la mayoría de campos de la minería de datos actual. R se podría considerar a día de hoy como una de las aplicaciones software más potentes para el análisis y visualización de datos. Uno de los principales causantes de su éxito es la posibilidad de ampliar su funcionalidad a través de paquetes software que la comunidad crea y distribuye libremente.

A pesar de su éxito, en R únicamente existe un paquete relacionado con métodos de SD, denominado *rsubgroup*[3]. Este paquete se trata de una interfaz para R del programa VIKAMINE, el cual contiene algoritmos de SD con enfoques clásicos. Por lo tanto, era necesario llevar a esta herramienta aquellos algoritmos de SD basados en sistemas difusos evolutivos más relevantes debido a la calidad del conocimiento que extraen [17].

[2] https://cran.r-project.org/
[3] http://www.rsubgroup.org/

Para ello, se ha realizado un paquete para el software R denominado SDEFSR[4] (*Subgroup Discovery with Evolutionary Fuzzy Systems for R*). En este paquete se han implementado cuatro algoritmos de SD basados en sistemas difusos evolutivos: SDIGA [50], MESDIF [51], NMEEF-SD [52] y FuGePSD [53]. Estos algoritmos son totalmente configurables, aunque se proponen valores por defecto para facilitar su uso. El objetivo es proveer a la comunidad de R con los algoritmos basados en sistemas difusos evolutivos para SD más relevantes de la literatura.

Los algoritmos incluidos en este paquete son capaces de leer datos desde diferentes formatos: ARFF, de la herramienta Weka [154], CSV, KEEL, así como desde la estructura de datos *data.frame* propia de R, permitiendo ampliar el espectro de formatos utilizando paquetes externos a SDEFSR.

Tras la ejecución de un método de SD en SDEFSR, se puede realizar un análisis y post-procesamiento de los resultados obtenidos. Esto es posible gracias a que el paquete proporciona una utilidad para visualizar los patrones en un gráfico mostrando su calidad respecto a las medidas TPR y FPR para determinar su fiabilidad y generalidad. Asimismo, el paquete aporta la funcionalidad suficiente para realizar operaciones básicas de post-procesamiento, como el filtrado por una medida de calidad o número de reglas de manera sencilla.

Finalmente, el paquete SDEFSR incluye una aplicación web que permite realizar todo el procedimiento de SD, así como una visualización básica de datos de manera gráfica. Un ejemplo de este interfaz se puede observar en la Figura 2.1. El objetivo es acercar toda la funcionalidad que ofrece este paquete a aquellas personas con muy poca experiencia de programación, o aquellas que prefieren hacer el proceso de manera gráfica.

A fecha de hoy, el impacto del paquete en la comunidad de R es aceptable, con un total de 5382 descargas[5] desde su lanzamiento en 2016. Actualmente se sigue trabajando activamente en la mejora del mismo, ampliando su funcionalidad y mejorando la eficiencia para producir un mayor impacto en la comunidad de R.

Como conclusión final, se destaca que el estudio de métodos y herramientas, junto al desarrollo de este paquete software, ha permitido realizar una profunda revisión bibliográfica sobre la tarea. Esta revisión ha permitido identificar las diferentes propuestas publicadas y conocer sus limitaciones, permitiendo detectar problemas abiertos para trabajos de investigación futuros.

El trabajo de investigación asociado a esta parte es:

[4]https://cran.r-project.org/web/packages/SDEFSR/index.html
[5]https://cranlogs.r-pkg.org/badges/grand-total/SDEFSR

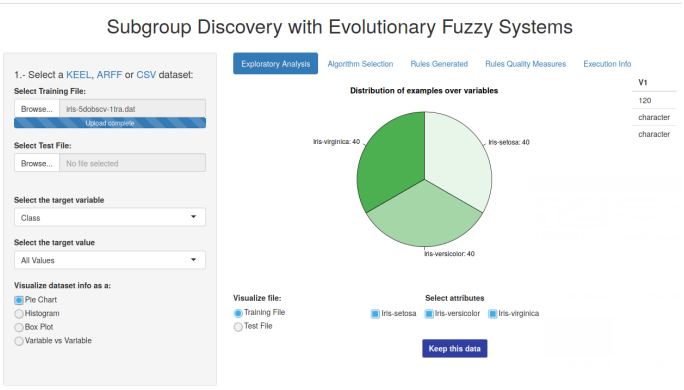


Figura 2.1: Ejemplo de la interfaz web para el empleo del paquete SDEFSR. Detalle del empleo del visualizador básico de datos. *Fuente: elaboración propia.*

- A. M. García, F. Charte, P. González, C. J. Carmona y M. J. del Jesús, «Subgroup Discovery with Evolutionary Fuzzy Systems in R: The SDEFSR Package,» *The R Journal*, vol. 8, n.º 2, págs. 307-323, 2016. DOI: 10.32614/RJ-2016-048, IF (JCR 2016): 1.075, Ranking: 55/124 (Statistics & Probability), Cuartil: Q2.
- A. M. García-Vico, F. Charte, P. González, C. J. Carmona y M. J. del Jesus, «Usando Algoritmos de Descubrimiento de Subgrupos en R: El Paquete SDR,» en *Proc. of the 16th Conference of the Spanish Association for Artificial Intelligence*, 2015, págs. 739-748.

2.3. Análisis de los principales enfoques para descubrimiento de subgrupos en problemas de aprendizaje multi-instancia

Debido a la creciente demanda de almacenamiento, a día de hoy el almacenaje de los mismos se realiza de muy diversas maneras. Una de ellas es la de agrupar varias instancias que de alguna manera se encuentran relacionadas por la misma causa en bolsas. Usando esta representación, conocida en la literatura como multi-instancia [105], la variable objetivo se asocia directamente con la

bolsa de instancias y no con cada una de las instancias individuales que la conforman. Este tipo de problemas han sido ampliamente estudiados por la comunidad de aprendizaje supervisado, en especial para problemas de clasificación debido a su utilidad [107]. Sin embargo, como resultado del trabajo de revisión realizado anteriormente, se descubrió que no existe ningún algoritmo específicamente desarrollado para la tarea de SD enfocado a resolver problemas MIL.

Por lo tanto, el objetivo de este trabajo es una primera aproximación a los problemas MIL en SD. Para ello, se realiza una adaptación de los enfoques más relevantes en SD a problemas MIL. En concreto, se utiliza un algoritmo exhaustivo, SD-Map [47], y dos algoritmos no exhaustivos: CGBA-SD [49] y NMEEF-SD [52]. SD-Map se basa en el conocido algoritmo FP-Growth [156], mientras que CGBA-SD se basa en programación genética y NMEEF-SD se basa en un sistema difuso evolutivo multi-objetivo.

En [107] se presenta una revisión de los paradigmas más destacados que se han utilizado en MIL. Desde un punto de vista descriptivo, los enfoques basados en el espacio de bolsa o en el espacio embebido no son relevantes pues se pierde la información relacionada con las instancias. Por esta razón, los algoritmos de SD se modifican basándose en el paradigma del espacio de instancia. A modo de resumen, la adaptación llevada a cabo en estos algoritmos se encuentra en el conteo de características o instancias cubiertas realizado para el cálculo de las medidas de calidad. En concreto, se emplea la asunción estándar del problema multi-instancia. Por un lado, el conteo de características se incrementará solo una vez por bolsa si dicha característica aparece al menos una vez en la bolsa. Por otro lado, una bolsa será marcada como cubierta por un patrón si al menos una de sus instancias es cubierta.

Del estudio experimental llevado a cabo, se observa que el enfoque exhaustivo produce unos resultados bastante pobres en calidad. Por su parte, CGBA-SD muestra un mejor comportamiento en general, presentando unos resultados de calidad competentes. Finalmente, el algoritmo NMEEF-SD presenta un comportamiento más variado, pues obtiene resultados de baja calidad en problemas de identificación de contenido en imágenes, mientras que en el resto se obtienen unos resultados interesantes y precisos.

Como conclusión, el trabajo realizado consiste en una primera aproximación a MIL de los principales algoritmos de SD basados en diferentes enfoques de extracción de patrones. Por tanto, este trabajo aporta a la sociedad una nueva posibilidad de resolución de problemas anteriormente no disponible. Asimismo,

el aporte realizado a la comunidad científica en este trabajo destaca por la posibilidad de continuar investigando en esta nueva línea de trabajo tras la obtención de resultados prometedores.

El trabajo de investigación asociado a esta parte es:

- J. M. Luna, C. J. Carmona, A. M. García-Vico, M. J. del Jesus y S. Ventura, «Subgroup Discovery on Multiple Instance Data,» *International Journal of Computational Intelligence Systems*, vol. 12, n.º 2, págs. 1602-1612, 2019. DOI: 10.2991/ijcis.d.191213.001, IF (JCR 2018): 2.153, Ranking: 55/106 (Computer Science, Interdisciplinary Applications), Cuartil: Q3.

2.4. Revisión de la minería de patrones emergentes desde el punto de vista descriptivo

EPM tiene como objetivo la búsqueda de patrones discriminatorios cuyo soporte entre diferentes clases o conjuntos de datos difiera de manera significativa [20]. De este modo, se pueden describir las características discriminadoras entre las diferentes clases de un problema o el descubrimiento de tendencias emergentes a lo largo del tiempo. Dado que EPM se basa principalmente en el empleo de la medida GR (Ecuación (1.9)), estos poseen un alto poder discriminatorio. Por esta razón, a lo largo de la literatura se han empleado y desarrollado modelos de extracción de patrones únicamente enfocados a maximizar la precisión de los modelos, ignorando en gran medida el carácter descriptivo que poseen los patrones.

En este trabajo se realiza una revisión y análisis de los principales enfoques desarrollados para EPM desde el punto de vista descriptivo. Para ello, se analiza el comportamiento de los principales algoritmos desarrollados a lo largo de la literatura. También se analiza la influencia de diferentes mecanismos de filtrado de patrones para la mejora de las capacidades descriptivas de los modelos extraídos. El objetivo es el establecimiento de una línea a seguir para el desarrollo de futuras propuestas enfocadas al aspecto descriptivo.

Fruto del trabajo de revisión realizado, se han identificado tres subconjuntos interesantes de patrones emergentes en función de sus características: patrones minimales y maximales, patrones *jumping* y patrones tolerantes a ruido. Además, se ha llevado a cabo una taxonomía de algoritmos de extracción de patrones emergentes en donde se han identificado cuatro enfoques principales: (1) basados en límites, (2) basados en árboles, (3) basados en árboles de decisión y (4)

basados en sistemas difusos evolutivos. Es importante destacar la evolución que han sufrido estos enfoques con respecto a las capacidades descriptivas de los modelos, en donde los primeros enfoques ignoran totalmente el aspecto descriptivo, mientras que el tercer y cuarto enfoque sí lo tienen en cuenta e introducen diversos mecanismos, como el empleo de lógica difusa, entre otros, para mantener cierto nivel de interpretabilidad.

Para la obtención de conclusiones fundamentadas, se ha realizado un estudio experimental con un amplio número de conjuntos de datos, así como el empleo de pruebas estadísticas para determinar el nivel de significación de las diferencias. Este procedimiento se ha llevado a cabo sobre cinco medidas de calidad, permitiendo determinar las características descriptivas de los patrones: fiabilidad, generalidad e interpretabilidad.

De los resultados obtenidos, se pueden obtener las siguientes conclusiones:

- Los patrones tipo Chi son muy interesantes desde el punto de vista descriptivo. Esto se debe principalmente al empleo de un umbral de soporte mínimo junto a la utilización de la prueba χ^2, el cual permite determinar si todos los selectores de un patrón aportan información de manera significativa. Sin embargo, a pesar de sus grandes cualidades, el alto número de restricciones hace que sea muy difícil extraer este tipo de patrón en problemas complejos.

- El post-procesamiento de los patrones extraídos es necesario en la mayoría de los casos para mejorar las cualidades descriptivas. Por lo tanto, es interesante integrar en futuros desarrollos estos mecanismos para mejorar las cualidades descriptivas. En concreto, se destaca la búsqueda de patrones minimales con altos niveles de confianza.

- El empleo de lógica difusa es vital para la mejora de los modelos en el enfoque descriptivo. Su uso permite obtener altos niveles de calidad en todos los aspectos analizados, mejorando la interpretabilidad de los modelos. Por lo tanto, se anima a su empleo para futuros desarrollos.

- El enfoque basado en sistemas difusos evolutivos es capaz de obtener unos niveles de calidad desde el punto de vista descriptivo similares a los mejores métodos analizados, con un número de patrones significativamente menor que el resto. Este reciente enfoque de EPM es muy interesante desde el punto de vista descriptivo e invita al desarrollo de nuevos algoritmos basados en métodos evolutivos o en otras metaheurísticas.

- Se identificaron los principales métodos para la extracción de patrones emergentes descriptivos en la literatura basándose en los tres objetivos principales de SDRD: precisión, interés y generalidad. El resultado de este estudio, mostrado en la Figura 2.2 destaca la calidad de los algoritmos FEPM, iEPM y EvAEP, por tanto, es interesante de cara a futuros desarrollos el análisis profundo de estos algoritmos para mejorar sus capacidades descriptivas.

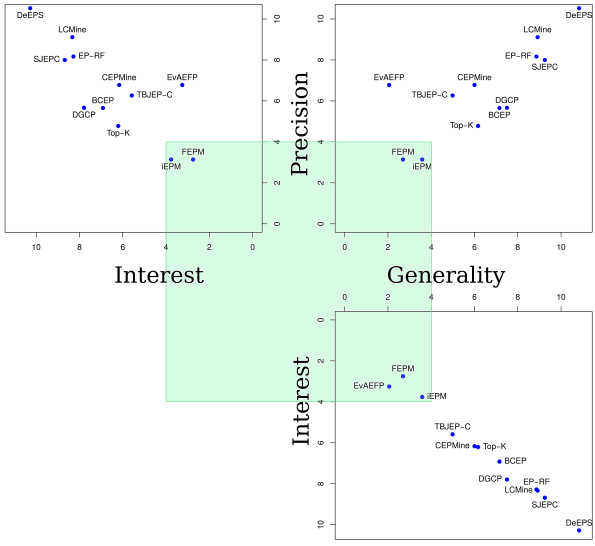


Figura 2.2: Comparación del ranking de Friedman de los diferentes algoritmos analizados en el estudio de revisión de EPM. La zona sombreada en verde indica la zona con mayor relevancia. *Fuente: [61].*

El trabajo de investigación asociado a esta parte es:

- A. M. García-Vico, C. J. Carmona, D. Martín, M. García-Borroto y 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, n.º 1, e1231, 2018. DOI: 10.1002/widm.1231, IF (JCR 2018): 2.541, Ranking: 26/105 (Computer Science, Theory & Methods), Cuartil: Q1.

