UNIVERSIDAD
DE GRANADA

Programa de Doctorado en Tecnologías
de la Información y la Comunicación

Estudio y Diseño de Técnicas para la Extracción de Reglas de Asociación Temporales

Tesis Doctoral

Luis Alberto Segura Delgado

Granada, pendiente de lectura en Junio de 2021

Departamento de Ciencias de la Computación
e Inteligencia Artificial

Editor: Universidad de Granada. Tesis Doctorales
Autor: Luis Alberto Segura Delgado
ISBN: 978-84-1306-939-5
URI: http://hdl.handle.net/10481/69644

UNIVERSIDAD
DE GRANADA

Programa de Doctorado en Tecnologías
de la Información y la Comunicación

Estudio y Diseño de Técnicas para la Extracción de Reglas de Asociación Temporales

MEMORIA QUE PRESENTA

Luis Alberto Segura Delgado

PARA OPTAR AL GRADO DE DOCTOR EN INFORMÁTICA

Junio de 2021

DIRECTORES

Jesús Alcalá Fernández
Rafael Alcalá Fernández

Departamento de Ciencias de la Computación
e Inteligencia Artificial

El desarrollo de esta tesis ha sido financiada por el Fondo Europeo de Desarrollo Regional de la Junta de Andalucia, en el marco de la Consejería de Transformación Económica, Industria, Conocimiento y Universidades. Bajo el proyecto titulado "*Explicabilidad de la Inteligencia Artificial para el Análisis Inteligente de Datos: Aplicaciones en Problemas de BioSalud y del Internet de las Cosas*", con código de proyecto: P18-RT-2248.

Agradecimientos

Quiero dedicar esta tesis a todas aquellas personas que de una u otra forma han contribuido a la realización de la misma. Especialmente a mi familia, ya que ha sido el principal apoyo durante este camino. Sin vosotros esto no hubiese sido posible.

A mis directores Jesús y Rafael, por toda la ayuda, sus sabios consejos y su apoyo.

A Ramón, por su ayuda durante mis primeros años de carrera y sus sabios consejos.

Por último, quiero dar las gracias a todos mis amigos y compañeros de trabajo que me han apoyado y, de forma cariñosa, me han llamado "doctor" sin serlo todavía.

GRACIAS A TODOS

Resumen

A lo largo de todos estos años, en los que la tecnología ha ido tomando cada vez más peso en nuestras vidas, se han ido generando gran cantidad de datos almacenados en diferentes bases de datos por todo el mundo. Este hecho ha provocado que la minería de datos sea uno de los ámbitos que mayor interés suscita en los últimos años. Y es que las técnicas de minería de datos han sido aplicadas en una gran variedad de problemas con el fin de extraer conocimiento útil e interesante de todos los datos que cada día se van almacenando.

Entre todas las técnicas de minería de datos existentes, aquellas que tratan de extraer reglas de asociación son de las más utilizadas en la actualidad, gracias a su eficacia para la extracción de conocimiento y a su facilidad para comprenderlo. Estas técnicas permiten extraer asociaciones entre los items o variables almacenados en una base de datos. Por ejemplo, en una base de datos de un supermercado con información sobre las compras de los clientes, estas técnicas serían capaces de extraer asociaciones entre los productos que suelen comprarse juntos de forma frecuente (ej: {pan, leche} → {café}; que equivale a: si compra pan y leche entonces también se comprará café).

Las técnicas de extracción de reglas de asociación permiten la obtención de conocimiento interesante que ocurre de forma frecuente, pero otro aspecto fundamental que debemos tener en cuenta es la información temporal que tenemos en los datos. En la mayor parte de problemas reales, el conocimiento no es correcto o válido para siempre, sino que a lo largo del tiempo pueden producirse cambios que afecten a dicho conocimiento. Esos cambios pueden acabar en conocimiento que acabe siendo incorrecto o poco útil. Igualmente, podemos encontrar bases de datos en las que no se detecten ciertas relaciones porque éstas se producen únicamente en ciertos intervalos de tiempo. Por ejemplo, ciertos productos solamente se venden de forma frecuente en ciertos periodos de tiempo, como los helados en verano. Por lo que la detección de

asociaciones entre dichos productos no se producirá al aplicar técnicas de extracción de reglas de asociación que no tienen en cuenta el tiempo.

En los últimos años, los investigadores han comenzado a darse cuenta de la importancia que tiene la componente temporal en el conocimiento que se extrae, lo que ha producido que se hayan desarrollado nuevas técnicas que tratan de incluir dicha componente en el proceso de extracción, tratando así de extraer reglas más interesantes y útiles para el usuario. Aunque el campo de extracción de reglas de asociación temporales se encuentra en pleno crecimiento, la novedad de este ámbito introduce algunos problemas. Se ha detectado la falta de un marco de trabajo bien definido, con una terminología estándar y una clasificación clara de las diferentes técnicas que podemos encontrar en la literatura especializada. En la actualidad se utilizan diferentes términos para referirse a lo mismo, lo que dificulta la búsqueda y comparación de propuestas existentes. Esto dificulta su expansión y su aplicación en diferentes problemas reales.

Aún teniendo en cuenta que en los últimos años se ha producido un aumento de propuestas de técnicas de extracción de reglas de asociación temporales, muchas de ellas son técnicas basadas en algoritmos clásicos de extracción de reglas de asociación. Estos algoritmos suelen adaptarse para tener en cuenta la componente temporal de diferentes formas. El uso de técnicas clásicas adaptadas junto al uso de medidas de calidad clásicas para evaluar las reglas nos sugiere que es necesario el desarrollo de nuevas propuestas que aprovechen las bondades de los algoritmos actuales en el proceso de extracción, además del desarrollo de nuevas medidas de calidad que permitan evaluar lo interesante o útil que es una regla para el usuario en función del problema.

Por todo lo anterior, en esta tesis se propone una taxonomía de dos niveles que permite clasificar las propuestas de extracción de reglas de asociación temporales existentes en la literatura, proporcionando así un marco de trabajo bien definido que permita a los investigadores conocer las propuestas existentes y detectar los problemas abiertos en los que puedan aportar nuevas soluciones. También se propone un nuevo algoritmo evolutivo multiobjetivo para extracción de reglas de asociación temporales, HAUS-rules, que hace uso de una medida de utilidad media de las reglas para guiar el proceso de búsqueda, lo que permite obtener reglas más interesantes, útiles y fáciles de comprender por el usuario. Por último, se aplica nuestra nueva propuesta en un problema bio-sanitario real para el análisis temporal sobre un estudio longitudinal in vivo de la expresión genética en tejido adiposo humano.

Índice

Índice de figuras

Índice de tablas

Tabla de Acrónimos

MD	—	Minería de Datos	1
BD	—	Base de Datos	2
RA	—	Regla de Asociación	1
RAT	—	Regla de Asociación Temporal	2
RAET	—	Reglas de Asociación Espacial-Temporal	35
RSAU	—	Regla Secuencial de Alta Utilidad	40
RSAUM	—	Regla Secuencial de Alta Utilidad Media	41
HAUSR	—	High Average Utility Sequential Rules	41
AG	—	Algoritmo Genético	40
AE	—	Algoritmo Evolutivo	40
AEMO	—	Algoritmo Evolutivo MultiObjetivo	41
MOEA	—	Multiple Objective Evolutionary Algorithms	41
VPA	—	Valor p Ajustado	53
ARNm	—	Ácido Ribonucleico Mensajero	56
GO	—	Genetic Oncology (Oncología Genética)	61
KEGG	—	Enciclopedia de Genes y Genomas de Kyoto	61
FM	—	Función Molecular	61
PB	—	Proceso Biológico	61
CS	—	Camino de Señalización	61
CC	—	Comportamiento Celular	61
FT	—	Factor de Transcripción	61

Capítulo 1

Introducción

A Planteamiento

Durante las últimas décadas, las técnicas de minería de datos (MD) han sido satisfactoriamente aplicadas para extraer conocimiento útil e interesante de grandes conjuntos de datos para ayudar a los expertos a tomar decisiones o para realizar predicciones sobre eventos futuros [HK06, LHTD02]. Las técnicas de extracción de reglas de asociación (RAs) son unas de las técnicas más utilizadas dentro del ámbito de la MD. Este tipo de técnicas han sido utilizadas con éxito para resolver problemas en una gran cantidad de problemas de diferentes ámbitos. Hoy en día encontramos una gran cantidad de problemas reales que pueden resolverse aplicando este tipo de técnicas, por lo que los avances en las propuestas ya conocidas además del diseño de propuestas novedosas es de gran interés para la resolución de estos problemas y futuros problemas que puedan aparecer.

Por ejemplo, el sector financiero es uno de los sectores en los que la aplicación de técnicas de minería de datos han permitido resolver problemas de forma eficaz. Conocer qué productos de inversión o de ahorro son los más adecuados para los clientes ha sido uno de los problemas clásicos que se han intentado resolver a través del uso de técnicas de minería de datos, y más concretamente a través de técnicas de extracción de RAs [ZZ02, ZH17, MCCC20].

Otro ejemplo muy conocido es el problema clásico de "la cesta de la com-

pra-[DAG19, XLT+19]. En este caso disponemos de bases de datos (BDs) transaccionales, en las que se almacenan las compras realizadas por una gran cantidad de clientes a lo largo del tiempo. El objetivo es definir dependencias entre los productos o hábitos de compra de los clientes que permitan al gestor del establecimiento tomar decisiones para maximizar sus ventas (distribuyendo los productos de forma adecuada, aumentando la oferta de determinados productos, etc) y mejorar el grado de satisfacción de sus clientes.

A lo largo de los últimos años, una importante cantidad de técnicas de extracción de RAs han sido publicadas en la literatura especializada. Las diferentes propuestas tratan de mejorar la calidad de las reglas obtenidas y optimizar el uso de recursos y el tiempo de ejecución del método. Sin embargo, una gran cantidad de ellas no tienen en cuenta la información temporal que hay en los datos en el proceso de extracción de conocimiento.

Cada vez más, en las distintas BDs encontramos información temporal en distintas formas que debe ser tenida en cuenta a la hora de extraer conocimiento a partir de ellas. Esta información temporal debe ser considerada en el proceso de extracción, ya que si no podemos extraer información que esté basada en hechos/eventos que sucedieron en el pasado y que no se dan en la actualidad, o no seremos capaces de extraer conocimiento importante que está asociado a eventos que suceden en momentos concretos en el tiempo (por ejemplo, temporadas de rebajas, huracanes, celebración de eventos deportivos, etc), o que suceden con una cierta frecuencia en el tiempo, etc. Por ejemplo, considerando nuestro problema de la cesta de la compra, productos que en el pasado se compraban de forma frecuente en la actualidad no se venden, o que haya ciertos productos que se venden con una cierta periodicidad (por ejemplo, cremas solares) pero que analizando todas las ventas del año no sea un producto que se venda de forma frecuente, etc. Si no tenemos en cuenta la componente temporal podemos tomar decisiones incorrectas y/o poco beneficiosas al no estar la información contextualizada en el tiempo.

En los últimos años, se han propuesto técnicas de extracción de reglas de asociación temporales (RATs) [AS95, MTV97, ALW+18] que introducen la información temporal en el proceso de extracción de forma satisfactoria. Si estudiamos la literatura especializada encontramos diversas formas de introducir la componente temporal en el proceso de extracción y, por tanto, en las reglas finales que se obtienen, y todo ello en función del problema y de cómo los autores han decidido resolverlo. Encontramos propuestas que introdu-

cen la componente temporal como una relación de orden [MTV97] entre los items que forman parte de la regla: "Si un cliente compra A y B, entonces comprará C en alguna de sus próximas compras". O por ejemplo, podemos encontrar trabajos que proponen la introducción de la variable temporal para indicar la periodicidad con la que se producen dichas relaciones entre los items [Car14]: "Cada sábado por la noche los clientes compran pizza y cerveza". Éstas son sólo dos de las formas más utilizadas por los investigadores para introducir la componente temporal en el proceso de extracción de RAs y en la reglas finales, sin embargo, en la literatura podemos encontrar más formas diferentes en función del problema y de las necesidades de sus autores [CPH98, AR00].

Entre las técnicas que se han ido proponiendo en los últimos años para introducir la componente temporal del problema en el proceso de extracción encontramos una gran cantidad de propuestas que extienden algoritmos clásicos de RAs (tales como Apriori [AS94] o FP-growth [HPY00]) para considerar de manera directa (por ejemplo, introducción restricciones temporal en las reglas para que se puedan aplicar) o indirecta (por ejemplo, dividiendo la BD en función de un intervalo de tiempo para generar reglas frecuentes en todas las subBDs) la información temporal. Aunque esta área de investigación ha recibido una atención significativa en los últimos años, el campo adolece de una falta de terminología estándar, lo que dificulta la búsqueda y comparación de propuestas y estudios anteriores en este campo. A modo de ejemplo, en la literatura podemos encontrar términos como RAs basadas en intervalos, reglas secuenciales, reglas de episodios, etc. Incluso, algunos de estos términos (como "reglas secuenciales" o "reglas de episodios") podrían ser incluso más generales que las RATs, ya que los datos pueden ordenarse secuencialmente según otros criterios distintos del tiempo. Debido a ello es necesario establecer un marco de trabajo bien definido que permita a los investigadores conocer las propuestas existentes y detectar problemas abiertos que les permitan proponer propuestas futuras de gran interés.

Por otro lado, a lo largo de los años se ha demostrado la necesidad de utilizar nuevas medidas de calidad que evalúen la calidad de las reglas extraídas [BBSV02], y ésto es aún más necesario en el caso de evaluar RATs, ya que en este caso no sólo es necesario evaluar la calidad de la regla como se ha hecho para RAs clásicas, sino que para las RATs es necesario tener en cuenta la posible influencia de la temporalidad en el cálculo de las medidas de calidad. La combinación de técnicas de extracción y medidas de calidad novedosas permitirían la evolución no solo a nivel teórico de las propias técnicas y medidas, sino también una evolución y mejora en los resultados,

y por lo tanto en el conocimiento que puede ser obtenido tras el proceso. De esta manera se obtendría conocimiento de mayor calidad, que puede ser mejor valorado según el problema a resolver y que puede obtenerse de una forma más eficiente.

Por todo lo anterior, es necesario no solamente el desarrollo de nuevas técnicas de extracción de RATs, sino también el uso y desarrollo de nuevas medidas de calidad que vayan más allá de medir la frecuencia o calidad de una regla de forma general, sino que es importante también ser capaz de medir el interés o la *utilidad* de una regla para un problema en concreto [HLW11, WLPF18, RMS98a, TKS02, GH06]. De esta forma, los algoritmos de extracción pueden utilizar estas medidas para guiar el proceso de búsqueda, obteniendo así reglas aún más interesantes para los usuarios finales [DDG+20].

B Objetivos

Como hemos comentado con anterioridad, aunque esta área de investigación ha recibido una atención significativa en los últimos años es necesario establecer un marco de trabajo bien definido que permita a los investigadores conocer las propuestas existentes y detectar problemas abiertos que les permitan proponer propuestas futuras de gran interés. Además, es necesario proponer nuevas propuestas basadas en modelos de búsqueda eficientes que permitan generar conocimiento interesante y útil para el problema en concreto que se esté resolviendo.

El propósito general de esta tesis se divide en dos lineas. Por un lado, definir una nueva clasificación que facilite a los investigadores establecer bases solidas para proporcionar soluciones a los problemas abiertos en el área. Por otro lado, desarrollar nuevos métodos para la extracción de RATs basados en modelos evolutivos multiobjetivo que permitan extraer RATs maximizando conjuntamente el interés y utilidad de las reglas generadas para el problema.

De acuerdo a la meta genérica de esta tesis, se establecen los siguientes objetivos:

- Estudiar y comprender el funcionamiento de las diferentes técnicas de extracción de RATs disponibles en la literatura especializada, ya que

dicho estudio del arte permite tener una idea clara de lo existente, sus puntos fuertes y sus debilidades, centrándonos de este modo en aportar algo que mejore lo existente y que realmente resuelva problemas reales.

- Proponer una clasificación de las diferentes técnicas de extracción de RATs en base al modo en el que estas introducen la componente temporal en el proceso de extracción y en las reglas finales, lo que permite comprender mejor las técnicas existentes y sus diferencias, además de aportar un marco de trabajo al resto de investigadores que deseen estudiar el estado del arte de extracción de RATs. Se pretende proponer una taxonomía a tal efecto, de la que el área adolece en este momento.

- Diseñar y desarrollar un nuevo algoritmo de extracción de reglas temporales que permita mejorar en aquellos puntos débiles que puedan existir en otras propuestas disponibles en la literatura especializada, poniendo especial interés en aquellos tipos de propuestas que tienen un futuro más prometedor en cuanto a su utilización para resolver problemas reales interesantes. Se prestará atención al uso de medidas de calidad novedosas, que tengan en cuenta la componente temporal, así como al desarrollo de técnicas de extracción más avanzadas y la posibilidad de integrar medidas de utilidad que mejoren la efectividad del proceso.

- Analizar la eficiencia del algoritmo sobre un problema real y comparar la calidad de los resultados obtenidos sobre diferentes fuentes de datos con los resultados obtenidos por otros algoritmos, de forma que podamos comprobar si nuestra propuesta supone un verdadero avance frente a otras ya disponibles. Para la aplicación del nuevo algoritmo a un problema real se utilizará un problema biológico, por lo que el desarrollo de dicho algoritmo estará diseñado para resolver de forma eficaz este problema y para obtener conocimiento útil de las BDs biológicas proporcionadas por los expertos, lo que les permitirá determinar las ventajas de diferentes tratamientos.

C Organización y Estructura de la memoria

Durante el desarrollo de la tesis se han completado cada uno de los objetivos que conforman el propósito general de la misma, quedando expuesta

la investigación completa en el presente documento. La estructura de este se introduce brevemente a continuación.

Capítulo 2: Se presenta una nueva taxonomía de dos niveles basada en la forma en la que las propuestas tienen en cuenta la componente temporal en el proceso de extracción de las reglas. Para ello se realiza un estudio del estado del arte que nos permite identificar las características principales de las propuestas que pertenecen a cada una de las categorías de la taxonomía. Además, se presenta un análisis crítico sobre la investigación en el area y se presentan algunas de las lineas de trabajo futuro.

Capítulo 3: Se lleva a cabo el desarrollo de la nueva propuesta para la extracción de RATs, concretamente reglas de asociación secuenciales de alta utilidad. Se realiza una introducción a las reglas de asociación secuenciales de alta utilidad, para posteriormente pasar a describir el diseño y desarrollo del nuevo algoritmo de extracción. Finalmente se realiza una comparación con otras propuestas existentes en la literatura para demostrar la efectividad de nuestra propuesta frente al resto.

Capítulo 4: Se lleva a cabo la aplicación de nuestra nueva propuesta en un problema biológico real, en el que los expertos han proporcionado una BD con información genética de dos grupos de pacientes con obesidad a los que se les aplican dos dietas diferentes, una más agresiva que la otra, y cuyo objetivo es tratar de detectar las diferencias en los resultados obtenidos por cada una de las dietas. Inicialmente se introduce el problema biológico junto a las medidas de calidad biológicas a tener en cuenta para la evaluación de las reglas obtenidas. Finalmente se presentan los resultados de la aplicación del nuevo algoritmo a este problema, y la evaluación de los mismos.

También se ha incluido una sección de "Comentarios Finales", en la que se resumen los resultados obtenidos en este memoria y se presentan las conclusiones finales que podemos extraer de éstos. Finalmente, se introducen algunas de las posibles líneas de investigación futuras relacionadas con este trabajo de investigación.

El presente documento termina con una recopilación bibliográfica que recoge las contribuciones principales que podemos encontrar en la literatura sobre la materia estudiada en nuestro trabajo.

Capítulo 2

Estado del Arte

En los últimos años, la cantidad de datos recopilados en diferentes áreas de aplicación ha llevado a una situación en la que la extracción de conocimiento interesante a partir de grandes BDs es un tarea muy atractiva, al mismo tiempo que representa un gran desafío para los investigadores. El descubrimiento de asociaciones entre los datos almacenados es una de las técnicas más comunes de la MD utilizadas para extraer conocimiento interesante de grandes BDs. Las RAs se representan como $X \rightarrow Y$, donde X e Y representan conjuntos de elementos que satisfacen $X \cap Y = \oslash$ [AIS93]. Estas reglas nos permiten generar modelos descriptivos o predictivos a partir de los datos, que no solo permiten explicarlos mejor, sino que, además, nos permiten realizar predicciones [HK06, ZZ02]. Las técnicas de extracción de RAs se han aplicado satisfactoriamente a una amplia variedad de problemas de diferentes áreas, como pueden ser la biomedicina [BVG+19], seguridad de tráfico [DAG19], ingeniería de software [MCC18], energía [XLT+19] o investigación demográfica [Do18].