- C. J. Carmona, F. J. Pulgar-Rubio, A. M. García-Vico, P. González y M. J. del Jesus, «Análisis descriptivo mediante aprendizaje supervisado basado en patrones emergentes,» en *Proc. of the 7th Simposio Teoría y Aplicaciones de Minería de Datos*, 2015, págs. 685-694.

- A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «Minería de Patrones Emergentes: Una oportunidad para la extracción evolutiva de conocimiento.,» en *Proc. of the 17th Conference of the Spanish Association for Artificial Intelligence*, 2016, págs. 149-159.

- A. M. García-Vico, C. J. Carmona y M. J. del Jesus, «Análisis de Diferentes Tipos de Reglas en Sistemas Difusos Evolutivos para Minería de Patrones Emergentes,» en *Proc. of the 12th Spanish Conference on Metaheuristics, Evolutive and Bioinspired Algorithms*, 2017, págs. 876-885.

Finalmente, como resultado adicional de la realización de este estudio se ha desarrollado la herramienta de código abierto EPM-Framework[6], que contiene los algoritmos de EPM utilizados en este estudio. El software está en constante desarrollo y tiene como objetivo permitir la aplicación de los principales algoritmos de extracción de patrones emergentes de una forma sencilla e intuitiva, permitiendo la expansión del mismo con nuevas propuestas fácilmente. Se espera que esta herramienta tenga un impacto importante en la comunidad de EPM, ampliando la visibilidad y aplicabilidad de estos modelos.

[6]https://github.com/SIMIDAT/epm-framework

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

Los resultados obtenidos en el trabajo anterior animan a seguir con el desarrollo de algoritmos basados en sistemas difusos evolutivos para la extracción de patrones emergentes, pues poseen un buen balance entre la fiabilidad de los patrones obtenidos, su generalidad y la interpretabilidad de los modelos. El único algoritmo que hasta la fecha estaba basado en un sistema difuso evolutivo era el algoritmo EvAEP [68], que utiliza un algoritmo evolutivo mono-objetivo.

La extracción de patrones descriptivos tiene en cuenta tres objetivos fundamentales y que son contrarios entre sí: fiabilidad, generalidad e interpretabilidad. Por lo tanto, un enfoque multi-objetivo resulta adecuado para esta tarea ya que nos permitirá obtener aquellos patrones con mejor balance entre los diferentes objetivos, es decir, los patrones que se encuentren en el frente de Pareto.

En este trabajo se presenta un sistema difuso evolutivo multi-objetivo denominado MOEA-EFEP (*Multi-Objective Evolutionary Algorithm for Extracting Fuzzy Emerging Patterns*) el cual introduce una serie de mecanismos que fomentan la extracción de patrones emergentes descriptivos, apoyándose en las conclusiones extraídas en el trabajo de revisión presentado anteriormente. En concreto, las principales características del método propuesto se presentan a continuación:

- Sistema difuso evolutivo basado en ordenación por dominancia [160].
- Representación "cromosoma = patrón" donde un individuo de la población representa un potencial patrón. Se permite además el uso represantación canónica o DNF en función de las necesidades del usuario.
- Esquema cooperativo-competitivo para la extracción del conjunto final de patrones. Se hace uso de una población élite en donde los individuos cooperarán para obtener un conjunto de patrones cuyo valor WRAcc medio sea máximo.
- Dicha población élite se actualiza mediante el operador de *token competition* [161] si el valor WRAcc medio de la población tras aplicar este operador es mejor que el de la población élite actual. Este operador permite la obtención de aquellos patrones de mayor calidad, cubriendo el mayor espacio de búsqueda posible, evitando la redundancia y reduciendo el solapamiento entre patrones.

- Uso de operadores genéticos orientados a impulsar la generalidad de los patrones, como el operador de inicialización basado en patrones generales, una mutación orientada a la generalidad y un operador de reinicialización guiada para moverse a zonas del espacio de búsqueda aún no exploradas.

- Al finalizar el proceso evolutivo, se emplea un filtro a elegir por parte del usuario entre filtro por confianza, por patrones maximales o patrones minimales, de acuerdo al estudio de revisión presentado anteriormente.

El estudio experimental llevado a cabo en este trabajo tiene tres objetivos fundamentales: (1) determinar la mejor representación del conocimiento para el algoritmo propuesto, (2) determinar el tipo de filtro a aplicar tras finalizar el proceso evolutivo, y (3) comparar la calidad de los patrones extraídos frente a los mejores métodos según el estudio de revisión presentado anteriormente. Para ello, los análisis se han realizado utilizando 50 conjuntos de datos, empleando pruebas estadísticas para determinar el nivel de significación de las diferencias.

Las conclusiones que se extraen del estudio experimental llevado a cabo se presentan a continuación:

- Existe un mejor balance fiabilidad-generalidad a favor de la representación DNF respecto a la representación canónica, con unos niveles de interpretabilidad similares. Esto se puede deber a la mayor flexibilidad que ofrece este tipo de representación al permitir varios valores para una misma variable.

- La aplicación de un filtro por confianza para el algoritmo propuesto permite la mejora de manera significativa de la fiabilidad de los patrones extraídos, manteniendo los niveles de generalidad e interpretabilidad.

- Comparado con otros métodos, el algoritmo propuesto es capaz de aumentar el balance generalidad-fiabilidad aumentando de manera significativa la generalidad de los patrones extraídos, mientras que la fiabilidad se mantiene similar con respecto al resto de algoritmos. Además, la interpretabilidad de los patrones extraídos es mejor que el resto al devolver un número muy reducido de reglas. Estos resultados implican una extracción de patrones más interesantes por parte del algoritmo propuesto respecto al resto de métodos analizados.

- Finalmente, el tiempo de ejecución es de media el más rápido con respecto al resto de métodos, ya que su complejidad depende del número de variables e instancias del problema, mientras que el resto depende del número de selectores e instancias, siendo habitualmente el número de selectores mucho mayor que el de variables.

El trabajo de investigación asociado a esta parte es:

- A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «MOEA-EFEP: Multi-Objective Evolutionary Algorithm for Extracting Fuzzy Emerging Patterns,» *IEEE Transactions on Fuzzy Systems*, vol. 26, n.º 5, págs. 2861-2872, 2018. DOI: 10.1109/TFUZZ.2018.2814577, IF (JCR 2018): 8.759, Ranking: 6/134 (Computer Science, Artificial Intelligence), Cuartil: Q1.
- A. M. García-Vico, J. Montes, J. Aguilera, C. J. Carmona y M. J. del Jesus, «Analysing Concentrating Photovoltaics Technology through the use of Emerging Pattern Mining,» en *Proc. of the 11th International Conference on Soft Computing Models in Industrial and Environmental Applications*, 2016, págs. 1-8.

2.6. BD-EFEP: un enfoque big data para la extracción de patrones emergentes difusos

Uno de los principales problemas de los enfoques de extracción de patrones emergentes es la falta de escalabilidad. Esto es especialmente grave en entornos *big data*. Los sistemas basados en reglas difusas son relevantes para la comunidad ya que poseen robustez frente a problemas de escalabilidad [100], [162]. De hecho, se han propuesto en la literatura varios desarrollos de sistemas basados en reglas difusas para la extracción de conocimiento en entornos *big data* en diversas áreas de la minería de datos, como clasificación [163], [164], regresión [165], preprocesamiento [166], reglas de asociación [167], SD [55] y una primera aproximación para EPM [168].

En este trabajo se presenta un algoritmo para la extracción de patrones emergentes difusos en entornos *big data* denominado BD-EFEP (*Big Data approach for the Extraction of Fuzzy Emerging Patterns*). Este método utiliza un sistema difuso evolutivo multi-objetivo basado en un esquema cooperativo-competitivo que permite obtener una descripción precisa del problema, abarcando la mayor

cantidad del espacio de búsqueda posible. Para ello, el algoritmo propuesto se apoya en el empleo de operadores genéticos que fomentan la extracción de este tipo de patrones, como el operador de inicialización y mutación orientada, junto al empleo del operador de *token competition* y un filtro de confianza para mejorar la precisión de los patrones obtenidos.

La principal aportación realizada en este trabajo es el enfoque distribuido basado en MapReduce de la evaluación de los individuos de la población. Este proceso, representado en la Figura 2.3, se aplica únicamente en el momento de realizar la evaluación de los individuos de la población. Esto se debe a que el proceso de evaluación es la operación más costosa del algoritmo evolutivo, ya que para cada individuo hay que realizar un recorrido completo del enorme conjunto de datos. Esta evaluación distribuida funciona de la siguiente manera, para cada individuo: en la fase *map*, se calcula la tabla de contingencia relativa a la partición de datos a procesar. Es importante destacar que este proceso se realiza de manera distribuida. Tras su finalización, estas tablas se agrupan en la fase *reduce* sumando sus valores para obtener la tabla de contingencia final de la que se podrán calcular las diferentes medidas de calidad. Con este proceso, el método propuesto es capaz de obtener los mismos resultados independientemente del número de particiones de datos realizadas.

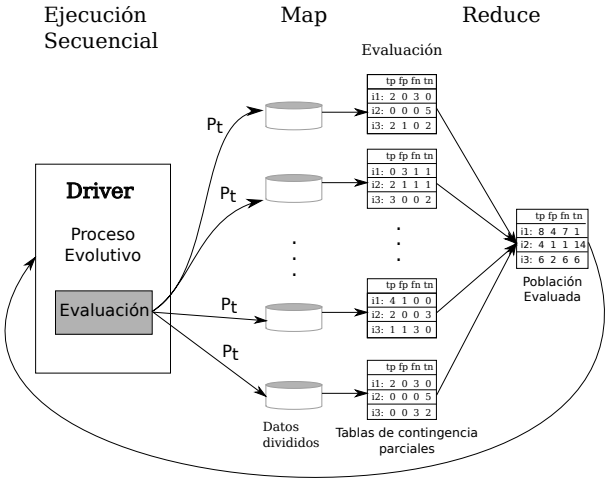


Figura 2.3: Procedimiento de evaluación basado en MapReduce del algoritmo BD-EFEP. *Fuente: elaboración propia a partir de [145].*

El estudio experimental llevado a cabo tiene dos objetivos: (1) determinar la calidad de los patrones extraídos por BD-EFEP frente a una adaptación para *big data* del algoritmo EvAEP [68], denominada EvAEP-Spark [168]; y (2) determinar la escalabilidad de la propuesta respecto a EvAEP-Spark.

Las conclusiones extraídas de este trabajo indican que el algoritmo propuesto obtiene un conjunto de patrones significativamente más preciso que EvAEP-Spark gracias a la aplicación del filtro de confianza final y al empleo de objetivos que promueven la fiabilidad. Sin embargo, la generalidad de estos patrones es peor. A pesar de ello, el balance generalidad-fiabilidad es mejor en BD-EFEP, por lo que los patrones extraídos son más interesantes para el experto. Respecto a la interpretabilidad, gracias al empleo del operador de *token competition* el método extrae bastantes patrones, pero simples. En EvAEP-Spark ocurre el resultado opuesto, se extraen menos patrones, pero más complejos. En este sentido, BD-EFEP mejora la interpretabilidad ya que simplifica de manera significativa el conocimiento obtenido.

Finalmente, la escalabilidad y tiempo de ejecución de BD-EFEP mejora significativamente a EvAEP-Spark. Además, se observa que el método escala de manera adecuada con respecto a la cantidad de particiones realizadas. No obstante, es importante destacar que el tiempo de ejecución para los algoritmos más grandes aún sigue siendo bastante elevado. Por lo tanto, aunque el enfoque MapReduce sea adecuado, estos resultados animan a seguir investigando en el desarrollo de mecanismos y métodos para la ejecución eficiente de algoritmos de extracción de patrones emergentes en entornos *big data*.

El trabajo de investigación asociado a este apartado es:

- A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «A Big Data Approach for Extracting Fuzzy Emerging Patterns,» *Cognitive Computation*, vol. 11, n.º 3, págs. 400-417, 2019. DOI: 10.1007/s12559-018-9612-7, IF (JCR 2018): 4.287, Ranking: 25/134 (Computer Science, Artificial Intelligence), Cuartil: Q1.

- A. M. García-Vico, P. González, M. J. del Jesus y C. J. Carmona, «A First Approach to Handle Emergining Patterns Mining on Big Data Problems: The EvAEFP-Spark Algorithm,» en *Proc. of the 2017 IEEE International Conference on Fuzzy Systems*, 2017, págs. 1-6.

- A. M. García-Vico, P. González, C. J. Carmona y M. J. del Jesus, «Impact of the type of rule in Fuzzy Emerging Pattern Mining on a Big Data Approach,» en *Proc. of the 2nd International Symposium on Fuzzy and Rough Sets*, 2017, págs. 1-10.

2.7. FEPDS: una propuesta para la extracción de patrones emergentes en flujos continuos de datos

Como se ha comentado anteriormente, el tipo de conocimiento extraído en EPM es interesante dentro del ámbito de la minería de flujo de datos ya que permite realizar decisiones de manera más sencilla, rápida y fiable. Sin embargo, como se ha visto en el Apartado 1.3.3 la extracción de conocimiento en este ámbito es un reto, debido entre otros, a fuertes restricciones en tiempo de respuesta, uso de memoria y actualización continua del modelo de conocimiento.

En este trabajo se presenta una primera aproximación al problema de extracción de patrones emergentes bajo el paradigma de la minería de flujo de datos, denominado FEPDS (*Fuzzy Emerging Patterns in Data Streams*). Los principales componentes del algoritmo FEPDS se describen a continuación:

- Se asume que los datos llegan al método en forma de bloques de datos de tamaño fijo. Para ello, FEPDS implementa un mecanismo que recoge y agrupa dichos datos provenientes del flujo.

- La adaptación del modelo al estado actual del flujo se lleva a cabo mediante una estrategia de adaptación ciega en donde el modelo se actualiza en cada bloque de datos. Así, se intenta minimizar los efectos de los problemas relacionados con el cambio de concepto

- Se emplea una memoria basada en ventana deslizante que almacena los conjuntos de patrones obtenidos recientemente por el algoritmo de aprendizaje. Esta estructura se utilizará para guiar el proceso de búsqueda hacia zonas del espacio que fueron prometedoras recientemente.