Muchas de las propuestas disponibles en la literatura especializada se centra en la extracción de RAs sin prestar atención a la componente temporal, considerando que el conocimiento obtenido no cambia a lo largo del tiempo, representando el mismo conocimiento para todas las situaciones temporales. Sin embargo, en las aplicaciones reales, los datos normalmente cambian a lo largo del tiempo, produciéndose cambios en las relaciones que existen entre los datos a lo largo del tiempo. Si la componente temporal no se tiene en cuenta o no se tiene en cuenta correctamente, el conocimiento

extraído podría no ser útil, puesto que no sabemos si la RAs son aplicables en el presente, o si serán aplicables en el futuro [LOW02, RS02]. Además, existe la posibilidad de que no obtengamos conocimiento interesante que solamente ocurre de forma frecuente en ciertos periodos de tiempo o en eventos especiales, como pueden ser los eventos deportivos [SM11]. Por lo que el diseño de algoritmos de MDs que sean capaces de manejar correctamente la información temporal es un reto para los investigadores.

En estos últimos años, se han propuesto en la literatura especializa una gran cantidad de métodos para la extracción de reglas de asociación temporales a partir de BDs que contienen información temporal, lo que proporciona un mayor poder predictivo y un mayor interés al conocimiento obtenido [ALW+18, Car14, HLS+16, MTV97, TCL90, LNWJ03, PKSG09]. Además de las dependencias entre elementos de la misma transacción (intratransacción), estas reglas también representan dependencias entre los elementos de diferentes transacciones (intertransacción), que se refieren a diferentes momentos temporales en el antecedente y consecuente de la regla [DJ05, LFH00, TLHF03]. Reglas como "después de recibir un tratamiento de radiación durante 7 días, los pacientes de cáncer sufren nauseas y deficiencia de magnesio" generalmente son obtenidas a partir de BDs secuenciales, en las que los eventos están ordenados de forma lineal. Los algoritmos clásicos suelen extenderse para extraer RATs estableciendo una ventana temporal en la que pueden encontrar conjuntos de elementos que aparecen frecuentemente en la BD (secuencias, episodios, etc.), generando después a partir de ellos las reglas [MTV97]. En otros casos, la variable temporal se considera de forma que representa algún tipo de restricción temporal, como una restricción de distancia temporal entre eventos [Höp01, HD04]. Otras propuestas introducen la componente temporal en el proceso de aprendizaje para analizar el momento temporal en el que las reglas ocurren. Por ejemplo, estas propuestas pueden extraer reglas que representan situaciones que se producen de forma periódica en función de unos intervalos de tiempo definidos por el usuario [ORS98, LNWJ03].

A pesar de que los investigadores de este área han sido muy activos durante los últimos años, actualmente no existe una terminología estandarizada, lo que dificulta la búsqueda y comparación de propuestas y estudios. Por ejemplo, podemos encontrar en la literatura términos como los siguientes para referirse al mismo campo: "reglas secuenciales" [TCL90], "reglas de episodios" [MTV97], "reglas de asociación inter-transaccionales" [LFH00], "reglas para predicción" [DJ05], "reglas de asociación cíclicas" [ORS98], "reglas de asociación de calendario" [RMS98b], "reglas de asociación pe-

riódicas“ [LD05], “reglas de asociación de intervalos temporales“ [HK01]. Observemos que algunos de estos términos (como “reglas secuenciales“ o “reglas de episodios“) podrían ser incluso más generales que las reglas de asociación temporales, puesto que los datos que pueden ser ordenados secuencialmente siguiendo otros criterios que van más allá de la componente temporal.

Debido a ello, en este capítulo hemos presentado una nueva clasificación del estado del arte actual en el ámbito de la extracción de RATs, recopilando, organizando y resumiendo las propuestas existentes con el objetivo de entender mejor este campo y proporcionar al resto de investigadores y estudiantes un marco de trabajo en el que se puedan localizar fácilmente las propuestas existentes y detectar problemas abiertos que puedan ser resueltos en propuestas futuras. Para ello, se ha diseñado una taxonomía de dos niveles en la que se agrupan las propuestas existentes. En un primer nivel, se agrupan aquellas propuestas en las que la variable temporal es considerada como el criterio por el cual están ordenados los datos en la BD, mientras que por otro lado se agrupan aquellas propuestas en las que la componente temporal es considerada como un atributo más en el proceso de aprendizaje. En un segundo nivel, se presentan diferentes categorías basadas en el tipo de información temporal o en cómo dicha información se incluye en el proceso de extracción.

Para ello, este capítulo se estructura de la siguiente manera. En primer lugar se introduce el problema de la extracción de RATs, definiendo conceptos fundamentales sobre este tipo de reglas, que facilitan la comprensión de la información consideradas en la secciones siguientes. A continuación, se introducirá la taxonomía propuesta, incluyendo información sobre buena parte de las propuestas existentes en la literatura que hemos estudiado, y separando en dos secciones diferentes los dos grandes grupos que forman la clasificación. A continuación hacemos un repaso sobre los problemas reales en los que se han aplicado algunas de estas propuestas, y sobre algunas de las herramientas software que tenemos disponibles con licencia de libre distribución. Finalmente se expondrán una serie de consideraciones finales y se mostrarán algunas de las lineas de trabajo futuro.

2.1. Conceptos fundamentales de las reglas de asociación temporales

Las RAs permiten identificar dependencias o correlaciones entre elementos o valores (items) de la BD, donde estos pueden ser de diferentes tipos (discretos, binarios, cuantitativos, etc). Estas reglas se definen como una expresión del tipo $X \rightarrow Y$ (con $X \cap Y = \oslash$), donde X e Y son conjuntos de items (itemsets) [AIS93]. Por ejemplo, la regla {precios altos de la aerolínea 1}→{bajos precios de la aerolínea 2} podría haberse extraído de una BD de agencias de viajes, indicando que la segunda aerolínea ofrece precios bajos cuando la primera de ellas sube sus precios. Las medidas clásicas utilizadas para medir la calidad de las reglas son el *soporte* y la *confianza*. El soporte de un itemset I (al que nos referiremos como $Sop(I)$) se define como la frecuencia con la que I aparece en la BD. Basada en esta definición, las medidas de soporte y confianza de una regla $X \rightarrow Y$ se definen como:

$$Soporte(X \rightarrow Y) = \frac{Sop(XY)}{\mid N \mid}, \quad Confianza(X \rightarrow Y) = \frac{Sop(XY)}{Sop(X)}, \quad (2.1)$$

donde $\mid N \mid$ es el número de ejemplos/transacciones en la BD, y $Sop(XY)$ y $Sop(Y)$ son el soporte de los itemsets XY y X, respectivamente. Muchos de los métodos propuestos en la literatura para extraer RAs están basados en el marco de trabajo clásico *soporte-confianza*. En ellos, el usuario define un soporte mínimo (*minSop*) y una confianza mínima (*minConf*) para, en primer lugar, generar todos los itemsets cuyo soporte alcanza el minSop definido, a los que se les conoce como itemsets frecuentes, y en segundo lugar, se generan a patir de ellos las RAs cuya confianza supere la minConf indicada [FVLV$^+$17]. Sin embargo, múltiples investigadores han puesto de manifiesto los distintos problemas que presentan estas medidas de calidad [BBSV02]. Por un lado, la confianza no es capaz de detectar la independencia estadística o negativa entre el antecedente y el consecuente de la reglas, puesto que no se tiene en cuenta el soporte del consecuente en su calculo. Por otro lado, los itemsets con un valor muy alto de soporte pueden resultar en reglas poco útiles, puesto que estos están presentes en una gran parte de los ejemplos de la BD, y cualquier otro itemset de la BD puede parecer un buen predictor de su presencia. Por ello, existen muchas propuestas en las que se presentan nuevas medidas que tratan de resolver estos problemas de las medidas clásicas, para intentar seleccionar y clasificar las reglas en base a su potencial interés para el usuario [GH06, TKS02].

Tabla 2.1: Resumen de algunas de las medidas más importantes para medir la calidad de las reglas.

Definición	*Descripción*	*Dominio*
$Lift(X \rightarrow Y)$ [RMS98a]		
$\frac{Sop(XY)}{Sop(X)Sop(Y)}$	Relación entre confianza y confianza esperada: Medida < 1: Correlación negativa entre X e Y Medida = 1: X e Y son independientes Medida > 1: Correlación positiva entre X e Y	[0,∞)
$Yule's Q(X \rightarrow Y)$[TKS02]		
$\frac{Sop(XY)Sop(\neg X\neg Y) - Sop(X\neg Y)Sop(\neg XY)}{Sop(XY)SUP(\neg X\neg Y) + Sop(X\neg Y)Sop(\neg XY)}$	Coeficiente de probabilidades: Medida < 0: Correlación negativa entre X e Y Medida = 0: X e Y son independientes Medida > 0: Correlación positiva entre X e Y	[-1,1]
Factor de Certeza$(X \rightarrow Y)$[SB75]		
si confianza$(X \rightarrow Y) > Sop(Y)$: $\frac{confianza(X \rightarrow Y) - Sop(Y)}{1 - Sop(Y)}$ *si confianza*$(X \rightarrow Y) < Sop(Y)$: $\frac{confianza(X \rightarrow Y) - Sop(Y)}{Sop(Y)}$ *En otro caso es* 0	Ganancia normalizada: Medida < 0: Correlación negativa entre X e Y Medida > 0: Correlación positiva entre X e Y Medida = 0: X e Y son independientes	[-1,1]

Muchos de métodos de la literatura para extraer de RAs no prestan especial atención a la información temporal que hay en las BDs. Sin embargo, como hemos comentado, en las aplicaciones de la vida real las asociaciones entre los datos suelen cambiar con el tiempo o se presentan en distintos momentos temporales. Por ello, en los últimos años han sido muchos los métodos que se han propuesto para extraer RATs a partir de BDs temporales. Estas BDs son distintas dependiendo del atributo o la componente temporal que contengan: cada transacción o secuencia de la BD esta asociada con un instante de tiempo (timestamp) concreto; cada item/evento en una transacción esta asociado con un intervalo temporal; y cada item/evento en una transacción/secuencia esta asociado con un instante temporal en el que este se produce. A diferencia de las RAs clásicas, las RATs normalmente introducen una restricción temporal (ya sea sobre un instante exacto en el tiempo o sobre un intervalo de tiempo) que determina el momento temporal en el que sucede. Estas reglas pueden definirse siguiendo distintos enfoques o a partir de distintos tipos de itemset temporal, pero dos de los itemset temporales más utilizados en la literatura para ello son: **patrones secuenciales** y **episodios**.

El término **patrón secuencial** (sequential pattern) es una secuencia de

itemsets ordenados en función de un criterio (tiempo, espacio, etc). Este término fue introducido por Agrawal y SriKant [AS95] a partir de un problema real en el que disponían de una BD con información de compras de clientes. Vamos a considerar esta misma BD de clientes como ejemplo para introducir este término, en la que cada ejemplo/transacción de nuestra BD esta compuesta de un identificador de cliente (*id*), la lista de productos que compró el cliente y un valor que representa el instante de tiempo (*timestamp*) en el que se produjo la compra (ver Tabla 2.2a). A partir de esta BD inicial, es posible generar una nueva BD en la que las transacciones con el mismo *id* se agrupan, y las listas de productos comprados se incluyen en una única secuencia ordenadas en función de momento temporal (timestamp) en el que se realizaron (ver Tabla 2.2a). A cada una de las filas incluidas en la nueva BD se le conoce con el nombre de $secuencia - cliente$ ($customer - sequence$). Destacar que cada una de estas $secuencias - cliente$ no es más que una secuencia de itemsets ordenados por el momento temporal en el que sucedieron. Las Tablas 2.2 muestran un ejemplo de este tipo de BDs, considerando 4 clientes y 5 productos: a, b, c, d y e. La Tabla 2.2a muestra los valores de las 9 transacciones de la BD ordenados por *id* y sus respectivos *timestamps*. La Tabla 2.2b muestra las $secuencias - cliente$ que se obtienen al agrupar las transacciones de la BD original.

ID	**Timestamp**	**Items comprados**
1	26 Abril 2019	a
1	27 Abril 2019	b
1	28 Abril 2019	c,d
2	27 Abril 2019	b,c
2	28 Abril 2019	a
3	26 Abril 2019	d,e
4	26 Abril 2019	a
4	27 Abril 2019	b
4	28 Abril 2019	c

(a) BD de transacciones de clientes.

ID	**Secuencias-Cliente**
1	$\langle$(a), (b), (c,d)$\rangle$
2	$\langle$(b,c), (a)$\rangle$
3	$\langle$(d,e)$\rangle$
4	$\langle$(a), (b), (c)$\rangle$

(b) BD transformada con *secuencias − cliente*.

Tabla 2.2: Un problema sencillo con 4 clientes y 5 productos.

Este enfoque se puede aplicar a cualquier BD en la que la información se almacena en una estructura similar a la de este ejemplo. Por ejemplo, una BD en la que se almacena la lista de síntomas detectados en pacientes en distintos momentos temporales de un tratamiento, etc. A partir de la BD de este tipo, se define que una secuencia de itemsets $\langle X_1, X_2, ..., X_n\rangle$ es una subsecuencia de otra $\langle Y_1, Y_2, ..., Y_m\rangle$ con $n \leqslant m$ si se cumple $X_1 \subseteq Y_{i_1}, X_2 \subseteq Y_{i_2}, ..., X_n \subseteq Y_{i_n}$ y $i_1 < i_2 < ... < i_n$. Por lo tanto, el soporte de una secuencia S ($Sop(S)$) se define como la frecuencia con la que S es una subsecuencia de las secuencias-cliente que forman la BD. Por ejemplo, el soporte de la secuencia $S = \langle(a), (c)\rangle$ es 0,5, ya que es una subsecuencia de las secuencias-cliente 1 y 4 de la Tabla 2.2b. Basada en esta definición, muchos métodos de la literatura hacen uso del enfoque clásico soporte-confianza para generar primero patrones secuenciales frecuentes a partir de la BD, y después generar las RATs a patir de ellos [HD04, LFW05].

Otro enfoque muy utilizado en la literatura consiste en obtener obtener RATs a partir de una secuencia de eventos ordenados en base al tiempo (como son registros de navegación en páginas web, etc) en los que los eventos pueden producirse en un instante de tiempo concreto o en un intervalo de tiempo [MTV97, NLPV18]. Esta tarea consiste en encontrar colecciones de eventos (denominadas episodios) que aparecen con la suficiente frecuencia en una secuencia de eventos para generar a partir de ellos reglas que nos permitan predecir la subsecuencia que puede ocurrir a continuación. Estos episodios pueden clasificarse dentro de tres clases diferentes [MTV97]: Episodios en serie o seriales (Serial episode), cuando el orden entre los eventos es total[1]; Episodios paralelos (Parallel episode), cuando el orden entre eventos no tiene restricciones; y Episodios No-Seriales y No-Paralelos, cuando solamente algunos de los eventos atienden a restricciones de orden total. Por ejemplo, supongamos una secuencia S con 4 eventos diferentes (a, b, c y d), donde cada uno de ellos tiene asociado un *timestamp* en el que se produce:

$$S = \langle (a,1), (b,2), (d,3), (c,4), (a,5), (c,6), (a,7), (b,8), (c,9) \rangle$$

la Figura 2.1 muestra un ejemplo sencillo de cada una de las clases de episodios con una subsecuencia S' de ejemplo en la que ocurre el episodio.

Un episodio A se considera un subepisodio de otro episodio B ($A \preceq B$)si todos los eventos y restricciones de orden de A están incluidos en B. En dicho caso, el episodio B es considerado un superepisodio del episodio A. Por ejemplo, el episodio en serie (a, c) es un subepisodio del episodio en serie (a, c, d), porque contiene los mismos eventos y las restricciones de orden de aparición se mantienen entre ellos. Un episodio A se produce en una secuencia de eventos cuando A es un subepisodio de una secuencia. Para calcular el soporte de un episodio A habitualmente se define estableciendo una ventana temporal w con un tamaño que indica el máximo instante de tiempo en el que debe producirse. El soporte se define como la fracción de ventanas temporales w en las que el episodio se produce [LSU05]. Por ejemplo, si tomamos una ventana de tamaño 5 para la secuencia de eventos S indicada anteriormente, el episodio temporal (a, b, c) se produce en 5 de las 13 ventanas temporales w posibles (ver Figura 2.2). Por lo tanto, el soporte de este episodio sería 0,38. De esta forma, podemos definir un episodio frecuente como aquel cuyo soporte sea superior al minSop definido por el usuario. Al igual que en el caso de los patrones secuenciales, en la literatura podemos

[1]El orden entre los eventos es total cuando todos los eventos que componen el episodio esta ordenados según el criterio establecido.

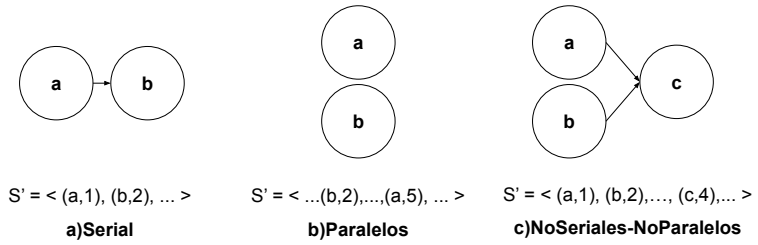

Figura 2.1: Ejemplo de las tres clases de episodios.

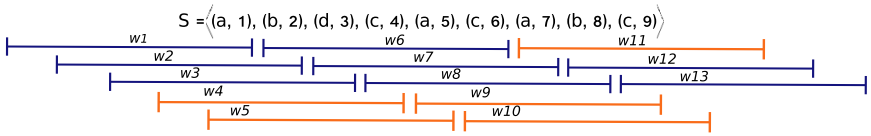

Figura 2.2: Ventanas en las que el episodio en serie (a, b, c) se produce en S con una ventana w de tamaño 5.

encontrar un gran número de propuesta que hacen uso del enfoque clásico de soporte-confianza para generar episodios frecuentes y después generar a partir de ellos las RATs [ALW+18].

En las aplicaciones del mundo real es normal que las dependencias entre los elementos de las BDs estén asociadas a un periodo de tiempo determinado o al desarrollo de un acontecimiento especial (meteorológico, deportivo, etc). El problema esta en cómo aprender o identificar ese intervalo de tiempo durante el cual las reglas se cumplen, pudiendo así descubrir la posible periodicidad con la que ocurren dichas reglas. La restricción de periodicidad puede relajarse con el objetivo de encontrar reglas interesantes que en ocasiones no se cumplen debido, por ejemplo, a que los elementos que la componen no suceden en el orden indicado debido a la presencia de ruido en los datos. Además, la periodicidad en los eventos de la vida real no siempre es tan regular, y, normalmente, contiene algunas perturbaciones. La relajación de esta restricción ha permitido a los investigadores extraer conocimiento útil para el usuario que las técnicas clásicas no son capaces de detectar.

Los diferentes marcos de trabajo y aplicaciones en los que las RATs son utilizadas han dado lugar a una falta de terminología estándar en la literatura, tal y como se ha introducido anteriormente. Es por ello que ante

esta falta de terminología estándar y de una clasificación clara de los diferentes tipos de propuestas existentes hemos decidido realizar una propuesta de taxonomía que permita a los investigadores disponer de un marco de trabajo bien definido sobre el que desarrollar sus nuevas propuestas. En las siguientes subsecciones se introduce la taxonomía de dos niveles que hemos propuesto durante nuestro estudio del estado del arte.

2.2. Taxonomía basada en la componente temporal

Con el objetivo de analizar y estudiar cuál es el estado del arte en el campo de la extracción de RATs, hemos revisado una amplia variedad de propuestas que tratan de obtener este tipo de reglas a partir de BDs con información temporal. Para clarificar su funcionamiento y clasificarlas hemos presentado una taxonomía de dos niveles, que podemos observar en la Figura 2.3.