- La principal aportación de este trabajo se encuentra en el algoritmo de aprendizaje de FEPDS, el cual se basa en un sistema difuso evolutivo multiobjetivo basado en ordenación por dominancia cuyas principales características son:

- Representación "cromosoma = patrón" utilizando un esquema cooperativo-competitivo. La representación se lleva a cabo mediante un vector binario para permitir una representación tipo DNF de los patrones, agilizando la aplicación de operadores genéticos gracias al empleo de operaciones de bits.
- Uso de operadores genéticos orientados a la extracción de patrones con buen balance entre precisión y fiabilidad.
- Empleo de mecanismos de reinicio en caso de estancamiento para guiar el proceso de búsqueda hacia zonas prometedoras del espacio. Esto se realiza mediante la aplicación del operador de *token competition* usando la información incluida en la memoria basada en ventana deslizante descrita anteriormente.

Todos estos componentes se relacionan entre sí para obtener conocimiento de manera continua. El esquema de funcionamiento se encuentra gráficamente representando en la Figura 2.4. A modo de resumen, el algoritmo propuesto se basa en: primero, recolectar los datos provenientes del flujo de datos hasta formar un bloque. Dicho bloque se emplea a continuación para evaluar el modelo actual, si existe. Tras esto, se actualiza el modelo de conocimiento actual empleando el algoritmo de aprendizaje junto a los nuevos datos. Finalmente, el conjunto de patrones extraído se añade a la memoria basada en ventana deslizante que se utilizará dentro del proceso de aprendizaje para guiar la búsqueda.

El algoritmo FEPDS ha sido analizado en un amplio estudio experimental con la finalidad de determinar tres objetivos fundamentales: (1) la adaptabilidad del método a posibles cambios de concepto, (2) la calidad media del conocimiento extraído y (3) la escalabilidad y estabilidad del algoritmo a lo largo del tiempo. Finalmente, para demostrar la utilidad real del mismo, se ha llevado a cabo un caso de estudio en donde se describe de manera continua el perfil de los usuarios de los taxis de la ciudad de Nueva York durante el año 2017.

Las conclusiones extraídas del estudio llevado a cabo sugieren que el enfoque de adaptación propuesto, unido al empleo de lógica difusa, permite una gran robustez del sistema frente a cambios de concepto reales. Asimismo, la calidad del conocimiento extraído posee un ratio generalidad-fiabilidad muy interesante para el experto, junto a una interpretabilidad muy buena debido al bajo número de patrones y su simplicidad. Por último, el rendimiento del algoritmo respecto al tiempo de respuesta permite procesar datos a una velocidad aproximada de

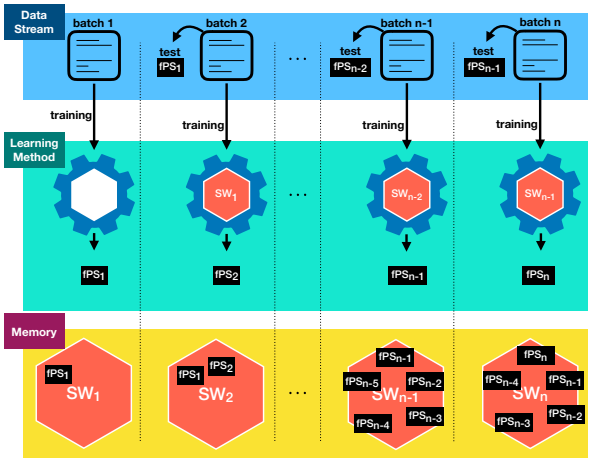


Figura 2.4: Esquema de funcionamiento general del algoritmo FEPDS. *Fuente: [146].*

hasta 5 KHz sin comprometer la estabilidad del sistema, por lo que hace que sea una alternativa viable para el procesamiento de datos en gran cantidad de aplicaciones.

Este hecho se ve reflejado en el caso de estudio abordado en este trabajo. En él, se muestra que la calidad media de los patrones extraídos es muy interesante para los expertos, pudiéndose realizar decisiones a corto plazo de manera sencilla. Además, la capacidad de FEPDS no se limita al corto plazo. Gracias a que es capaz de almacenar el conocimiento extraído, se pueden realizar análisis para decisiones a largo plazo. En concreto, en este estudio se analizan patrones recurrentes para la extracción de un perfil general de usuario a lo largo del año. Por lo tanto, el algoritmo FEPDS es una primera aproximación prometedora que anima a continuar esta línea de investigación en el futuro.

El trabajo de investigación asociado a esta parte es:

- A. M. García-Vico, C. J. Carmona, P. González, H. Seker y M. J. del Jesus, «FEPDS: A proposal for the Extraction of Fuzzy Emerging Patterns in Data Streams,» *IEEE Transactions on Fuzzy Systems*, Submitted (Major revision). DOI: No disponible, IF (JCR 2018): 8.759, Ranking: 6/134 (Computer Science, Artificial Intelligence), Cuartil: Q1.
- A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «Una primera aproximación para la extracción de patrones emergentes en flujos continuos de datos,» en *Proc. of the 18th Conferencia de la Asociación Española para la Inteligencia Artificial*, 2018, págs. 1093-1098.

3

Conclusiones y trabajos futuros

En este capítulo se resumen los resultados obtenidos en los trabajos presentados en esta memoria y se extraen las conclusiones del trabajo realizado. Asimismo, se comentan aspectos relacionados con los trabajos futuros que se pueden realizar siguiendo las líneas de investigación desarrolladas en esta memoria.

Tal y como se ha presentado a lo largo del presente estudio:

1. Para la técnica de SD existe un amplio abanico de desarrollos a lo largo de la literatura, pero muy pocos están enfocados a problemas complejos. De hecho, hasta donde llega nuestro conocimiento, no existe un método de SD enfocado a la extracción de conocimiento en problemas MIL ni en entornos de minería de flujo de datos. Para análisis de datos voluminosos, a pesar de existir métodos enfocados a este tipo de datos, no se ha analizado el efecto de factores externos que pueden influir en la calidad de los mismos, como es la presencia de ruido.

2. Las propiedades descriptivas de EPM son muy interesantes para el experto, pero el aspecto descriptivo de esta tarea no ha sido impulsado por parte de la literatura especializada. En este sentido, EPM se ha utilizado como un medio de clasificación en donde se maximiza la precisión, teniendo muy poco en cuenta el carácter descriptivo.

3. La revisión de la bibliografía del tema muestra que el desarrollo de métodos de minería de patrones emergentes para problemas complejos está muy poco desarrollado.

4. Finalmente, es importante señalar que las técnicas SDRD no se encuentran disponibles en los principales software de análisis de datos. En este sentido es interesante incluir este tipo de métodos dentro de estos entornos software para poder proporcionar una respuesta a la necesidad de analizar datos mediante este enfoque.

Para afrontar esta situación, la investigación realizada se centra principalmente en el análisis y desarrollo de soluciones algorítmicas con el objetivo de poder aplicar este tipo de técnicas en problemas complejos. En concreto, la tesis gira en torno a las tareas de SD y EPM aplicadas a este tipo de problemas.

3.1. Conclusiones relacionadas con descubrimiento de subgrupos

La primera parte de la investigación se orienta a SD. Las conclusiones de los trabajos realizados en este área se presentan a continuación:

- Una de las principales características de los problemas complejos es, entre otras, la diversidad de fuentes de donde provienen los datos. Esta diversidad aumenta la probabilidad de introducir ruido en el sistema, reduciendo su rendimiento. A pesar de que existen varios desarrollos basados en sistemas difusos evolutivos para SD, no existe un análisis formal de su robustez en SD. Para solucionar esto, se presenta un análisis acerca de la influencia de ruido en este tipo de métodos y la eficacia de aplicar filtros de ruido clásicos, orientados a clasificación, sobre los mismos. Los resultados destacan el funcionamiento de FuGePSD, el cual es capaz de aislar e ignorar el ruido, mientras que en el resto la aplicación de un filtro ayuda a paliar estos efectos negativos.

- Uno de los problemas complejos actualmente en auge son los problemas MIL, entre los cuales se encuentran, por ejemplo, problemas de segmentación de imágenes o síntesis de moléculas, entre otros. Este tipo de problemas son especialmente interesantes en SD, pero no pueden ser abordados por los enfoques actuales. Por ello, se presenta un estudio con la adaptación a problemas MIL de los principales enfoques desarrollados para SD. Los resultados muestran que los enfoques clásicos no presentan buena calidad, mientras que los enfoques basados en algoritmos evolutivos permiten obtener una mejor calidad en la mayoría de problemas. Por lo tanto, este trabajo abre una nueva línea de investigación enfocada al desarrollo de métodos SD aplicados a problemas MIL.
- Como resultado del estudio realizado de la literatura relacionada con SDRD para conocer todo lo realizado y la identificación de problemas abiertos, se analiza una gran cantidad de software específico que contiene los algoritmos más destacados de SD. No obstante, la falta de este tipo de paquetes de software en las herramientas de análisis de datos más populares, como R o Python, impide que el alcance de la tarea sea mucho más amplio. Para paliar este déficit, se desarrolla al inicio de esta tesis como resultado de la revisión bibliográfica realizada, el paquete SDEFSR. Este paquete contiene los principales algoritmos de SD basados en sistemas difusos evolutivos, así como herramientas específicas de procesamiento básico de patrones, junto a una interfaz web para su manejo.

En esta parte de la tesis doctoral se ha realizado un profundo trabajo de análisis de diferentes problemas de interés para la comunidad de SD. La principal conclusión obtenida es la capacidad de los algoritmos evolutivos en general, y de los sistemas difusos evolutivos en particular, para adaptarse a nuevos entornos y tipos de problemas manteniendo un buen nivel de calidad. De hecho, los trabajos realizados en SD abren nuevas líneas de investigación para la mejora de los problemas que se presentan en los mismos.

3.2. Conclusiones relacionadas con minería de patrones emergentes

La segunda parte del desarrollo de esta memoria de tesis está orientada a EPM. Las conclusiones extraídas de los trabajos realizados se indican a continuación:

- Uno de los primeros objetivos de esta parte es el análisis de los diferentes enfoques desarrollados a lo largo de la literatura para EPM desde el punto de vista descriptivo. El resultado de este estudio es una agrupación de los diferentes subconjuntos de patrones emergentes en función de ciertas características interesantes y una taxonomía de algoritmos en función del enfoque que tienen. Además, estos métodos son analizados desde el punto de vista descriptivo. Las conclusiones obtenidas destacan las capacidades descriptivas que presentan los patrones tipo Chi, y el empleo de lógica difusa para mejorar las características descriptivas de los patrones emergentes. Finalmente, el estudio refleja las altas capacidades descriptivas que poseen los enfoques basados en sistemas difusos evolutivos.

- A continuación, se desarrolla una propuesta para la obtención de sistemas difusos evolutivos que explote todas las características positivas para la extracción de patrones emergentes altamente descriptivos propuestas en el estudio anterior, el algoritmo MOEA-EFEP. Este método se basa en un algoritmo evolutivo multiobjetivo que mediante el empleo de operadores genéticos orientados específicamente a la extracción de patrones simples, y el empleo del operador de *token competition*, fomentan la extracción de patrones emergentes altamente descriptivos. Los resultados obtenidos demuestran un buen balance entre generalidad y fiabilidad de los patrones extraídos, sobrepasando significativamente a los principales enfoques desarrollados en la literatura. Por tanto, el algoritmo MOEA-EFEP es una interesante alternativa para la extracción de conocimiento de calidad para la tarea de EPM.

- Se ha propuesto el algoritmo BD-EFEP, un sistema difuso evolutivo para EPM enfocado a abordar problemas *big data*. La principal novedad de este método es el empleo de un procedimiento exacto para la gestión de datos a gran escala gracias a que se distribuye el proceso computacionalmente más costoso a lo largo del algoritmo: la evaluación de los individuos. Así, el método permite la obtención del mismo modelo de conocimiento independientemente de las particiones empleadas. Los resultados obtenidos demuestran por un lado la buena escalabilidad del método y un tiempo de ejecución aceptable. Además, se demuestra que el método obtiene un mejor balance entre generalidad y fiabilidad que otras propuestas. El enfoque propuesto es una alternativa interesante para la extracción de patrones emergentes de gran calidad en entornos *big data* en un tiempo de ejecución aceptable.

- Finalmente, se desarrolla el algoritmo FEPDS, que presenta un compromiso entre las diferentes características y restricciones de la minería de flujo de datos. Este método hace uso de un sistema difuso evolutivo muy eficiente para la extracción continua de patrones emergentes. Para ello, el algoritmo actualiza el modelo mediante un enfoque ciego, donde se añade el conjunto de patrones a la población inicial para ser actualizado, junto al empleo de ventanas deslizantes que permiten guiar la búsqueda a zonas prometedoras basándose en datos históricos. El enfoque propuesto es muy interesante para la extracción de conocimiento en este tipo de entornos, ya que extrae patrones de muy alta calidad, en un tiempo aceptable y tratando adecuadamente el cambio de concepto para evitar perder rendimiento.

El trabajo de investigación llevado a cabo para EPM en esta tesis ha permitido realizar un profundo avance en el conocimiento relacionado con las capacidades descriptivas de la tarea. Como resultado, se han identificado los tipos de patrones y enfoques de minería más interesantes desde este punto de vista. Además, se ha avanzado en la extracción de este tipo de conocimiento en problemas complejos como *big data* y minería de flujo de datos con resultados muy prometedores.

3.3. Trabajos futuros

En base al trabajo de investigación realizado en esta tesis, se proponen una serie de trabajos futuros. Estos se basan en la ampliación y mejora de las propuestas presentadas, así como en afrontar cuestiones sin resolver que han surgido durante el desarrollo de la investigación.

- Desarrollo de métodos aún más eficientes para SDRD en entornos de grandes volúmenes de datos. En este trabajo de investigación se ha presentado una propuesta para abordar el problema de extracción de este tipo de conocimiento en entornos *big data*. No obstante, es interesante seguir profundizando en esta tarea para la obtención de métodos aún más eficientes y con mayor escalabilidad.

- Desarrollo de técnicas para el análisis distribuido de flujo de datos. A día de hoy la cantidad de datos generados continuamente puede ser tan grande que es necesario realizar un procesamiento distribuido de los mismos para cumplir las restricciones de tiempo de ejecución impuestas en este proble-

ma. Actualmente ya existen herramientas como Apache Spark Streaming[1] o Apache Flink[2], entre otras, capaces de realizar esta tarea. Sin embargo, aún no se ha desarrollado ninguna técnica de extracción de modelos SDRD en este ámbito.

- Eliminación de redundancias en modelos SDRD. Los patrones obtenidos en SDRD deben ser tratados de manera individual, por lo tanto es interesante evitar en la medida de lo posible patrones con alto nivel de solapamiento. Aunque los algoritmos presentados en este trabajo contienen el mecanismo de *token competition* para evitar redundancias, este no es lo suficientemente potente para eliminar patrones con alto nivel de solapamiento. En este sentido, se hace necesario el desarrollo de métodos capaces de eliminar de manera eficaz y eficiente dichos patrones para poder aplicarse en problemas complejos, especialmente en flujos de datos.

- Desarrollo de modelos SDRD enfocados a problemas singulares en aprendizaje automático. En este trabajo de investigación se ha presentado un análisis introductorio del comportamiento de los principales enfoques SD en problemas MIL. Este resultado anima al desarrollo de nuevos modelos SDRD enfocados a este tipo de problema. Asimismo, existe otro tipo de problemas, denominados multi-etiqueta [171], en los que existe más de una clase asignada a cada instancia. Hasta donde alcanza nuestro conocimiento, este tipo de problema aún no se ha analizado desde el punto de vista descriptivo.

- Descripción del cambio de concepto. No sólo es interesante la descripción del estado subyacente de los datos. En algunos problemas también es interesante la identificación de los elementos que han producido un cambio. A este concepto se le denomina en la literatura como *change mining* [13] y las técnicas de EPM son capaces de realizar esta tarea. Por lo tanto, es una línea de trabajo futuro a tener en cuenta.

- Mejora en el tratamiento del cambio de concepto. En el trabajo propuesto en esta investigación, únicamente se han tratado cambios de conceptos reales, pues son los que afectan a la calidad del modelo. Sin embargo, también es interesante desde el punto de vista descriptivo el análisis del

[1] https://spark.apache.org/streaming/
[2] https://flink.apache.org/

cambio de concepto virtual, así como otros tipos de cambio como la aparación/desaparición de variables y/o clases. Por lo tanto, existe una amplia línea de trabajo futuro en el desarrollo de técnicas para el tratamiento eficaz y eficiente de los diversos tipos de cambio de concepto dentro de SDRD.

3.4. Publicaciones relacionadas con la memoria

Finalmente, en esta sección se presenta una lista de publicaciones derivadas de los resultados presentados en este documento:

3.4.1. Revistas internacionales indexadas en JCR

- J. Luengo, A. M. García-Vico, M. D. Pérez-Godoy y C. J. Carmona, «The influence of noise on the evolutionary fuzzy systems for subgroup discovery,» *Soft Computing*, vol. 20, n.º 11, págs. 4313-4330, 2016. DOI: 10.1007/s00500-016-2300-1, IF (JCR 2016): 2.472, Ranking: 33/105 (Computer Science, Interdisciplinary Applications), Cuartil: Q2.

- A. M. García, F. Charte, P. González, C. J. Carmona y M. J. del Jesús, «Subgroup Discovery with Evolutionary Fuzzy Systems in R: The SDEFSR Package,» *The R Journal*, vol. 8, n.º 2, págs. 307-323, 2016. DOI: 10.32614/RJ-2016-048, IF (JCR 2016): 1.075, Ranking: 55/124 (Statistics & Probability), Cuartil: Q2.

- A. M. García-Vico, C. J. Carmona, D. Martín, M. García-Borroto y 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, n.º 1, e1231, 2018. DOI: 10.1002/widm.1231, IF (JCR 2018): 2.541, Ranking: 26/105 (Computer Science, Theory & Methods), Cuartil: Q1.

- A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «MOEA-EFEP: Multi-Objective Evolutionary Algorithm for Extracting Fuzzy Emerging Patterns,» *IEEE Transactions on Fuzzy Systems*, vol. 26, n.º 5, págs. 2861-2872, 2018. DOI: 10.1109/TFUZZ.2018.2814577, IF (JCR 2018): 8.759, Ranking: 6/134 (Computer Science, Artificial Intelligence), Cuartil: Q1.

- A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «A Big Data Approach for Extracting Fuzzy Emerging Patterns,» *Cognitive Computation*, vol. 11, n.º 3, págs. 400-417, 2019. DOI: 10.1007/s12559-018-9612-7, IF (JCR 2018): 4.287, Ranking: 25/134 (Computer Science, Artificial Intelligence), Cuartil: Q1.

- J. M. Luna, C. J. Carmona, A. M. García-Vico, M. J. del Jesus y S. Ventura, «Subgroup Discovery on Multiple Instance Data,» *International Journal of Computational Intelligence Systems*, vol. 12, n.º 2, págs. 1602-1612, 2019. DOI: 10.2991/ijcis.d.191213.001, IF (JCR 2018): 2.153, Ranking: 55/106 (Computer Science, Interdisciplinary Applications), Cuartil: Q3.

- A. M. García-Vico, C. J. Carmona, P. González, H. Seker y M. J. del Jesus, «FEPDS: A proposal for the Extraction of Fuzzy Emerging Patterns in Data Streams,» *IEEE Transactions on Fuzzy Systems*, Submitted (Major revision). DOI: No disponible, IF (JCR 2018): 8.759, Ranking: 6/134 (Computer Science, Artificial Intelligence), Cuartil: Q1.

  .

3.4.2. Congresos internacionales

- A. M. García-Vico, J. Montes, J. Aguilera, C. J. Carmona y M. J. del Jesus, «Analysing Concentrating Photovoltaics Technology through the use of Emerging Pattern Mining,» en *Proc. of the 11th International Conference on Soft Computing Models in Industrial and Environmental Applications*, 2016, págs. 1-8.

- A. M. García-Vico, P. González, M. J. del Jesus y C. J. Carmona, «A First Approach to Handle Emergining Patterns Mining on Big Data Problems: The EvAEFP-Spark Algorithm,» en *Proc. of the 2017 IEEE International Conference on Fuzzy Systems*, 2017, págs. 1-6.

- A. M. García-Vico, P. González, C. J. Carmona y M. J. del Jesus, «Impact of the type of rule in Fuzzy Emerging Pattern Mining on a Big Data Approach,» en *Proc. of the 2nd International Symposium on Fuzzy and Rough Sets*, 2017, págs. 1-10.

- C. J. Carmona, A. M. García-Vico, P. González y M. J. del Jesus, «Extracting Emerging Patterns with Change Detection in Time for Data Streams,» en *Proc. of the 1st International 'Alan Turing' Conference on Decision Support and Recommender Systems*, 2019.

3.4.3. Congresos nacionales

- C. J. Carmona, F. J. Pulgar-Rubio, A. M. García-Vico, P. González y M. J. del Jesus, «Análisis descriptivo mediante aprendizaje supervisado basado en patrones emergentes,» en *Proc. of the 7th Simposio Teoría y Aplicaciones de Minería de Datos*, 2015, págs. 685-694.

- A. M. García-Vico, F. Charte, P. González, C. J. Carmona y M. J. del Jesus, «Usando Algoritmos de Descubrimiento de Subgrupos en R: El Paquete SDR,» en *Proc. of the 16th Conference of the Spanish Association for Artificial Intelligence*, 2015, págs. 739-748.

- A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «Minería de Patrones Emergentes: Una oportunidad para la extracción evolutiva de conocimiento,» en *Proc. of the 11th Congreso Español de Metaheurísticas, Algoritmos Evolutivos y Bioinspirados*, 2016, págs. 1-10.

- A. M. García-Vico, C. J. Carmona y M. J. del Jesus, «Análisis de Diferentes Tipos de Reglas en Sistemas Difusos Evolutivos para Minería de Patrones Emergentes,» en *Proc. of the 12th Spanish Conference on Metaheuristics, Evolutive and Bioinspired Algorithms*, 2017, págs. 876-885.

- A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «Una primera aproximación para la extracción de patrones emergentes en flujos continuos de datos,» en *Proc. of the 18th Conferencia de la Asociación Española para la Inteligencia Artificial*, 2018, págs. 1093-1098.

- A. M. García-Vico, «Modelos descriptivos basados en aprendizaje supervisado para el tratamiento de grandes volúmenes de datos y flujos continuos de datos,» en *Proc. of the 18th Conferencia de la Asociación Española para la Inteligencia Artificial*, 2018, págs. 1402-1407.

  Este trabajo obtuvo el premio al mejor trabajo en II Workshop en Big Data y Análisis de Datos Escalable, XVIII Conferencia de la Asociación Española para la Inteligencia Artificial, Granada, Octubre 2018.

3.4.4. Seminarios impartidos

- *Supervised Descriptive Rule Models for the Extraction of Knowledge in Big Data and Data Streams*. Intelligent Systems Lab, University of Bristol, Bristol (Reino Unido). Noviembre 2018.
- *Supervised Descriptive Rule Models for the Extraction of Knowledge in Big Data and Data Streams*. School of Computing and Informatics, DeMontfort University, Leicester (Reino Unido). Diciembre 2018.

3.4.5. Premios asociados a la tesis

- Mejor trabajo de iniciación a la investigación en IV Premios Ada Lovelace de la Universidad de Jaén. Septiembre 2018.
- Premio al mejor trabajo en II Workshop en Big Data y Análisis de Datos Escalable, XVIII Conferencia de la Asociación Española para la Inteligencia Artificial, Granada, Octubre 2018.
- Mejor trabajo de investigación en V Jornadas Doctorales en TIC de la Universidad de Jaén. Mayo 2019.

3
Concluding remarks

This chapter summarises the results presented in this thesis and draws conclusions from the work carried out. It also comments on the aspects related to future work that can be carried out along the research lines developed here. As presented throughout this study:

1. For the SD technique there is a wide range of developments throughout the literature, but very few are focused on complex problems. In fact, to the best of our knowledge, there is no SD method focused on extracting knowledge from MIL problems or data stream mining environments. Despite the existence of methods focused on big data, the effect of external factors that can influence the quality of the data, such as the presence of noise, has not yet been analysed.

2. The descriptive properties of EPM are very interesting for the expert, but the descriptive aspect of this task has not been driven by the specialized literature. In this way, EPM has been used as a means of classification where accuracy is maximised, with little regard for the descriptive capacity.

3. To the best of our knowledge, the development of EPM methods for complex problems is still very limited.

4. Finally, it is important to remark that the SDRD techniques are not available in the main data analysis software. In this sense it is interesting to include these kinds of methods within this software in order to provide an answer to the need for data analysis through this approach.

To alleviate this situation the research carried out focuses mainly on the analysis and development of algorithmic solutions with the aim of being able to apply this type of technique to complex problems. Specifically, the thesis focuses on the tasks of SD and EPM applied to this type of problem.

3.1. Conclusions related to subgroup discovery

The first part of the investigation is geared towards SD. The conclusions of the work carried out in this area are presented below:

- One of the main characteristics of complex problems is the diversity of sources from which the data come. This increases the probability of introducing noise into the system, reducing its performance. Although there are several developments based on evolutionary fuzzy systems for SD, there is no formal analysis of their robustness. In order to resolve this, an initial analysis is presented of the influence of noise on these types of methods and the effectiveness of applying classical, classification-oriented noise filters on them. The results highlight the FuGePSD algorithm, which is able to isolate and ignore noise. With respect to the remaining methods, the application of a filter helps to alleviate these negative effects.

- Some of the complex problems currently on the rise are MIL problems such as image segmentation and molecule synthesis, among others. These types of problems are interesting in SD, but cannot be addressed using current approaches. Therefore, a study is presented with the adaptation to MIL problems of the main approaches developed for SD, based on the MIL standard assumption. The results show that the classical approaches do not present good quality results, while those approaches based on evolutionary algorithms allow us to obtain better results for most problems. Therefore, this study opens up a new line of research focused on the development of SD methods applied to MIL problems.

- As a result of the in-depth study carried out on the literature related to SDRD, a large amount of specific software containing the most outstanding SDRD algorithms is analysed. However, the lack of such software packages in the most popular data analysis tools, such as R or Python, prevents the scope of the task from being much wider. In order to alleviate this deficit, the SDEFSR package is developed of the bibliographic review carried out. This package contains the main evolutionary fuzzy system-based algorithms, as well as specific basic pattern processing tools, along with a web interface for easy use. The response of the community to this package has been acceptable and encourages the continuous improvement of the package to expand its influence.

In this part of the doctoral thesis, an in-depth analysis of different problems that are of interest to the SD community has been carried out. The main conclusion obtained is the capacity of evolutionary algorithms in general, and evolutionary fuzzy systems in particular, to adapt to new environments and types of problems while maintaining a good level of quality. In fact, the work carried out opens up new research lines for the improvement of the problems presented in them.

3.2. Conclusions related to emerging pattern mining

The second part of the development of this thesis is oriented towards EPM. The conclusions drawn from the work carried out are set out below:

- One of the first objectives of this part is an analysis of the different approaches developed throughout the literature for EPM from a descriptive point of view. The result of this study is a grouping of the different subsets of emerging patterns according to certain interesting characteristics, together with an algorithmic taxonomy depending on the approach taken. In addition, these methods are analysed from a descriptive point of view. The conclusions obtained highlight the descriptive capabilities of Chi-type patterns and the use of fuzzy logic to improve the descriptive characteristics of the emerging patterns. Finally, the study reflects the high descriptive capabilities of evolutionary fuzzy system approaches.
- Next, an evolutionary fuzzy system is developed in order to exploit all the positive features for the extraction of highly descriptive emergent patterns proposed in the previous study, the MOEA-EFEP algorithm. This

method is based on a multi-objective evolutionary algorithm. It uses genetic operators specifically oriented to the extraction of simple patterns together with the use of the token competition operator, which encourages the extraction of highly descriptive emerging patterns. The results obtained show a good balance between generality and reliability of the extracted patterns, significantly surpassing the main approaches developed in the literature. Therefore, we conclude that the MOEA-EFEP algorithm is an interesting alternative for the extraction of quality knowledge for the EPM task.

- The BD-EFEP algorithm, an evolutionary fuzzy system for EPM focused on addressing big data problems, has been proposed. The main novelty of this method is the use of an exact procedure for large-scale data management because only the most computationally expensive process is distributed throughout the algorithm: the evaluation of individuals. Thus, the method allows the extraction of the same knowledge model regardless of the partitions used. The results obtained demonstrate the good scalability of the method and an acceptable execution time. Furthermore, it is shown that the method achieves a better balance between generality and reliability than other proposals. The conclusion drawn from this work is that the proposed approach is an interesting alternative for the extraction of high quality emerging patterns in big data environments at an acceptable execution time.