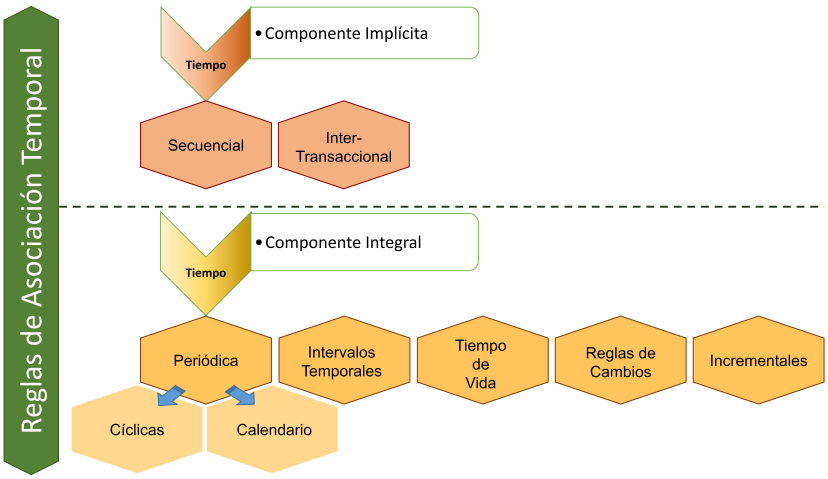

Figura 2.3: Taxonomía propuesta para las RATs.

El primer nivel agrupa las propuestas en función de si la variable tiempo es considerada como una componente implícita o integral/explícita. Este nivel ha dado lugar a 2 categorías:

- Tiempo como componente Implícita. La variable temporal se conside-

ra como una relación de orden entre los datos que forman parte del conjunto y/o para establecer restricciones temporales que determinen la importancia de un evento/dato/item con respecto a otro.

- Tiempo como componente Integra. La variable temporal se considera como un atributo más dentro del proceso de aprendizaje que, además, podría introducirse como una variable más en el modelo de la regla.

En el segundo nivel de la taxonomía, los métodos de la primera categoría (*componente implícita*) se clasifican en dos subcategorias en función del tipo de los datos temporales a partir de los cuales se extraen las reglas: *secuencial* e *inter-transaccional*. La categoría *secuencial* agrupa aquellas propuestas que tratan de extraer episodios o reglas secuenciales a partir de una o varias secuencias en las que el tiempo se utiliza para ordenar linealmente las apariciones de los datos/eventos de la BD. La categoría *inter-transaccionales* engloba a todas aquellas propuestas que tratan de obtener reglas de asociación inter-transaccionales, es decir, reglas que se obtiene a partir de BDs de transacciones en los que la componente temporal se incluye en cada una de las transacciones como una fecha o instante de tiempo en el que los datos/eventos ocurren.

Por otro lado, las propuestas de la segunda categoría del primer nivel se organizan dependiendo de como se incluye la información temporal en el proceso de aprendizaje, agrupando las propuestas en cinco categorías diferentes: periódicas, intervalos temporales, tiempo de vida (lifespan), cambios e incrementales. La categoría de *periódicas*, hace referencia a grupos de propuestas en las que se consideran restricciones de periodicidad en el proceso de extracción. Esta categoría se subdivide a su vez en dos subcategorías (*cíclicas* y *de calendario*) dependiendo de si la partición temporal se realiza haciendo uso de intervalos de tiempo uniformes o mediante un esquema de calendario. La categoría de *intervalos temporales* incluye propuestas que representan las asociaciones temporales considerando la duración de los eventos. La categoría de propestas de *tiempo de vida* (*LifeSpan*) recoge las propuestas que tienen en consideración durante el proceso de extracción los intervalos de tiempo en los que los items aparecen en la BD. La categoría *cambios* hace referencia a propuestas que generan meta-reglas temporales, para representar los cambios que se producen en las RAs a lo largo del tiempo. Finalmente, la categoría *incremental* agrupa aquellas propuestas de la literatura que resuelven el problema de extraer RATs a partir de bases de datos incrementales, actualizando las reglas extraídas o generando nuevas a partir de los datos nuevos que se van incorporando en la BD de forma

incremental.

Además de los aspectos indicados anteriormente, que tienen que ver con la componente temporal, en nuestra propuesta de taxonomía hemos considerado otros aspectos adicionales tales como si las propuestas están diseñadas para una sola secuencia o para múltiples secuencias, o si hacen uso de técnicas particulares como la lógica difusa, etc., para ordenar las propuestas disponibles en cada una de las categorías. Esta taxonomía nos permite clasificar y organizar el espacio de búsqueda en el que podemos encontrar las diferentes técnicas existentes en la literatura. En las siguientes subsecciones se proporcionará una breve descripción para algunas de las propuestas más importantes de cada categoría.

2.3. Técnicas de extracción de reglas considerando el tiempo como componente implícita

En esta subsección presentaremos una breve introducción sobre las dos subcategorías de propuestas en las que se considera la variable de tiempo como una relación de orden entre los datos de la BD y/o para establecer restricciones temporales que determinen la importancia de un evento/dato/item con respecto a otro.

Categoría Secuencial

En muchas colecciones de datos encontramos secuencias de valores o de eventos ordenadas linealmente atendiendo al instante de tiempo en el que estos se produjeron. Algunos ejemplos de este tipo de colecciones de datos son: secuencias de alarmas en una red de telecomunicaciones, actividades humanas, estados en la evolución de una enfermedad, series temporales, microarrays de expresión de genes, etc. Los usuarios que se enfrentan a este tipo de problemas podrían estar interesados en obtener RATs de estas colecciones de datos, en las que el antecedente y el consecuente hacen referencia a instantes de tiempo diferentes, en los que uno representa aquellos eventos que se producen antes en el tiempo que los otros. Por lo que este tipo de reglas representan algún tipo de relación de orden entre los datos o eventos de las secuencias, y que con una cierta confianza pueden ser utilizadas para realizar predicciones [LKW09, FVFNN12]. Las reglas secuenciales son similares a las RAs clásicas, pero introducen esas dependencias temporales de

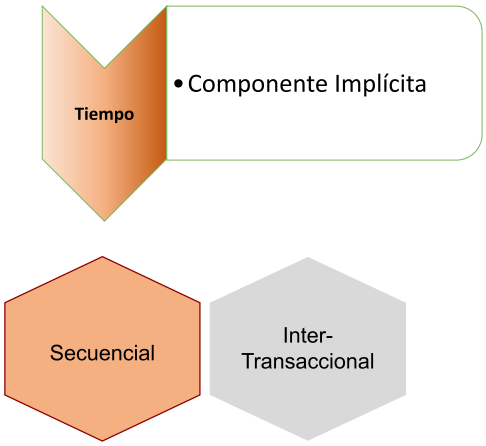

Figura 2.4: Componente Implícita: Secuenciales.

orden en las que el antecedente debe ocurrir antes que el consecuente. Esta diferencia junto a la necesidad de utilizar BDs secuenciales para obtener este tipo de reglas [TCL90] es lo que las diferencia de las RAs clásicas.

Este tipo de reglas normalmente se generan a partir de patrones secuenciales o episodios [FHHP12, FVLK$^+$17, MR13, Zim14] que aparecen en una única secuencia [MTV97, DJ05], entre secuencias diferentes [LRRV12, DLM$^+$98], o comunes a múltiples secuencias [LKW09, FWT$^+$15]. Lo habitual es que las propuestas que encontramos en la literatura estén basadas en algoritmos clásicos de extracción de RAs, que normalmente se extienden haciendo uso ventanas temporales deslizantes que permiten encontrar itemsets frecuentes que serán utilizados posteriormente para generar las reglas finales [MTV97, FVWTN12]. Algunos de estos algoritmos también incluyen otras restricciones temporales como pueden ser un máximo o mínimo en el intervalo de tiempo en el que las reglas deben producirse [HD04, NKD09] para que puedan tenerse en cuenta; o un tiempo máximo que puede pasar entre dos eventos consecutivos [NKD09, ALW$^+$18]. Algunas variantes intentan relajar las restricciones de orden de los eventos en el proceso de extracción de las reglas [FWT$^+$15, SY18], o definen una plantilla o máscara para las reglas, en la que se involucran los items o eventos que deben incluir, con el objetivo de obtener relaciones en las que el usuario este especialmente interesado [BWJ98, Sud05, CSC$^+$11]. La mayor parte de los algoritmos que encontramos para extraer reglas secuenciales se centran en generar todas las reglas posibles a partir de la BD, lo que hace el proceso

de aprendizaje ineficiente y poco útil, debido a las gran cantidad de reglas redundantes generadas. Con el objetivo de resolver este problema, algunas propuestas se han diseñado específicamente para extraer directamente reglas secuenciales no redundantes de la BD secuencial [TLVH16]. Otros métodos hacen uso de estructuras de datos específicas que les permiten mejorar en gran medida la eficiencia del proceso de extracción y la interpretabilidad de las reglas [WMXP18, FVGZT14].

En la literatura especializada encontramos como las reglas secuenciales o de episodios han sido aplicadas de forma satisfactoria a una amplia variedad de aplicaciones y problemas. Por ejemplo, este tipo de reglas se han utilizado para identificar rápidamente las bandas del espectro no utilizadas, aliviando así los problemas generados por el aumento constante de la demanda de ancho de banda de las comunicaciones [HT19]. También han sido utilizadas para mejorar la seguridad de los sistemas, aplicando este tipo de técnicas para analizar los registros de eventos con información histórica sobre posibles brechas de seguridad [KP18]. Por último, otro de los muchos casos de uso de este tipo de reglas es la predicción del nivel de congestión de tráfico [WZS+19].

Categoría Inter-Transaccionales

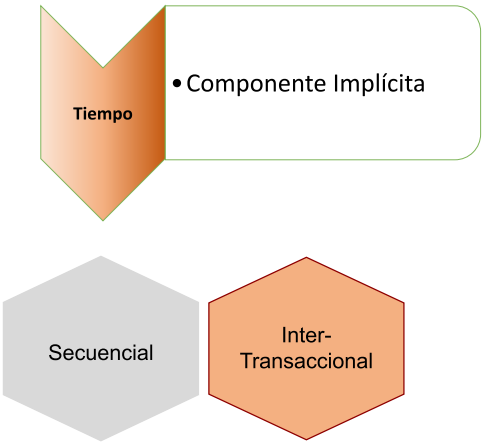

Figura 2.5: Componente Implícita: Inter-Transaccionales.