- Finally, the FEPDS algorithm is developed, which presents a compromise between the different characteristics and constraints of data stream mining. This method uses a very efficient evolutionary fuzzy system for the continuous extraction of emerging patterns. In order to do this the algorithm updates the model using a blind approach, where the set of patterns is added to the initial population to be updated together with the use of sliding windows that allow the search to be guided to promising areas based on historical data. The conclusion obtained from this work suggests that the proposed approach is very interesting for knowledge extraction in this type of environment. The method is able to extract very high quality patterns, in an acceptable execution time, while the concept drift is handled adequately.

The research work carried out in this thesis has allowed an advance in knowledge related to the descriptive capacities of the EPM task. As a result, the most interesting types of patterns and mining approaches have been identified from this point of view. In addition, progress has been made in extracting this type of knowledge in complex problems such as big data and data stream mining with very promising results.

3.3. Future work

Based on the research work carried out in this thesis a series of future studies are proposed. These are mainly based on extending and improving the presented proposals, as well as addressing unresolved issues that have arisen during the development of the research.

- Development of efficient methods for SDRD in big data environments. In this study, a proposal has been presented to address the problem of extracting this type of knowledge in big data environments. However, it is interesting to continue the research on this task in order to obtain even more efficient and scalable methods.

- Development of techniques for distributed data stream analysis. Today the amount of data generated can be so large that distributed processing of the data is necessary to meet the runtime restrictions imposed on this problem. Currently, there are tools such as Apache Spark Streaming[1] and Apache Flink[2], among others, capable of performing this task. However, no SDRD model extraction technique has been developed in this area yet.

- Elimination of redundancies in SDRD models. The patterns obtained in SDRD must be treated individually, so it is interesting to avoid as much as possible redundant or highly overlapping patterns. Although the algorithms presented in this study contain the mechanism of token competition in order to avoid redundancies, this is not powerful enough to eliminate patterns with a high level of overlap. In this way, it is necessary to develop methods capable of effectively and efficiently removing such patterns so that they can be applied to complex problems, especially in data streams.

[1] https://spark.apache.org/streaming/
[2] https://flink.apache.org/

- Development of models focused on new complex problems. In this study, an introductory analysis of the behaviour of the main SD approaches in MIL problems has been presented. This result encourages the development of new models focused on this type of problem. There is also another type of complex problem, called multi-label [171], when there is more than one class assigned to each instance. To the best of our knowledge, this type of problem has not been analysed from the SDRD point of view yet.

- Description of the concept drift. Not only the description of the underlying state of the data is interesting. In some problems it is also interesting to identify the elements that have produced a change. This concept is referred to in the literature as *change mining* [13] and EPM techniques are capable of performing this task. Therefore, it is future work to be considered.

- Improvement in the treatment of concept drift. In the work proposed in this research, only real drifts have been dealt with, as these affect the quality of the model. However, it is also interesting from a descriptive point of view to analyse the virtual drift, as well as other types of change such as the appearance/disappearance of variables and/or classes, amongst others. Therefore, there is a broad future research line in developing techniques for the effective and efficient treatment of various types of concept drifts in this area.

3.4. Associated publications

Finally, this section presents a list of publications derived from the results presented in this document:

3.4.1. International journals indexed in JCR

- 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, no. 11, pp. 4313–4330, 2016.
DOI: 10.1007/s00500-016-2300-1, IF (JCR 2016): 2.472, Ranking: 33/105 (Computer Science, Interdisciplinary Applications), Quartile: Q2.

- A. M. García, F. Charte, P. González, C. J. Carmona and M. J. del Jesús, «Subgroup discovery with evolutionary fuzzy systems in R: The SDEFSR package», *The R Journal*, vol. 8, no. 2, pp. 307–323, 2016. DOI: 10.32614/RJ-2016-048, IF (JCR 2016): 1.075, Ranking: 55/124 (Statistics & Probability), Quartile: Q2.

- 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, e1231, 2018. DOI: 10.1002/widm.1231, IF (JCR 2018): 2.541, Ranking: 26/105 (Computer Science, Theory & Methods), Quartile: Q1.

- A. M. García-Vico, C. J. Carmona, P. González and M. J. del Jesus, «Moeaefep: Multi-objective evolutionary algorithm for extracting fuzzy emerging patterns», *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 5, pp. 2861–2872, 2018. DOI: 10.1109/TFUZZ.2018.2814577, IF (JCR 2018): 8.759, Ranking: 6/134 (Computer Science, Artificial Intelligence), Quartile: Q1.

- 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. DOI: 10.1007/s12559-018-9612-7, IF (JCR 2018): 4.287, Ranking: 25/134 (Computer Science, Artificial Intelligence), Quartile: Q1.

- J. M. Luna, C. J. Carmona, A. M. García-Vico, M. J. del Jesus and S. Ventura, «Subgroup discovery on multiple instance data», *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, pp. 1602–1612, 2019. DOI: 10.2991/ijcis.d.191213.001, IF (JCR 2018): 2.153, Ranking: 55/106 (Computer Science, Interdisciplinary Applications), Quartile: Q3.

- A. M. García-Vico, C. J. Carmona, P. González, H. Seker and M. J. del Jesus, «Fepds: A proposal for the extraction of fuzzy emerging patterns in data streams», *IEEE Transactions on Fuzzy Systems*, Submitted (Major revision). DOI: Not available, IF (JCR 2018): 8.759, Ranking: 6/134 (Computer Science, Artificial Intelligence), Quartile: Q1.

.

3.4.2. International congresses

- 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*, 2016, pp. 1–8.

- 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 *Proc. of the 2017 IEEE International Conference on Fuzzy Systems*, 2017, pp. 1–6.

- 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.

- C. J. Carmona, A. M. García-Vico, P. González and M. J. del Jesus, «Extracting emerging patterns with change detection in time for data streams», in *Proc. of the 1st International 'Alan Turing' Conference on Decision Support and Recommender Systems*, 2019.

3.4.3. National congresses

- C. J. Carmona, F. J. Pulgar-Rubio, A. M. García-Vico, P. González and M. J. del Jesus, «Análisis descriptivo mediante aprendizaje supervisado basado en patrones emergentes», in *Proc. of the 7th Simposio Teoría y Aplicaciones de Minería de Datos*, 2015, pp. 685–694.

- A. M. García-Vico, F. Charte, P. González, C. J. Carmona and M. J. del Jesus, «Usando algoritmos de descubrimiento de subgrupos en r: El paquete sdr», in *Proc. of the 16th Conference of the Spanish Association for Artificial Intelligence*, 2015, pp. 739–748.

- 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 11th Congreso Español de Metaheurísticas, Algoritmos Evolutivos y Bioinspirados*, 2016, pp. 1–10.

- A. M. García-Vico, C. J. Carmona and M. J. del Jesus, «Análisis de diferentes tipos de reglas en sistemas difusos evolutivos para minería de patrones emergentes», in *Proc. of the 12th Spanish Conference on Metaheuristics, Evolutive and Bioinspired Algorithms*, 2017, pp. 876–885.

- A. M. García-Vico, C. J. Carmona, P. González and M. J. del Jesus, «Una primera aproximación para la extracción de patrones emergentes en flujos continuos de datos», in *Proc. of the 18th Conferencia de la Asociación Española para la Inteligencia Artificial*, 2018, pp. 1093–1098.

- A. M. García-Vico, «Modelos descriptivos basados en aprendizaje supervisado para el tratamiento de grandes volúmenes de datos y flujos continuos de datos», in *Proc. of the 18th Conferencia de la Asociación Española para la Inteligencia Artificial*, 2018, pp. 1402–1407.

  This study won the prize for the best paper at the II Workshop on Big Data and Scalable Data Analysis, XVIII Conference of the Spanish Association for Artificial Intelligence, Granada, October 2018.

3.4.4. Seminars delivered

- *Supervised Descriptive Rule Models for the Extraction of Knowledge in Big Data and Data Streams.* Intelligent Systems Lab, University of Bristol, Bristol (Reino Unido). Noviembre 2018.

- *Supervised Descriptive Rule Models for the Extraction of Knowledge in Big Data and Data Streams.* School of Computing and Informatics, DeMontfort University, Leicester (Reino Unido). Diciembre 2018.

3.4.5. Awards received

- Best initiation to research work in the 4th Ada Lovelace Awards of the University of Jaén. September 2018.

- Award to the best work in 2nd Workshop on Big Data and Scalable Data Analysis, 18th Conference of the Spanish Association for Artificial Intelligence, Granada, October 2018.

- Best research paper at the 5th PhD Conference on Information and Communication Technologies at the University of Jaén. May 2019.

4

Publicaciones indexadas en JCR asociadas

En virtud con lo establecido en el artículo 25, punto 2, de la normativa vigente de los Estudios de Doctorado de la Universidad de Jaén, correspondiente al programa establecido en el RD. 99/2011, en este capítulo se presenta una colección de trabajos publicados por el doctorando durante su pertenencia al programa de doctorado. Estas publicaciones constituyen el núcleo de la tesis doctoral.

Dichas publicaciones se corresponden con siete artículos científicos publicados en revistas internacionales indexadas en JCR. Para cada una de las publicaciones presentadas, se muestra información básica sobre la publicación, así como los principales índices de calidad de las revistas en donde se han publicado los trabajos. Los trabajos mostrados en esta sección son los siguientes:

- J. Luengo, A. M. García-Vico, M. D. Pérez-Godoy y C. J. Carmona, «The influence of noise on the evolutionary fuzzy systems for subgroup discovery», *Soft Computing*, vol. 20, n.º 11, págs. 4313-4330, 2016.
  DOI: 10.1007/s00500-016-2300-1, IF (JCR 2016): 2.472, Ranking: 33/105 (Computer Science, Interdisciplinary Applications), Cuartil: Q2.

- A. M. García, F. Charte, P. González, C. J. Carmona y M. J. del Jesús, «Subgroup Discovery with Evolutionary Fuzzy Systems in R: The SDEFSR Package», *The R Journal*, vol. 8, n.º 2, págs. 307-323, 2016. DOI: 10.32614/RJ-2016-048, IF (JCR 2016): 1.075, Ranking: 55/124 (Statistics & Probability), Cuartil: Q2.

- A. M. García-Vico, C. J. Carmona, D. Martín, M. García-Borroto y 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, n.º 1, e1231, 2018. DOI: 10.1002/widm.1231, IF (JCR 2018): 2.541, Ranking: 26/105 (Computer Science, Theory & Methods), Cuartil: Q1.

- A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «MOEA-EFEP: Multi-Objective Evolutionary Algorithm for Extracting Fuzzy Emerging Patterns», *IEEE Transactions on Fuzzy Systems*, vol. 26, n.º 5, págs. 2861-2872, 2018. DOI: 10.1109/TFUZZ.2018.2814577, IF (JCR 2018): 8.759, Ranking: 6/134 (Computer Science, Artificial Intelligence), Cuartil: Q1.

- A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «A Big Data Approach for Extracting Fuzzy Emerging Patterns», *Cognitive Computation*, vol. 11, n.º 3, págs. 400-417, 2019. DOI: 10.1007/s12559-018-9612-7, IF (JCR 2018): 4.287, Ranking: 25/134 (Computer Science, Artificial Intelligence), Cuartil: Q1.

- J. M. Luna, C. J. Carmona, A. M. García-Vico, M. J. del Jesus y S. Ventura, «Subgroup Discovery on Multiple Instance Data», *International Journal of Computational Intelligence Systems*, vol. 12, n.º 2, págs. 1602-1612, 2019. DOI: 10.2991/ijcis.d.191213.001, IF (JCR 2018): 2.153, Ranking: 55/106 (Computer Science, Interdisciplinary Applications), Cuartil: Q3.

- A. M. García-Vico, C. J. Carmona, P. González, H. Seker y M. J. del Jesus, «FEPDS: A proposal for the Extraction of Fuzzy Emerging Patterns in Data Streams», *IEEE Transactions on Fuzzy Systems*, Submitted (Major revision). DOI: No disponible, IF (JCR 2018): 8.759, Ranking: 6/134 (Computer Science, Artificial Intelligence), Cuartil: Q1.

4.1. The influence of noise on the evolutionary fuzzy systems for subgroup discovery

- J. Luengo, A. M. García-Vico, M. D. Pérez-Godoy y C. J. Carmona, «The influence of noise on the evolutionary fuzzy systems for subgroup discovery», *Soft Computing*, vol. 20, n.º 11, págs. 4313-4330, 2016.

  - Estado: **Publicado**.
  - ISSN: 1432-7643.
  - DOI: 10.1007/s00500-016-2300-1.
  - Factor de Impacto (JCR 2016): **2.472**.
  - Categoría: **Computer Science, Interdisciplinary Applications**.
  - Ranking: **33/105**.
  - Cuartil: **Q2**.

Resumen

External factors such as the presence of noise in data can affect the data mining process. This is a common problem that produces several negative consequences which involves errors in the data collection, preparation and, above all, in the results obtained by the data mining techniques employed. The capabilities of the models built under such cir- cumstances will depend heavily on the quality of the training data. Hence, problems containing noise are complex prob- lems and accurate solutions are often difficult to achieve. A particular supervised learning field like subgroup discovery has overlooked the analysis of noise and its impact on the descriptions obtained. This paper presents an analysis of the impact of noise on the most relevant evolutionary fuzzy sys- tems for subgroup discovery. We also focus on how filtering techniques, devised for predictive tasks, may alleviate the impact of noise on descriptive fields such as subgroup dis- covery. Specifically, the analysis is carried out using recent filtering techniques for several class noise levels. The results obtained show two different behaviours, on the one hand, the SDIGA and NMEEFSD algorithms present a decrease in the quality of the subgroups when the noise is increased, making necessary the application of noise filtering in order to compensate for this loss of quality. On the other hand, the FuGePSD algorithm demonstrates its great

capacity to work in noisy environments without the necessity of using a preliminary filter. The study is completed with an analysis of the interpretability under the influence of noise focused on the number of rules and variables.

4.2. Subgroup Discovery with Evolutionary Fuzzy Systems in R: The SDEFSR Package

- A. M. García, F. Charte, P. González, C. J. Carmona y M. J. del Jesús, «Subgroup Discovery with Evolutionary Fuzzy Systems in R: The SDEFSR Package», *The R Journal*, vol. 8, n.º 2, págs. 307-323, 2016.

  - Estado: **Publicado**.
  - ISSN: 2073-4859.
  - DOI: 10.1007/s00500-016-2300-1.
  - Factor de Impacto (JCR 2016): **1.075**.
  - Categoría: **Statistics & Probability**.
  - Ranking: **55/124**.
  - Cuartil: **Q2**.

Resumen

Subgroup discovery is a data mining task halfway between descriptive and predictive data mining. Nowadays it is very relevant for researchers due to the fact that the knowledge extracted is simple and interesting. For this task, evolutionary fuzzy systems are well suited algorithms because they can find a good trade-off between multiple objectives in large search spaces. In fact, this paper presents the SDEFSR package, which contains all the evolutionary fuzzy systems for subgroup discovery presented throughout the literature. It is a package without dependencies on other software, providing functions with recommended default parameters. In addition, it brings a graphical user interface to avoid the user having to know all the parameters of the algorithms.

4.3. Subgroup Discovery on Multiple Instance Data

- J. M. Luna, C. J. Carmona, A. M. García-Vico, M. J. del Jesus y S. Ventura, «Subgroup Discovery on Multiple Instance Data», *International Journal of Computational Intelligence Systems*, vol. 12, n.º 2, págs. 1602-1612, 2019.

    - Estado: **Publicado**.
    - ISSN: 1875-6891.
    - DOI: 10.2991/ijcis.d.191213.001.
    - Factor de Impacto (JCR 2018): **2.153**.
    - Categoría: **Computer Science, Interdisciplinary Applications**.
    - Ranking: **55/106**.
    - Cuartil: **Q3**.

Resumen

To date, the subgroup discovery (SD) task has been considered in problems where a target variable is unequivocally described by a set of features, also known as instance. Nowadays, however, with the increasing interest in data storage, new data structures are being provided such as the multiple instance data in which a target variable value is ambiguously defined by a set of instances. Most of the proposals related to multiple instance data are based on predictive tasks and no supervised descriptive analysis can be provided when data is organized in this way. At this point, the aim of this work is to extend the SD task to cope with this type of data. SD is a really interesting task that aims at discovering interesting relationships between different features with respect to a specific target variable that is of interest for the user or the problem under study. In this regard, this paper presents three different approaches for mining interesting subgroups in multiple instance problems. The proposed models represent three different ways of tackling the problem and they are based on three well-known algorithms in the SD field: SD-Map (exhaustive search approach), CGBA-SD (Comprehensible Grammar-Based Algorithm for Subgroup Discovery) and NMEEF-SD (multi- objective evolutionary fuzzy system). The proposals have been tested on a wide set of datasets, including 10 real-world and 20 synthetic datasets, aiming at describing how the three methodologies behave on different scenarios. Any comparison is unfair since they are completely different methodologies.

4.4. An Overview of Emerging Pattern Mining in Supervised Descriptive Rule Discovery: Taxonomy, Empirical Study, Trends and Prospects

- A. M. García-Vico, C. J. Carmona, D. Martín, M. García-Borroto y 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, n.º 1, e1231, 2018.

  - Estado: **Publicado**.
  - ISSN: 1942-4795.
  - DOI: 10.1002/widm.1231.
  - Factor de Impacto (JCR 2018): **2.541**.
  - Categoría: **Computer Science, Theory & Methods**.
  - Ranking: **26/105**.
  - Cuartil: **Q1**.

Resumen

Emerging pattern mining is a data mining task that aims to discover discriminative patterns, which can describe emerging behavior with respect to a property of interest. In recent years, the description of datasets has become an interesting field due to the easy acquisition of knowledge by the experts. In this review, we will focus on the descriptive point of view of the task. We collect the existing approaches that have been proposed in the literature and group them together in a taxonomy in order to obtain a general vision of the task. A complete empirical study demonstrates the suitability of the approaches presented. This review also presents future trends and emerging prospects within pattern mining and the benefits of knowledge extracted from emerging patterns.

4.5. MOEA-EFEP: Multi-Objective Evolutionary Algorithm for Extracting Fuzzy Emerging Patterns

- A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «MOEA-EFEP: Multi-Objective Evolutionary Algorithm for Extracting Fuzzy Emerging Patterns», *IEEE Transactions on Fuzzy Systems*, vol. 26, n.º 5, págs. 2861-2872, 2018.

  - Estado: **Publicado**.
  - ISSN: 1063-6706.
  - DOI: 10.1109/TFUZZ.2018.2814577.
  - Factor de Impacto (JCR 2018): **8.759**.
  - Categoría: **Computer Science, Artificial Intelligence**.
  - Ranking: **6/134**.
  - Cuartil: **Q1**.

Resumen

Emerging pattern mining is a data mining task that belongs to the supervised descriptive rule discovery framework. Its objective is to find rules that describe emerging behaviour or differentiating characteristics with respect to a property of inter- est. A Multi-Objective Evolutionary Algorithm for the Extraction of Fuzzy Emerging Patterns (MOEA-EFEP) is described and analysed in this paper. MOEA-EFEP is the first multi-objective evolutionary algorithm proposed for emerging pattern mining. This approach allows us to get rules whose descriptions of the emerging phenomena are simpler than previous approaches. It is based on the well-known NSGA-II algorithm adapted for the extraction of emerging patterns. The proposal also uses fuzzy logic to deal with numeric variables in order to obtain a knowledge representation close to human reasoning. An experimental study was performed to verify the validity of the proposal. Firstly, it presents a comparison of different rule representations and post-processing filter strategies, in order to determine an optimal configuration of the proposal. Finally, it is compared with other algorithms for emerging pattern mining in order to determine the quality of the knowledge extracted. The results show that MOEA-EFEP obtains rules with a better description of the emerging or discriminative behaviour than other algorithms of the task. The conclusions of this study are supported by the use of statistical tests.

4.6. A Big Data Approach for Extracting Fuzzy Emerging Patterns

- A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «A Big Data Approach for Extracting Fuzzy Emerging Patterns», *Cognitive Computation*, vol. 11, n.º 3, págs. 400-417, 2019.

  - Estado: **Publicado**.
  - ISSN: 1866-9956.
  - DOI: 10.1007/s12559-018-9612-7.
  - Factor de Impacto (JCR 2018): **4.287**.
  - Categoría: **Computer Science, Artificial Intelligence**.
  - Ranking: **25/134**.
  - Cuartil: **Q1**.

Resumen

Nowadays, the growth of available data, known as big data, and machine learning techniques are changing our lives. The extraction of insights related to the underlying phenomena in data is key in order to improve decision-making processes. These underlying phenomena are described in emerging pattern mining by means of the description of the discriminative characteristics between the outputs of interest, which is a very important characteristic in machine learning. However, emerging pattern mining algorithms for big data environments have not been widely developed yet. This paper presents the first multi-objective evolutionary algorithm for emerging pattern mining in big data environments called BD-EFEP. BD-EFEP implements novelties for emerging pattern mining such as the MapReduce approach to improve the efficiency of the evaluation of the individuals, or the use of a token-competition-based procedure in order to boost the extraction of simple, general and reliable emerging pattern models. The experimental study performed using datasets with high number of examples shows the advantages of the algorithm proposed for the emerging pattern mining task in big data problems. Results show that the approach used by BD-EFEP opens new research lines for the extraction of high descriptive emerging patterns in big data environments.

4.7. FEPDS: A proposal for the Extraction of Fuzzy Emerging Patterns in Data Streams

- A. M. García-Vico, C. J. Carmona, P. González, H. Seker y M. J. del Jesus, «FEPDS: A proposal for the Extraction of Fuzzy Emerging Patterns in Data Streams», *IEEE Transactions on Fuzzy Systems*, Submitted (Major revision).

  - Estado: **Sometido en segunda vuelta con *major revision***.
  - ISSN: 1063-6706.
  - DOI: No disponible.
  - Factor de Impacto (JCR 2018): **8.759**.
  - Categoría: **Computer Science, Artificial Intelligence**.
  - Ranking: **6/134**.
  - Cuartil: **Q1**.

Resumen

Nowadays, most data is generated by devices that produce data continuously. These kinds of data can be categorised as data streams and valuable insights can be extracted from them. In particular, the insights extracted by emerging patterns are interesting in a data stream context as easy, fast, reliable decisions can be made. However, their extraction is a challenge due to the necessary response time, memory and continuous model updates. In this paper, an approach for the extraction of emerging patterns in data streams is presented. It processes the instances by means of batches following an adaptive approach. The learning algorithm is an evolutionary fuzzy system where previous knowledge is employed in order to adapt to concept drift. A wide experimental study has been performed in order to show both the suitability of the approach in combating concept drift and the quality of the knowledge extracted. Finally, the proposal is applied to a case study related to the continuous determination of the profiles of New York City cab customers according to their fare amount, in order to show its potential.

Bibliografía

[1] V. Dhar, «Data Science and Prediction», *Communications of the ACM*, vol. 56, n.º 12, págs. 64-73, 2013.

[2] D. M. Blei y P. Smyth, «Science and data science», *Proc. of the National Academy of Sciences*, vol. 114, n.º 33, págs. 8689-8692, 2017.

[3] C. Shearer, «The CRISP-DM: The new blueprint for data mining», *Journal of data warehousing*, vol. 5, n.º 4, 2000.

[4] V. Cherkassky y F. Mulier, *Learning from Data. Concepts, Theory and Methods*, 2nd. IEEE Press, 2007.

[5] G. Box, G. Jenkins y G. Reinsel, *Time series analysis: forecasting and control*, 4th. Wiley, 2008.

[6] J. B. MacQueen, «Some Methods for classification and Analysis of Multivariate Observations», en *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, págs. 281-297.

[7] R. Agrawal, T. Imieliski y A. Swami, «Mining association rules between sets of items in large databases», en *Proc. of the 1993 ACM SIGMOD International Conference on Management of Data*, 1993, págs. 207-216.

[8] D. M. Hawkins, *Identification of outliers*. Springer, 1980, ISBN: 978-94-015-3994-4.

[9] P. Kralj-Novak, N. Lavrac y G. I. Webb, «Supervised Descriptive Rule Discovery: A Unifying Survey of Constrast Set, Emerging Pattern and Subgroup Mining», *Journal of Machine Learning Research*, vol. 10, págs. 377-403, 2009.

[10] S. D. Bay y M. J. Pazzani, «Detecting group differences: Mining contrast sets», *Data Mining and Knowledge Discovery*, vol. 5, n.º 3, págs. 213-246, 2001.

[11] C.-Y. Tsai e Y.-C. Shieh, «A change detection method for sequential patterns», *Decision Support Systems*, vol. 46, págs. 501-511, 2009.

[12] J. Reps, Z. Guo, H. Zhu y U. Aickelin, «Identifying Candidate Risk Factors for Prescription Drug Side Effects Using Causal Contrast Set Mining», en *Proc. of the 4th International Conference on Health Information Science*, 2015, págs. 45-55.

[13] M. Boettcher, «Contrast and change mining», *WIREs Data Mining and Knowledge Discovery*, vol. 1, n.º 3, págs. 215-230, 2011.

[14] W. Kloesgen, «Explora: A Multipattern and Multistrategy Discovery Assistant», en *Advances in Knowledge Discovery and Data Mining*, 1996, págs. 249-271, ISBN: 978-0-262-56097-9.

[15] S. Wrobel, «An Algorithm for Multi-relational Discovery of Subgroups», en *Proc. of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery*, 1997, págs. 78-87.

[16] F. Herrera, C. J. Carmona, P. González y M. J. del Jesus, «An overview on Subgroup Discovery: Foundations and Applications», *Knowledge and Information Systems*, vol. 29, n.º 3, págs. 495-525, 2011.

[17] C. J. Carmona, P. González, M. J. del Jesus y 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, n.º 2, págs. 87-103, 2014.

[18] M. Atzmueller, «Subgroup discovery», *WIREs Data Mining and Knowledge Discovery*, vol. 5, págs. 35-49, 2015.

[19] S. Helal, «Subgroup Discovery Algorithms: A Survey and Empirical Evaluation», *Journal of Computer Science and Technology*, vol. 31, n.º 3, págs. 561-576, 2016.

[20] G. Z. Dong y J. Y. Li, «Efficient Mining of Emerging Patterns: Discovering Trends and Differences», en *Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, págs. 43-52.

[21] G. Z. Dong y J. Y. Li, «Mining border descriptions of emerging patterns from dataset pairs», *Knowledge and Information Systems*, vol. 8, n.º 2, págs. 178-202, 2005.

[22] M. García-Borroto, J. F. Martínez-Trinidad y J. A. Carrasco-Ochoa, «A survery of emerging patters for supervised classification», *Artificial Intelligence Review*, vol. 42, n.º 4, págs. 705-721, 2014.

[23] A. Fernández, S. Río, V. López, A. Bawakid, M. del Jesus, J. Benítez y 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, n.º 4, págs. 380-409, 2014.

[24] Z. Han, J. Wu, C. Huang, Q. Huang y M. Zhao, «A review on sentiment discovery and analysis of educational big-data», *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1328, 2019.

[25] Y. Xia, J. Chen, X. Lu, C. Wang y C. Xu, «Big traffic data processing framework for intelligent monitoring and recording systems», *Neurocomputing*, vol. 181, págs. 139-146, 2016.

[26] K. Soomro, M. N. M. Bhutta, Z. Khan y M. A. Tahir, «Smart city big data analytics: An advanced review», *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, n.º 5, e1319, 2019.

[27] M. Pramanik, R. Y. Lau, W. T. Yue, Y. Ye y C. Li, «Big data analytics for security and criminal investigations», *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, n.º 4, e1208, 2017.

[28] W. Ding, C.-T. Lin, S. Chen, X. Zhang y B. Hu, «Multiagent-consensus-MapReduce-based attribute reduction using co-evolutionary quantum PSO for big data applications», *Neurocomputing*, vol. 272, págs. 136-153, 2018.

[29] N. Bharill, A. Tiwari, A. Malviya, O. P. Patel, A. Gupta, D. Puthal, A. Saxena y M. Prasad, «Fuzzy knowledge based performance analysis on big data», *Neurocomputing*, In press.

[30] M. Makkie, H. Huang, Y. Zhao, A. V. Vasilakos y T. Liu, «Fast and scalable distributed deep convolutional autoencoder for fMRI big data analytics», *Neurocomputing*, vol. 325, págs. 20-30, 2019.

[31] J. Dean y S. Ghemawat, «MapReduce: Simplified data processing on large clusters», *Communications of the ACM*, vol. 51, n.º 1, págs. 107-113, 2008.

[32] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari y M. Ayyash, «Internet of things: A survey on enabling technologies, protocols, and applications», *IEEE communications surveys & tutorials*, vol. 17, n.º 4, págs. 2347-2376, 2015.

[33] J. Gama, *Knowledge discovery from data streams*. CRC Press, 2010, ISBN: 978-1-4398-2611-9.

[34] H. Fan y K. Ramamohanarao, «An efficient single-scan algorithm for mining essential jumping emerging patterns for classification», en *Proc. of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2002, págs. 456-462.