Por otro lado, en una buena parte de las aplicaciones reales se proporcionan BDs formadas por transacciones, en las que la componente temporal se

incluye en cada una de estas transacciones en forma de fecha y/o momento temporal en el que ocurrieron. En estos casos, los usuarios podrían estar interesandos en obtener dependencias entre los items de diferentes transacciones en lugar de dependencias entre items de una misma transacción. Aunque las transacciones dentro de un contexto temporal, espacial, etc., estos contextos son ignorados en las técnicas clásicas para la extracción de RAs. Cuando los items de la BD transaccional se organizan por el momento temporal en el que ocurrieron las transacciones, las RAs inter-transaccionales permiten representar asociaciones a través de la dimensión del tiempo [LFH00]. Estos tipos de reglas se generan normalmente haciendo uso de ventanas deslizantes que permiten dividir la BD por intervalos de tiempo de la misma longitud, generando reglas interesantes que abarcan un determinado número de intervalos. Con este objetivo, algunas de las propuestas de la literatura presentan extensiones del algoritmo clásico **Apriori** [AIS93], para obtener RAs inter-transaccionales [LFH00, FDL01, LFW05, HDC07]. Con el objetivo de mejorar la eficiencia en el proceso de aprendizaje, algunos de los métodos se han sido específicamente diseñados para generar este tipo de reglas. Por ejemplo, Tung [TLHF03] y otros, proponen un método en el que en primer lugar generan eficientemente los itemsets inter-transaccionales frecuentes para posteriormente generar a partir de ellos las reglas inter-transaccionales; en Wang C. S. [Wan15], los autores definen que son las reglas inter-transaccionales no redundantes y presenta un método para obtener eficientemente este tipo de reglas. Algunos investigadores también han presentado propuestas para aprender el tamaño más adecuado para la ventana de tiempo, con el objetivo de optimizar la frecuencia de las dependencias que son obtenidas [XTZ14].

Algunos investigadores han hecho uso de la lógica difusa para extraer este tipo de reglas debido a su simplicidad y similitud con respecto al razonamiento humano [INN04]. Por ejemplo, en [HK05, HKS07] los autores proponen una extensión de algunos algoritmos clásicos (Apriori y PrefixSpan) en la cual hacen uso de una ventana deslizante definida por el usuario para generar solamente las reglas difusas cuyo span es menor o igual a esta ventana. Algunos autores han hecho uso de algoritmos evolutivos para aprender las funciones de pertenencia antes de extraer las reglas de asociación inter-transacción difusas [MGH11, MGH13].

Estas reglas se han aplicado satisfactoriamente en una gran cantidad de aplicaciones reales. Por ejemplo, han sido utilizadas para representar características dinámicas de la congestión del tráfico en diferente regiones [XWZ20], o para generar reglas a partir de BDs transaccionales en sistemas dinámicos complejos de mercados financieros [HYWC16].

2.4. Técnicas de extracción de reglas considerando el tiempo como componente integral

A lo largo de esta sección se presenta una breve descripción de las diferentes subcategorías en las que se han agrupado aquellas propuestas que incluyen la componente temporal como un atributo más en el proceso de aprendizaje, y que podría introducirse como una variable más en el modelo de la regla. Estas propuestas a su vez pueden ser: periódicas, de intervalos de tiempo, de tiempo de vida, de expresión de cambio en las propias reglas e incrementales.

Categoría Periódica

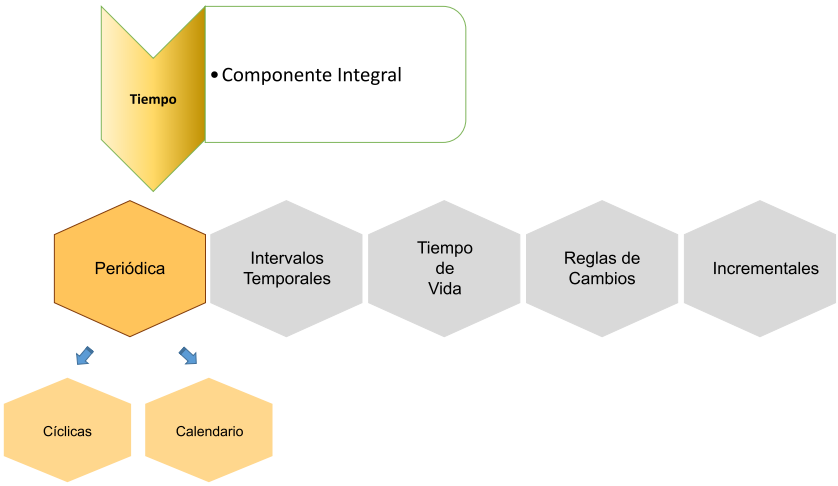

Figura 2.6: Componente Integral: Periódica.

En algunas ocasiones, los usuarios pueden estar interesados en el descubrimiento de RAs que ocurren de una forma recurrente. Por ejemplo, pueden estar interesados en reglas que representan comportamientos cíclicos de los consumidores en tiendas de comercio electrónico, y que permiten a estos comercios comprender mejor esos comportamientos para incrementar las ventas y mejorar sus beneficios [TTT12]; en reglas que nos permitan representar comportamientos de personas que se producen de forma cíclica a partir de la información que se encuentra almacenada en las grandes BDs de diferentes

redes sociales [TAKG15]. Estas reglas cíclicas solamente ocurren en determinados intervalos y no se dan en el resto de intervalos de tiempo. Las RAs cíclicas son similares a las RAs clásicas, pero representan dependencias que se producen de forma periódica a lo largo del tiempo [ORS98]. Un enfoque clásico para generar este tipo de reglas es utilizar el algoritmo Apriori para generar todas las RAs posibles en cada intervalo de tiempo definido por el usuario (el usuario define la partición del tiempo en intervalos de tiempo fijos). Después se aplica un simple algoritmo de búsqueda de coincidencias para detectar ciclos en las reglas generadas [ORS98]. Con el objetivo de mejorar la eficiencia del proceso de extracción de reglas, algunos métodos también introducen restricciones del usuario en dicho proceso [TAKG15], o son específicamente diseñados para obtener RAs cíclicas no redundantes [TK12]. Un enfoque más general fue propuesto por Chen [CP99], donde para cada regla se determina el conjunto de intervalos de tiempo contiguos más largo en el que se produce de forma periódica.

Sin embargo, dependiendo de los intervalos de tiempo que sean definidos (más generales o más específicos), las reglas cíclicas que se generan pueden ser diferentes. Para abordar este enfoque, las particiones temporales pueden definirse utilizando múltiples granularidades haciendo uso de un esquema de calendario (por ejemplo: semanas, días, horas, meses, etc.), reduciendo así la necesidad de conocimiento experto a priori para definir los intervalos de tiempo [RMS98b]. Los métodos clásicos para extraer RAs de calendario extienden el enfoque de extracción de RATs cíclicas [ORS98], generando RAs para cada unos los intervalos temporales definidos por el esquema de calendario y buscando a continuación cuales de estás reglas tienen un comportamiento periódico dentro de dichos intervalos [RMS98b]. Algunas propuestas introducen diferentes técnicas para mejorar la eficiencia en el proceso de extracción, como es la utilización de nuevas estructuras de datos que les permiten procesar de una forma más eficiente cada partición temporal [VV05, VVV05]. Sin embargo, las periodicidades que podemos encontrar en las BDs podrían no ser suficientemente precisas y podrían contener perturbaciones. Además, los usuarios tienden a definir una serie de restricciones temporales de una forma vaga y poco precisa. Por esta razón, algunos autores han hecho uso de la lógica difusa para definir el calendario, de una forma más cercana al lenguaje natural [INN04]. Por ejemplo, Lee y otros [LL04, LJL08] propusieron dos algoritmos en el que el esquema del calendario es definido haciendo uso de particiones difusas, proporcionando una mayor flexibilidad a la hora de introducir las restricciones impuestas por el usuario en el proceso de extracción.

Muchos métodos han sido utilizados para extraer reglas que presentan un comportamiento periódico, pero el concepto de periodicidad se define de una forma muy estricta y muchas de las reglas acaban siendo eliminadas en el proceso de extracción porque no siempre cumplen esta restricción. Con el objetivo de resolver este problema podemos encontrar algunas propuestas en la literatura que presentan el concepto de periodicidad de una forma más flexible [HGY98, FVYL+19]. Por ejemplo, Li y otros [LD05] extiende el enfoque utilizado por Han [HDY99], haciendo uso de ventanas deslizantes para determinar el momento temporal en el que la regla debe ocurrir para ser considerada periódica.

Categoría Intervalos Temporales

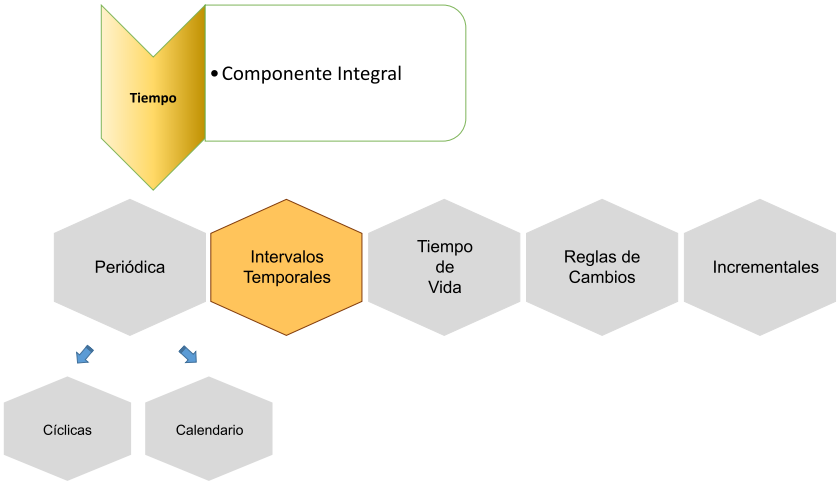

Figura 2.7: Componente Integral: Intervalos Temporales.