[35] H. Fan y 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, n.º 6, págs. 721-737, 2006.

[36] Q. Liu, P. Shi, Z. Hu e Y. Zhang, «A novel approach of mining strong jumping emerging patterns based on BSC-tree», *International Journal of Systems Science*, vol. 45, n.º 3, págs. 598-615, 2014.

[37] R. S. Michalski y R. Stepp, «Revealing conceptual structure in data by inductive inference», *Machine Intelligence*, vol. 10, págs. 173-196, 1982.

[38] M. Buckland y F. Gey, «The relationship between recall and precision», *Journal of the American society for information science*, vol. 45, n.º 1, págs. 12-19, 1994.

[39] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa y M. García-Borroto, «Effect of class imbalance on quality measures for contrast patterns: An experimental study», *Information Sciences*, vol. 374, págs. 179-192, 2016.

[40] M. García-Borroto, O. Loyola-González, J. F. Martínez-Trinidad y J. A. Carrasco-Ochoa, «Evaluation of quality measures for contrast patterns by using unseen objects», *Expert Systems with Applications*, vol. 83, págs. 104-113, 2017.

[41] C. J. Carmona, M. J. del Jesus y F. Herrera, «A Unifying Analysis for the Supervised Descriptive Rule Discovery via the Weighted Relative Accuracy», *Knowledge-Based Systems*, vol. 139, págs. 89-100, 2018.

[42] M. García-Borroto, O. Loyola-Gonzalez, J. F. Martínez-Trinidad y J. A. Carrasco-Ochoa, «Comparing Quality Measures for Contrast Pattern Classifiers», en *Proc. of the 18th Iberoamerican Congress, CIARP 2013*. 2013, págs. 311-318.

[43] U. M. Fayyad, G. Piatetsky-Shapiro y P. Smyth, «From data mining to knowledge discovery: an overview», en *Advances in knowledge discovery and data mining*, 1996, págs. 1-34, ISBN: 978-0-262-56097-9.

[44] D. Gamberger y N. Lavrac, «Expert-Guided Subgroup Discovery: Methodology and Application», *Journal Artificial Intelligence Research*, vol. 17, págs. 501-527, 2002.

[45] P.-N. Tan, V. Kumar y J. Srivastava, «Selecting the right interestingness measure for association patterns», en *Proc. of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, págs. 32-41.

[46] B. Kavsek y N. Lavrac, «APRIORI-SD: Adapting association rule learning to subgroup discovery», *Applied Artificial Intelligence*, vol. 20, págs. 543-583, 2006.

[47] M. Atzmueller y F. Puppe, «SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery», en *Proc. of the 17th European Conference on Machine Learning and 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2006, págs. 6-17.

[48] H. Grosskreutz, S. Rueping y S. Wrobel, «Tight optimistic estimates for fast subgroup discovery», en *Proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2008, págs. 440-456.

[49] J. M. Luna, J. R. Romero, C. Romero y S. Ventura, «On the Use of Genetic Programming for Mining Comprehensible Rules in Subgroup Discovery», *IEEE Transactions on Cybernetics*, vol. 44, n.º 12, págs. 2329-2341, 2014.

[50] M. J. del Jesus, P. González, F. Herrera y M. Mesonero, «Evolutionary Fuzzy Rule Induction Process for Subgroup Discovery: A case study in marketing», *IEEE Transactions on Fuzzy Systems*, vol. 15, n.º 4, págs. 578-592, 2007.

[51] M. J. del Jesus, P. González y F. Herrera, «Subgroup Discovery with Linguistic Rules», en *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models*, vol. 220, Springer, 2007, págs. 411-430, ISBN: 978-3-540-73723-0.

[52] C. J. Carmona, P. González, M. J. del Jesus y F. Herrera, «NMEEF-SD: Non-dominated Multi-objective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery», *IEEE Transactions on Fuzzy Systems*, vol. 18, n.º 5, págs. 958-970, 2010.

[53] C. J. Carmona, V. Ruiz-Rodado, M. J. del Jesus, A. Weber, M. Grootveld, P. González y 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, págs. 180-197, 2015.

[54] F. Padillo, J. M. Luna y S. Ventura, «Exhaustive search algorithms to mine subgroups on big data using apache spark», *Progress in Artificial Intelligence*, vol. 6, n.º 2, págs. 145-158, 2017.

[55] F. Pulgar-Rubio, A. J. Rivera-Rivas, M. D. Pérez-Godoy, P. González, C. J. Carmona y 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, págs. 70-78, 2017.

[56] C. J. Carmona, J. Luengo, P. González y M. J. del Jesus, «An analysis on the use of pre-processing methods in evolutionary fuzzy systems for subgroup discovery», *Expert Systems with Applications*, vol. 39, págs. 11 404-11 412, 2012.

[57] J. Li, J. Liu, H. Toivonene, K. Satou, Y. Sun y B. Sun, «Discovering statistically non-redundant subgroups», *Knowledge-Based Systems*, vol. 67, págs. 315-327, 2014.

[58] R. Li, R. Perneczky, A. Drzezga y S. Kramer, «Efficient redundancy reduced subgroup discovery via quadratic programming», *Journal of Intelligent Information Systems*, vol. 44, n.º 2, págs. 271-288, 2015.

[59] I. Triguero, S. González, J. M. Moyano, S. García, J. Alcalá-Fde, J. Luengo, A. Fernández, M. J. del Jesus, L. Sánchez y F. Herrera, «KEEL 3.0: An Open Source Software for Multi-Stage Analysis in Data Mining», *International Journal of Computational Intelligence Systems*, vol. 10, págs. 1238-1249, 2017.

[60] J. Demšar, T. Curk, A. Erjavec y col., «Orange: Data Mining Toolbox in Python», *Journal of Machine Learning Research*, vol. 14, págs. 2349-2353, 2013.

[61] A. M. García-Vico, C. J. Carmona, D. Martín, M. García-Borroto y 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, n.º 1, e1231, 2018.

[62] J. Y. Li, G. Z. Dong, K. Ramamohanarao y L. Wong, «DeEPs: A New Instance-Based Lazy Discovery and Classification System», *Machine Learning*, vol. 54, n.º 2, págs. 99-124, 2004.

[63] J. Bailey, T. Manoukian y K. Ramamohanarao, «Fast Algorithms for Mining Emerging Patterns», en *Principles of Data Mining and Knowledge Discovery*, vol. 2431, Springer, 2002, págs. 187-208.

[64] J. Bailey, T. Manoukian y K. Ramamohanarao, «A fast algorithm for computing hypergraph transversals and its application in mining emerging patterns», en *Proc. of the 3th International Conference on Data Mining*, 2003, págs. 485-488.

[65] H. Fan y K. Ramamohanarao, «Noise Tolerant Classification by Chi Emerging Patterns», en *Proc. of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2004, págs. 201-206.

[66] K. Ramamohanarao y H. Fan, «Patterns Based Classifiers», *World Wide Web*, vol. 10, n.º 1, págs. 71-83, 2007.

[67] M. García-Borroto, J. Martínez-Trinidad y J. Carrasco-Ochoa, «Fuzzy emerging patterns for classifying hard domains.», *Knowledge and Information Systems*, vol. 28, n.º 2, págs. 473-489, 2011.

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

[69] A. Konar, *Computational Intelligence: Principles, Techniques and Applications*. Springer, 2005, ISBN: 978-3-540-27335-6.

[70] L. A. Zadeh, «Fuzzy sets», *Information Control*, vol. 8, págs. 338-353, 1965.

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

[72] N. Alon y J. Spencer, *The probabilistic method*. Wiley-Interscience, 2000, ISBN: 978-1-119-06195-3.

[73] T. Bäck, D. Fogel y Z. Michalewicz, *Handbook of evolutionary computation*. Oxford University Press, 1997, ISBN: 978-0-7503-0392-7.

[74] R. M. Golden, *Mathematical Methods for Neural Network Analysis and Design*. Cambridge, MA: MIT Press, 1996, ISBN: 978-0-262-07174-1.

[75] G. J. Klir y B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall, 1994, ISBN: 978-0-13-101171-7.

[76] E. Hüllermeier, «Fuzzy methods in machine learning and data mining: Status and prospects», *Fuzzy Sets and Systems*, vol. 156, n.º 3, págs. 387-406, 2005.

[77] E. Hüllermeier, «Fuzzy sets in machine learning and data mining», *Applied Soft Computing*, vol. 11, n.º 2, págs. 1493-1505, 2011.

[78] L.-X. Wang y J. M. Mendel, «Generating fuzzy rules by learning from examples», *IEEE Transactions on systems, man, and cybernetics*, vol. 22, n.º 6, págs. 1414-1427, 1992.

[79] W. Pedrycz, *Fuzzy Modelling: Paradigms and Practices*. Kluwer Academic Publishers, 1996, ISBN: 978-0-7923-9703-8.

[80] R. Palm, D. Driankov y H. Hellendoorn, *Model Based Fuzzy Control*. Springer, 1997, ISBN: 978-3-662-03401-9.

[81] L. Kuncheva, *Fuzzy classifier design*. Springer, 2000, ISBN: 978-3-7908-1850-5.

[82] H. Ishibuchi, T. Nakashima y M. Nii, *Classification and Modelling with Linguistic Information Granules. Advanced Approaches to Linguistic Data Mining*. Springer, 2004, ISBN: 978-3-540-26875-8.

[83] J. H. Holland, *Adaptation in Natural and Artificial Systems*. The MIT Press, 1975, ISBN: 9780262275552.

[84] D. E. Goldberg, *Genetic Algorithms in search, optimization and machine learning*. Addison-Wesley Longman Publishing Co., Inc., 1989, ISBN: 978-0-201-15767-3.

[85] J. R. Koza, *Genetic Programming: On the Programming of computers by Means of Natural Selection*. MIT Press, 1992, ISBN: 978-0-262-11170-6.

[86] J. R. Koza, «Genetic programming as a means for programming computers by natural selection», *Statistics and Computing*, vol. 4, n.º 2, págs. 191-198, 1994.

[87] H. P. Schwefel, *Evolution and Optimum Seeking*. Wiley, 1993, ISBN: 978-0-471-57148-3.

[88] Z. Michalewicz, *Genetic algorithms + Data Structures = Ev. Programs*. Springer, 1992.

[89] D. E. Goldberg y K. Deb, «A comparative analysis of selection schemes used in genetic algorithms», en *Foundations of genetic algorithms*, vol. 1, 1991, págs. 69-93.

[90] M. Črepinšek, S.-H. Liu y M. Mernik, «Exploration and exploitation in evolutionary algorithms: A survey», *ACM computing surveys*, vol. 45, n.º 3, págs. 1-33, 2013.

[91] A. E. Eiben y J. E. Smith, *Introduction to evolutionary computation*. Springer, 2003, ISBN: 978-3-662-05094-1.

[92] S. Guillaume, «Designing fuzzy inference systems from data: An interpretability-oriented review», *IEEE Transactions on fuzzy systems*, vol. 9, n.º 3, págs. 426-443, 2001.

[93] M. J. Gacto, R. Alcalá y F. Herrera, «Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures», *Information Sciences*, vol. 181, n.º 20, págs. 4340-4360, 2011.

[94] A. Fernandez, F. Herrera, O. Cordon, M. J. del Jesus y F. Marcelloni, «Evolutionary Fuzzy Systems for Explainable Artificial Intelligence: Why, When, What for, and Where to?», *IEEE Computational Intelligence Magazine*, vol. 14, n.º 1, págs. 69-81, 2019.

[95] A. Fernandez, V. Lopez, M. J. del Jesus y F. Herrera, «Revisiting evolutionary fuzzy systems: Taxonomy, applications, new trends and challenges», *Knowledge-Based Systems*, vol. 80, págs. 109-121, 2015.

[96] O. Cordón, F. Herrera, F. Hoffmann y L. Magdalena, *Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases*. World Scientific, 2001, ISBN: 978-981-4494-45-8.

[97] O. Cordón, F. A. C. Gomide, F. Herrera, F. Hoffmann y L. Magdalena, «Ten years of genetic fuzzy systems. Current framework and new trends», *Fuzzy Sets and Systems*, vol. 14, págs. 5-31, 2004.

[98] F. Herrera, «Genetic fuzzy systems: taxomony, current research trends and prospects», *Evolutionary Intelligence*, vol. 1, págs. 27-46, 2008.

[99] O. Cordón, «A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems», *International Journal of Approximate Reasoning*, vol. 52, n.º 6, págs. 894-913, 2011.

[100] A. Fernández, C. J. Carmona, M. J. del Jesus y F. Herrera, «A View on Fuzzy Systems for Big Data: Progress and Opportunities», *International Journal of Computational Intelligence Systems*, vol. 9, n.º 1, págs. 69-80, 2016.

[101] S. F. Smith, «A learning system based on genetic adaptive algorithms», Tesis doct., Pittsburgh, PA, USA, 1980.

[102] J. H. Holland y J. S. Reitman, «Cognitive Systems Based on Adaptive Algorithms», en *Pattern directed inference systems*, Academic Press, 1978, págs. 313-329, ISBN: 978-0-12-737550-2.

[103] G. Venturini, «SIA: A Supervised Inductive Algorithm with Genetic Search for Learning Attributes based Concepts», en *Proc. of the European Conference on Machine Learning*, 1993, págs. 280-296.

[104] D. P. Greene y S. F. Smith, «Competition-based induction of decision models from examples», *Machine Learning*, vol. 13, n.º 2-3, págs. 229-257, 1993.

[105] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. S. Tarragó y S. Vluymans, *Multiple Instance Learning - Foundations and Algorithms*. Springer, 2016, ISBN: 978-3-319-47758-9.

[106] J. M. Luna, C. J. Carmona, A. M. García-Vico, M. J. del Jesus y S. Ventura, «Subgroup Discovery on Multiple Instance Data», *International Journal of Computational Intelligence Systems*, vol. 12, n.º 2, págs. 1602-1612, 2019.

[107] J. Amores, «Multiple instance classification: Review, taxonomy and comparative study», *Artificial Intelligence*, vol. 201, págs. 81-105, 2013.

[108] N. Weidmann, E. Frank y B. Pfahringer, «A Two-Level Learning Method for Generalized Multi-instance Problems», en *Proc. of the 14th European Conference on Machine Learning*, N. Lavrač, D. Gamberger, H. Blockeel y L. Todorovski, eds., 2003, págs. 468-479.

[109] D. Laney, «3D Data Management: Controlling Data Volume, Velocity, and Variety», *Application Delivery Strategies*, 2001.