En muchos dominios de aplicación, las BDs temporales no solo incluyen datos con información sobre el instante de momento concreto en el que suceden, sino que también incluyen datos/eventos con información del tiempo que duran como parte de los datos de entrada (por ejemplo, la administración de vitaminas o medicamentos durante un mes), o incluyen abstracciones o interpretaciones derivadas de ellos como parte de un proceso de análisis de los mismos (por ejemplo, una semana de anemia) [BBJ98, MER08]. Esta información temporal debe considerarse en los procesos de extracción de RATs con el objetivo de obtener conocimiento en forma de reglas que sea consisten-

te a lo largo del tiempo. Por ejemplo, la duración de los fallos que se producen en los procesos industriales es una información esencial cuando intentamos encontrar correlaciones entre ellos. Por ello, Laxman [LUS02, LSU07] extendió la definición de clásica de episodio [MTV97] para incorporar restricciones en la duración de los eventos.

El descubrimiento de reglas temporales a partir de este tipo de BDs es más complejo y requiere enfoques diferentes a los que se utilizan con BDs temporales que contienen solo momentos temporales concretos en los que ocurren la información que contienen. Un intervalo involucra una duración y, por lo tanto, los patrones que se extraen en el proceso de aprendizaje deberían incluir una semántica diferente a la de simplemente "antes" y "después". La lógica de intervalos temporales de Allen y sus extensiones [All83, RM05] se utilizan habitualmente para representar las relaciones entre intervalos [HK01, MS15]. Por ejemplo, Winarko presentó ARMADA [WR07], un método basado en el algoritmo para extraer patrones secuenciales MEMISP [LL02]. ARMADA genera reglas de asociación temporales más ricas a partir de datos con intervalos de tiempo, teniendo en cuenta una restricción de tiempo de intervalo máximo para reducir el número de reglas generadas. Papapetrou y otros [PKSG09] propusieron tres métodos basados en el algoritmo clásico Apriori, los cuales introducen una serie de restricciones para descubrir relaciones temporales a partir de secuencias de eventos basadas en intervalos. Estos métodos generan un conjunto de reglas que son ordenadas en base a una medida de interés dada por el usuario y devuelven las K mejores reglas según el valor de la medida de interés. Lee y otros [LLC+09] presentaron una propuesta en la que mediante un pre-procesamiento de los datos generaban los intervalos de tiempo asociados a ellos, y a continuación extraen las reglas a partir de estos. Nazerfard [NRC11] propuso TEREDA [NRC11] para extraer aspectos temporales (como el instante de tiempo en el que suele iniciarse un evento y su duración) de las actividades diarias que realizan las personas. En [HZLH12] Huang y otros propusieron una extensión del algoritmo bien conocido "Progressive-Partition-Miner" (PPM) [LCL03] para identificar relaciones entre actividades académicas teniendo en cuenta la duración de las mismas. Finalmente, Coello y otros [CSC+11] presentaron una propuesta que permitía considerar requisitos del usuario durante el proceso de extracción para generar reglas interesantes para el usuario a partir de BDs de eventos que ocurren en momentos concretos o durante un intervalo de tiempo.

Algunos investigadores han utilizado también la lógica difusa dentro de esta subcategoría para proporner nuevos métodos de minado de reglas a

partir de datos basados en intervalos temporales. Sudkamp [Sud05] propuso un método basado en Apriori para extraer RATs difusas a partir de varias fuentes de datos haciendo uso de la lógica difusa para representar la duración temporal de los eventos (por ejemplo: un corto periodo) y las restricciones temporales (por ejemplo: poco después) de una forma cercana al lenguaje natural. Wu [Wu10] presentó una propuesta para extraer RATs difusas a partir de los registros de acceso de los usuarios a una página web, utilizando particiones difusas para representar el tiempo que los usuarios pasan visitando las diferentes secciones de la página web.

Este tipo de reglas han sido utilizadas satisfactoriamente en una gran cantidad de aplicaciones reales. Algunos ejemplos son: reconocimiento de actividades complejas del día a día en el área de la computación móvil, en la que se registran una buena cantidad de datos a través de los sensores de los dispositivos móviles de los usuarios, y que contienen una información realmente interesante [LWP+16]; predicción de diagnósticos médicos a partir de datos almacenados en BDs de hospitales que se recopilan a lo largo de las estancias de los pacientes en el hospital mientras estos se encuentran en tratamiento, ya que dichos datos incluyen información sobre eventos que se producen a lo largo del tratamiento y que pueden ser de gran interés para futuros tratamientos de pacientes [VJT+17]; predicción de inestabilidad cardiorespiratoria en pacientes a partir de los datos de monitorización continua que registran las medidas de los signos vitales fisiológicos [GBDW+17].

Categoría Tiempo de Vida (Lifespan)

Un aspecto importante en el descubrimiento de RATs es tener en cuenta la información temporal sobre la existencia de un item en la BD. El tiempo de vida (lifespan) de un item se define como el periodo de tiempo entre la primera ocurrencia del item en una transacción de la BD y la última vez que lo encontramos. Si esta información temporal se tiene en cuenta al realizar el cálculo del soporte de las reglas, el número de reglas que alcanza el umbral de minSop será mayor. Por ejemplo, consideremos una BD sencilla con 1000 transacciones asociadas a un comercio electrónico (como Amazon, etc.), en el que se han registrado alrededor de 100 transacciones de clientes por año durante un periodo total de 10 años, y seleccionamos como minSop de 0,5 % (50 transacciones) para el proceso de extracción. Al generar los itemset frecuentes, un producto que ha sido comprado 5 veces al año (50 veces en total) será considerado un item frecuente. Sin embargo, un nuevo producto que haya empezado a venderse en el último año y que haya sido comprado 40

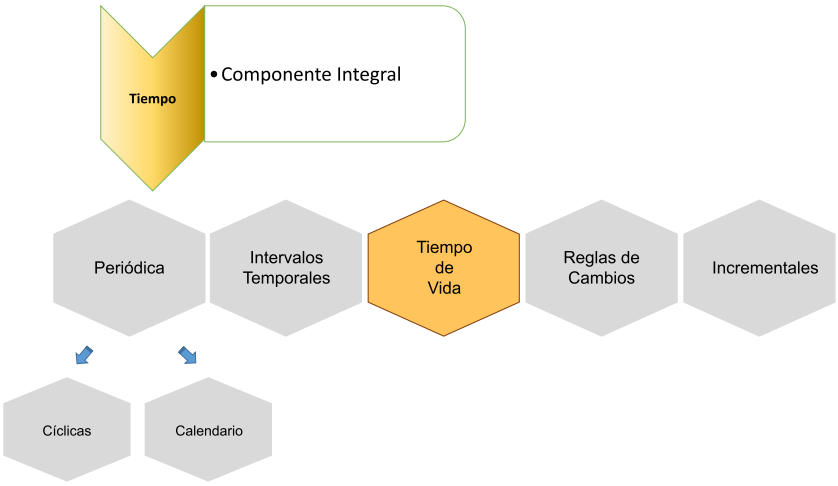

Figura 2.8: Componente Integral: Tiempo de Vida (Lifespan).

veces será un item infrecuente, debido a que no alcanza el umbral de minSop si consideramos la BD completa. Aunque el nuevo producto haya sido un éxito y durante su corto periodo de tiempo en el mercado haya logrado una gran cantidad de ventas (ha sido 8 veces más popular que el primero), este aparece en una pequeña cantidad de transacciones con respecto al total, por lo que no se considerará frecuente y no formará parte del conjunto final de reglas. Además, esta información temporal es utilizada para eliminar de la BD items/eventos poco frecuentes en la actualidad, reduciendo el tamaño de la BD y centrando el proceso de búsqueda en la información más relevante en la actualidad [AR00].

Esta información temporal también ha sido considerada por diversos investigadores para extraer RATs difusas teniendo en cuenta el lifespan de los items al calcular el soporte de las reglas [CLHL16]. Sin embargo, las funciones de pertenencia dadas podrían tener un influencia crítica en el conjunto final de reglas obtenidas. Por este motivo, algunas propuestas también han incorporado un aprendizaje o ajuste de las funciones de pertenencia en el que los lifespan de los items se tienen en cuenta al evaluar las soluciones generadas [CM19].

Este tipo de reglas han sido fundamentalmente generadas a partir en BDs de transacciones de diferentes tipos de compañías, para detectar y analizar las preferencias de compra de los clientes [CLHL16, CM19]. Sin embargo,

estos métodos pueden utilizarse en muchas otros tipos de aplicaciones y problemas reales. Por ejemplo, estos métodos podrían utilizarse en aplicaciones relacionadas con Internet de las Cosas (Internet of Things; IoT) para detectar fallos en las redes de sensores, o para detectar sensores cuya actividad se ha reducido de una forma considerable, de modo que estos pueden ser recolocados o eliminados de la red de sensores.

Categoría Cambios

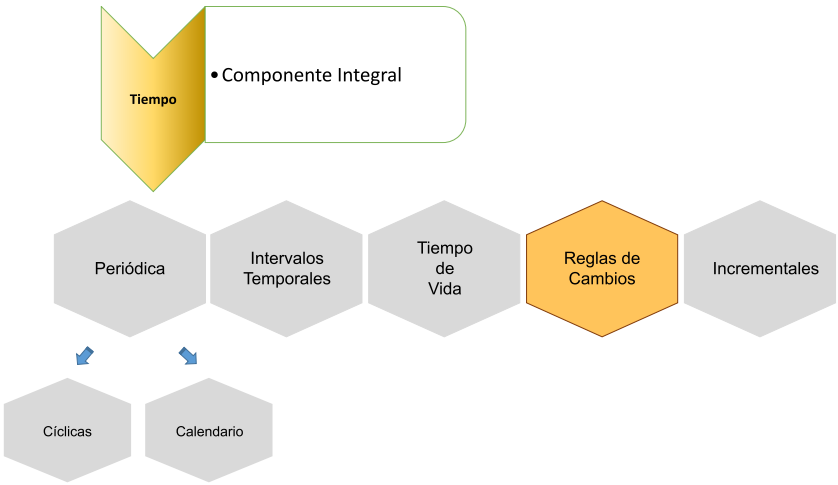

Figura 2.9: Componente Integral: Cambios.