[110] J. Dean y S. Ghemawat, «MapReduce: Simplified data processing on large clusters», en *Proc. of the Operating Systems Design and Implementation*, 2004, págs. 137-150.

[111] T. White, *Hadoop: The Definitive Guide*, 4.ª ed. Beijing: O'Reilly, 2015, ISBN: 978-1-4919-0163-2.

[112] J. Lin, «Mapreduce is good enough? if all you have is a hammer, throw away everything that's not a nail!», *Big Data*, vol. 1, n.º 1, págs. 28-37, 2013.

[113] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. Franklin, S. Shenker e I. Stoica, «Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing», en *Proc. of the 9th USENIX Symposium on Networked Systems Design and Implementation*, 2012, págs. 2-2.

[114] A. Spark. «MLlib: RDD-based API». (2020), dirección: http : / / spark . apache.org/docs/latest/mllib-guide.html (visitado 20-02-2020).

[115] M. M. Gaber, «Advances in data stream mining», *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, n.º 1, págs. 79-85, 2012.

[116] A. Bifet, «Adaptive learning and mining for data streams and frequent patterns», Tesis doct., Universitat Politècnica de Catalunya, 2009.

[117] I. Khamassi, M. Sayed Mouchaweh, M. Hammami y K. Ghédira, «Discussion and review on evolving data streams and concept drift adapting», *Evolving Systems*, vol. 9, n.º 1, págs. 1-23, 2018.

[118] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Wozniak y F. Herrera, «A survey on data preprocessing for data stream mining: Current status and future directions», *Neurocomputing*, vol. 239, págs. 39-57, 2017.

[119] R. Klinkenberg y T. Joachims, «Detecting concept drift with support vector machines.», en *Proc. of the 17th International Conference on Machine Learning*, 2000, págs. 487-494.

[120] W. Street e Y. Kim, «A streaming ensemble algorithm (SEA) for large-scale classification», en *Proc. of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, págs. 377-382.

[121] G. Zeira, O. Maimon, M. Last y L. Rokach, «Change detection in classification models induced from time series data», en *Data mining in time series databases*, World Scientific, 2004, págs. 101-125.

[122] J. Gama, P. Medas, G. Castillo y P. Rodrigues, «Learning with Drift Detection», en *Proc. of the 17th Brazilian Symposium on Artificial Intelligence*, 2004, págs. 286-295.

[123] A. Bifet y R. Gavaldà, «Learning from Time-Changing Data with Adaptive Windowing», en *Proc. of the 2007 SIAM International Conference on Data Mining*, 2007, págs. 443-448.

[124] P. Sobolewski y M. Wozniak, «Concept Drift Detection and Model Selection with Simulated Recurrence and Ensembles of Statistical Detectors.», *Journal of Universal Computer Science*, vol. 19, n.º 4, págs. 462-483, 2013.

[125] R. Vimieriro y P. Moscato, «A new method for mining disjunctive emerging patterns in high-dimensional datasets using hypergraphs», *Information Sciences*, vol. 40, págs. 1-10, 2014.

[126] G. Widmer y M. Kubat, «Learning in the presence of concept drift and hidden contexts», *Machine learning*, vol. 23, n.º 1, págs. 69-101, 1996.

[127] G. Hulten, L. Spencer y P. Domingos, «Mining time-changing data streams», en *Proc. of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, págs. 97-106.

[128] B. Babcock, S. Babu, M. Datar, R. Motwani y J. Widom, «Models and Issues in Data Stream Systems», en *Proc. of the 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 2002, págs. 1-16.

[129] J. Gama y G. Castillo, «Learning with Local Drift Detection», en *Proc. of the 2nd International Conference in Advanced Data Mining and Applications*, 2006, págs. 42-55.

[130] I. Khamassi y M. Sayed Mouchaweh, «Drift detection and monitoring in non-stationary environments», en *Proc. of the 2014 IEEE Conference on Evolving and Adaptive Intelligent Systems*, 2014, págs. 1-6.

[131] D.-H. Tran, «Automated change detection and reactive clustering in multivariate streaming data», en *Proc. of the 2019 IEEE-RIVF International Conference on Computing and Communication Technologies*, 2019, págs. 1-6.

[132] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski y M. Woźniak, «Ensemble learning for data stream analysis: A survey», *Information Fusion*, vol. 37, págs. 132-156, 2017.

[133] E. Ruiz y J. Casillas, «Adaptive fuzzy partitions for evolving association rules in big data stream», *International Journal of Approximate Reasoning*, vol. 93, págs. 463-486, 2018.

[134] S. Ren, B. Liao, W. Zhu, Z. Li, W. Liu y K. Li, «The Gradual Resampling Ensemble for mining imbalanced data streams with concept drift», *Neurocomputing*, vol. 286, págs. 150-166, 2018.

[135] E. Soares, P. Costa Jr, B. Costa y D. Leite, «Ensemble of evolving data clouds and fuzzy models for weather time series prediction», *Applied Soft Computing*, vol. 64, págs. 445-453, 2018.

[136] J. R. B. Junior y M. do Carmo Nicoletti, «An iterative boosting-based ensemble for streaming data classification», *Information Fusion*, vol. 45, págs. 66-78, 2019.

[137] A. Morshed, P. P. Jayaraman, T. Sellis, D. Georgakopoulos, M. Villari y R. Ranjan, «Deep osmosis: Holistic distributed deep learning in osmotic computing», *IEEE Cloud Computing*, vol. 4, n.º 6, págs. 22-32, 2018.

[138] C.-C. Lin, L. Shu, D.-J. Deng, T.-L. Yeh, Y.-H. Chen y H.-L. Hsieh, «A MapReduce-based ensemble learning method with multiple classifier types and diversity for condition-based maintenance with concept drifts», *IEEE Cloud Computing*, vol. 4, n.º 6, págs. 38-48, 2017.

[139] M. M. Finucane, I. Martinez y S. Cody, «What works for whom? A Bayesian approach to channeling big data streams for public program evaluation», *American Journal of Evaluation*, vol. 39, n.º 1, págs. 109-122, 2018.

[140] O. Carr, K. Saunders, A. Tsanas, A. Bilderbeck, N. Palmius, J. Geddes, R. Foster, G. Goodwin y M. De Vos, «Variability in phase and amplitude of diurnal rhythms is related to variation of mood in bipolar and borderline personality disorder», *Scientific reports*, vol. 8, n.º 1, pág. 1649, 2018.

[141] I. Khan, J. Z. Huang, Z. Luo y M. A. Masud, «CPLP: An algorithm for tracking the changes of power consumption patterns in load profile data over time», *Information Sciences*, vol. 429, págs. 332-348, 2018.

[142] J. Luengo, A. M. García-Vico, M. D. Pérez-Godoy y C. J. Carmona, «The influence of noise on the evolutionary fuzzy systems for subgroup discovery», *Soft Computing*, vol. 20, n.º 11, págs. 4313-4330, 2016.

[143] A. M. García, F. Charte, P. González, C. J. Carmona y M. J. del Jesús, «Subgroup Discovery with Evolutionary Fuzzy Systems in R: The SDEFSR Package», *The R Journal*, vol. 8, n.º 2, págs. 307-323, 2016.

[144] A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «MOEA-EFEP: Multi-Objective Evolutionary Algorithm for Extracting Fuzzy Emerging Patterns», *IEEE Transactions on Fuzzy Systems*, vol. 26, n.º 5, págs. 2861-2872, 2018.

[145] A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «A Big Data Approach for Extracting Fuzzy Emerging Patterns», *Cognitive Computation*, vol. 11, n.º 3, págs. 400-417, 2019.

[146] A. M. García-Vico, C. J. Carmona, P. González, H. Seker y M. J. del Jesus, «FEPDS: A proposal for the Extraction of Fuzzy Emerging Patterns in Data Streams», *IEEE Transactions on Fuzzy Systems*, Submitted (Major revision).

[147] S. García, J. Luengo y F. Herrera, *Data Preprocessing in Data Mining*. Springer, 2015, ISBN: 978-3-319-10247-4.

[148] C. E. Brodley y M. A. Friedl, «Identifying Mislabeled Training Data», *Journal of Artificial Intelligence Research*, vol. 11, págs. 131-167, 1999.

[149] T. M. Khoshgoftaar y P. Rebours, «Improving software quality prediction by noise filtering techniques», *Journal of Computer Science and Technology*, vol. 22, págs. 387-396, 2007.

[150] C.-M. Teng, «Correcting Noisy Data», en *Proc. of the 16th International Conference on Machine Learning*, 1999, págs. 239-248.

[151] S. Verbaeten y A. V. Assche, «Ensemble methods for noise elimination in classification problems», en *Proc. of the 4th International Workshop on Multiple Classifier Systems*, 2003, págs. 317-325.

[152] M. Atzmueller y F. Lemmerich, «VIKAMINE–open-source subgroup discovery, pattern mining, and analytics», en *Proc. of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2012, págs. 842-845.

[153] M. Meeng y A. Knobbe, «Flexible enrichment with Cortana–software demo», en *Proc. of the BeneLearn*, 2011, págs. 117-119.

[154] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann e I. H. Witten, «The WEKA data mining software: an update», *ACM SIGKDD explorations newsletter*, vol. 11, n.º 1, págs. 10-18, 2009.

[155] A. M. García-Vico, F. Charte, P. González, C. J. Carmona y M. J. del Jesus, «Usando Algoritmos de Descubrimiento de Subgrupos en R: El Paquete SDR», en *Proc. of the 16th Conference of the Spanish Association for Artificial Intelligence*, 2015, págs. 739-748.

[156] J. Han, J. Pei, Y. Yin y R. Mao, «Mining frequent patterns without candidate generation: A frequent-pattern tree approach», *Data mining and knowledge discovery*, vol. 8, n.º 1, págs. 53-87, 2004.

[157] C. J. Carmona, F. J. Pulgar-Rubio, A. M. García-Vico, P. González y M. J. del Jesus, «Análisis descriptivo mediante aprendizaje supervisado basado en patrones emergentes», en *Proc. of the 7th Simposio Teoría y Aplicaciones de Minería de Datos*, 2015, págs. 685-694.

[158] A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «Minería de Patrones Emergentes: Una oportunidad para la extracción evolutiva de conocimiento.», en *Proc. of the 17th Conference of the Spanish Association for Artificial Intelligence*, 2016, págs. 149-159.

[159] A. M. García-Vico, C. J. Carmona y M. J. del Jesus, «Análisis de Diferentes Tipos de Reglas en Sistemas Difusos Evolutivos para Minería de Patrones Emergentes», en *Proc. of the 12th Spanish Conference on Metaheuristics, Evolutive and Bioinspired Algorithms*, 2017, págs. 876-885.

[160] K. Deb, A. Pratap, S. Agrawal y T. Meyarivan, «A fast and elitist multiobjective genetic algorithm: NSGA-II», *IEEE Transactions Evolutionary Computation*, vol. 6, n.º 2, págs. 182-197, 2002.

[161] K. S. Leung, Y. Leung, L. So y K. F. Yam, «Rule Learning in Expert Systems Using Genetic Algorithm: 1, Concepts», en *Proc. of the 2nd International Conference on Fuzzy Logic and Neural Networks*, 1992, págs. 201-204.

[162] A. Fernández, A. Altalhi, S. Alshomrani y F. Herrera, «Why Linguistic Fuzzy Rule Based Classification Systems perform well in Big Data Applications?», *International Journal of Computational Intelligence Systems*, vol. 10, n.º 1, págs. 1211-1225, 2017.

[163] S. Río, V. López, J. M. Benítez y F. Herrera, «A MapReduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules», *International Journal of Computational Intelligence Systems*, vol. 8, n.º 3, págs. 442-437, 2015.

[164] M. Elkano, M. Galar, J. Sanz y H. Bustince, «CHI-BD: A fuzzy rule-based classification system for Big Data classification problems», *Fuzzy Sets and Systems*, vol. 348, págs. 75-101, 2018.

[165] I. Rodríguez-Fdez, M. Mucientes y A. Bugarín, «S-FRULER: Scalable fuzzy rule learning through evolution for regression», *Knowledge-Based Systems*, vol. 110, págs. 255-266, 2016.

[166] D. Peralta, S. Río, S. Ramíez-Gallego, I. Triguero, J. M. Beníez y F. Herrera, «Evolutionary Feature Selection for Big Data Classification: A MapReduce Approach», *Mathematical Problems in Engineering*, vol. 2015, págs. 1-11, 2015.

[167] F. Padillo, J. M. Luna y S. Ventura, «An evolutionary algorithm for mining rare association rules: A Big Data approach», en *Proc. of the 2017 IEEE Congress on Evolutionary Computation*, 2017, págs. 2007-2014.

[168] A. M. García-Vico, P. González, M. J. del Jesus y C. J. Carmona, «A First Approach to Handle Emergining Patterns Mining on Big Data Problems: The EvAEFP-Spark Algorithm», en *Proc. of the 2017 IEEE International Conference on Fuzzy Systems*, 2017, págs. 1-6.

[169] A. M. García-Vico, P. González, C. J. Carmona y M. J. del Jesus, «Impact of the type of rule in Fuzzy Emerging Pattern Mining on a Big Data Approach», en *Proc. of the 2nd International Symposium on Fuzzy and Rough Sets*, 2017, págs. 1-10.

[170] A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «Una primera aproximación para la extracción de patrones emergentes en flujos continuos de datos», en *Proc. of the 18th Conferencia de la Asociación Española para la Inteligencia Artificial*, 2018, págs. 1093-1098.

[171] F. Herrera, F. Charte, A. J. Rivera y M. J. Del Jesus, «Multilabel classification», en *Multilabel Classification*, Springer, 2016, págs. 17-31, ISBN: 978-3-319-41111-8.

[172] C. J. Carmona, A. M. García-Vico, P. González y M. J. del Jesus, «Extracting Emerging Patterns with Change Detection in Time for Data Streams», en *Proc. of the 1st International 'Alan Turing' Conference on Decision Support and Recommender Systems*, 2019.

[173] A. M. García-Vico, C. J. Carmona, P. González y M. J. del Jesus, «Minería de Patrones Emergentes: Una oportunidad para la extracción evolutiva de conocimiento», en *Proc. of the 11th Congreso Español de Metaheurísticas, Algoritmos Evolutivos y Bioinspirados*, 2016, págs. 1-10.

[174] A. M. García-Vico, «Modelos descriptivos basados en aprendizaje supervisado para el tratamiento de grandes volúmenes de datos y flujos continuos de datos», en *Proc. of the 18th Conferencia de la Asociación Española para la Inteligencia Artificial*, 2018, págs. 1402-1407.