El mundo real se encuentra en constante cambio, y dichos cambios se producen cada vez de una forma más rápida debido al enorme desarrollo tecnológico que vivimos a día de hoy. Es por ello que se vuelve esencial obtener información que nos permita detectar y evaluar estos cambios de condiciones a tiempo y de una forma inteligente, lo que nos permitirá responder a ellos de una forma adecuada [Boe11]. Por esta razón, un buen número de métodos han sido propuestos para detectar cambios en reglas y patrones extraídos desde colecciones de datos de diferentes periodos de tiempo. La idea principal consiste en introducir un nivel de abstracción que permita identificar cambios en las reglas y/o patrones extraídos haciendo uso de meta-reglas. Por ejemplo, Huang y otros [HHC16] propusieron un método para obtener RAs que representan cambios en el rendimiento de los estudiantes. Song [SKK01] presenta un método que detecta cambios en el

comportamiento de los consumidores a partir de las RAs generadas a partir de colecciones de datos de diferentes periodos de tiempo. Au y otros [AC02] presentaron un método para obtener meta-reglas difusas que detecten cambios en el soporte y/o la confianza de las reglas a lo largo del tiempo, siendo útiles para predecir cómo las RAs cambiaran en un futuro cercano.

Categoría Incrementales

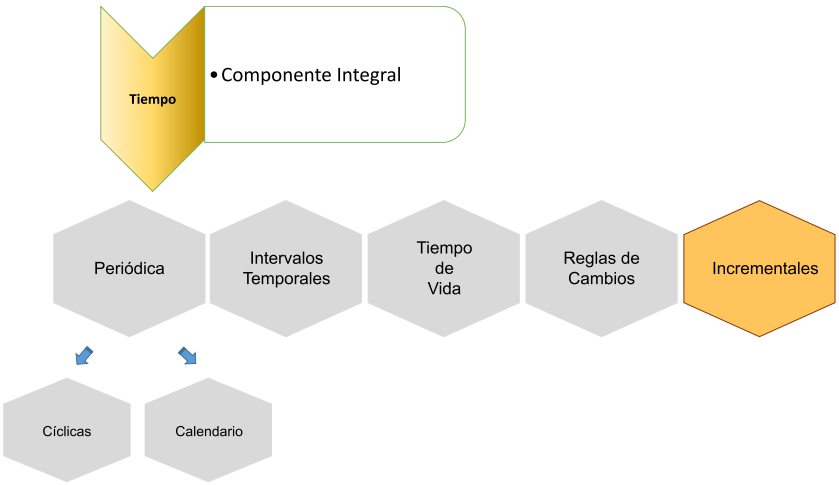

Figura 2.10: Componente Integral: Incrementales.

En la actualidad, las BDs temporales se actualizan constantemente, incrementando la cantidad de datos que almacenan. Debido a ello, el conjunto de RATs que haya sido obtenido a partir de este tipo de BDs necesita ser actualizado constantemente, eliminando las reglas que han dejado de ser frecuentes, y añadiendo reglas nuevas que representan relaciones entre los datos que no se daban anteriormente. Estos enfoques permiten que la información que disponemos de la BD este constantemente actualizada, proporcionando información útil en todo momento para ayudar a los expertos en la toma de decisiones. Re-ejecutar los algoritmos de extracción de RATs constatemente para actualizar los conjuntos de reglas ya existentes es un enfoque ineficiente, ya que al no considerar las reglas actuales en el proceso de extracción es necesario repetir parte del procesamiento que ya había sido realizado con anterioridad. Por ello, varios métodos en la literatura presentan soluciones para abordar de manera eficiente la extracción de RATs a partir

de BD incrementales. Gharib y otros [GNTA10] propusieron un algoritmo incremental basado en el algoritmo "Sliding-Window Filtering-[LLC01], en el cual los itemsets temporales frecuentes son almacenados para reducir el tiempo de procesamiento cuando la BD sea actualizada y haya que volver a generar reglas candidatas. Fouad y otros [FM17] propusieron un nuevo método eficiente de extracción de RATs incrementales, reduciendo el tiempo de cómputo haciendo uso de una estructura de datos eficiente para almacenar la información y almacenando los itemsets temporales frecuentes generados en ejecuciones anteriores.

Este tipo de reglas han sido utilizadas satisfactoriamente en una gran cantidad de aplicaciones reales. Teng y otros [TCL90] propusieron un método para extraer de reglas secuenciales que representaban anomalías en el comportamiento de los usuarios, haciendo uso de un motor inductivo basado en tiempo para adaptar las reglas a lo largo del tiempo. Zhou y otros [ZH17] propusieron utilizar un algoritmo genético para extraer reglas secuenciales que representen preferencias de los consumidores, las cuales son adaptadas mediante un algoritmo de optimización de colonia de hormigas a los cambios que se detectan en el comportamiento de los consumidores con el paso del tiempo.

2.5. Aplicación a problemas reales y herramientas software disponibles

La cantidad de datos que se recopila en diferentes áreas de aplicación en las que la información temporal está incluida en los datos, ya sea de forma explícita o de forma implícita, nos lleva a una situación en la que la extracción de conocimiento interesante y útil a partir de estas BDs temporales es muy atractiva y necesaria, suponiendo un desafío para los investigadores. Las RATs nos permiten generar modelos con un enorme poder predictivo y descriptivo, además de un mayor grado de interés, que puede aplicarse de forma satisfactoria en una amplia variedad de aplicaciones y problemas reales, como hemos visto en las distintas subsecciones en las que se introducían las categorías de la taxonomía. En la Tabla 2.3 podemos observar una lista reciente de aplicaciones y contribuciones que han sido abordadas haciendo uso de técnicas de extracción de RATs. A simple vista, podemos observar una gran cantidad de aplicaciones diferentes que están relacionadas con una amplia diversidad de áreas. Si nos fijamos en detalle, podemos

encontrar aplicaciones en la industria, seguridad, medicina y atención sanitaria, entre muchas otras áreas, con un número notable de propuestas aplicadas a las áreas de la medicina y la salud. Esto demuestra el gran potencial de las RATs, que pueden ser aplicadas de forma satisfactoria a una gran cantidad de problemas distintos. Además, el número de propuestas que se han aplicado a problemas reales ha ido incrementándose en los últimos años, proporcionando una mayor evidencia de la necesidad y del interés de estás técnicas para la resolución de problemas del mundo real.

Aunque un gran número de propuestas para extracción de RATs han sido publicados a lo largo de estos años, solamente unos pocos investigadores comparten el código fuente de sus propuestas y los códigos o descripciones disponibles en los libros y revistas a menudo presentan inconsistencias [Thi03]. Este problema, unido a la alta complejidad de algunas de las propuestas, impide que se extienda el uso de las técnicas de extracción de RATs en diferentes aplicaciones del mundo real. Además, a menudo los investigadores necesitan reimplementar los métodos propuestos por otros investigadores para realizar comparativas entre sus propuestas y las existentes en la literatura, lo que les lleva a perder una cantidad enorme de tiempo reimplementando algoritmos que en muchos casos no son tan eficientes como debería ser las versiones originales. Con el objetivo de abordar estos problemas, algunas herramientas de software de libre distribución y de código abierto han sido desarrolladas con el objetivo de facilitar el desarrollo de nuevas propuestas y su aplicación a problemas nuevos. Destacar que el modelo de liberación de código abierto facilita la aplicación y adaptación de estas técnicas a nuevos problemas y es esencial para su difusión en el sector empresarial [osi98].

SPMF [FVGG$^+$14] es una de las librerías de código abierto para tareas de minería de datos. Distribuida bajo la licencia GPL v3 y desarrollada en Java, por lo que puede utilizarse en una gran cantidad de plataformas y dispositivos. Esta librería incluye más de 110 algoritmos para la extracción de reglas secuenciales, predicción de secuencias, extracción de reglas de asociación, extracción de patrones secuenciales, extracción de patrones periódicos, extracción de episodios, y muchos más, además de incluir una buena cantidad de BDs temporales. SPMF puede ser embebida en otros programas desarrollados en Java de una forma sencilla, o puede ser ejecutado como un programa independiente desde linea de comandos o utilizando su interfaz gráfica, más accesible para usuarios menos avanzados. También podemos encontrar varios paquetes de R en el repositorio CRAN que permiten extraer reglas secuenciales, patrones secuenciales, etc, además de incluir multiples

Dominio de Aplicación	**Descripción (Referencias) Categoría**
Industria	Procesamiento de la señal [HT19]• Detección de patrones de uso de energía a partir de datos de contadores inteligentes [FDPD19] ⋄ Estudio del efecto domino en accidentes marítimos [WHSZ21]• Descubrimiento de asociaciones en el proceso de producción de hierro [HYYZ20]↑
Economía	Sistemas dinámicos de los mercados financieros [HYWC16] ◎
Seguridad	Seguridad a través del análisis de registro de eventos [KP18]•
Administración y análisis del Tráfico	Administración inteligente del tranporte [WZS+19]•, Congestión del tráfico [XWZ20] ◎ Análisis de la movilidad y demanda de taxis [GGB20]∘
Comportamiento humano	Reconocimiento de actividades del día a día [LWP+16]∘, Cambios en el rendimiento y características del aprendizaje de estudiantes [HHC16]↺, Detección de preferencias de consumidores [ZH17]↑ [PP18] ◅ [CM19] ⋄, Predicción de comportamientos humanos en compañías de fabricación [SY18]•, Análisis de diarios de viajes [VLLZ18]• Estudio de intereses de usuarios para el desarrollo de sistemas de recomendación [YNY21]∘ Predicción de comportamiento humano en el uso de smartphone [SK20]⋄
Medicina y Salud	Búsqueda de posibles interacciones entre medicamentos [JYT+16]•, Aplicaciones en bases de datos hospitalarios [VJT+17]∘, Detección de señales de toxicidad de fármacos [ABT+17], Previsión de inestabilidad respiratoria [GBDW+17]∘, Tratamientos de cáncer [NLPV18]•, Diagnóstico de enfermedades coronarias [ODS+18]∘, Predicción de fases en el abuso de fármacos [KKA+19]•, Sistemas de conocimiento sanitario [GKM19]∘, Enfermedades del corazón [BRBHEHA19]•

• Secuencial, ◎ Inter-Transaccional, ◅ Periódica, ∘ Intervalos Temporales,
⋄ Tiempo de Vida, ↺ Cambios, ↑ Incrementales.

Tabla 2.3: Aplicaciones recientes de propuestas de extracción de RATs agrupadas por dominio de aplicación.

BDs temporales [2]. Por ejemplo, el paquete *arulesSequences* proporciona in-

[2] `https://cran.r-project.org/web/packages/`

terfaces para las implementaciones en C++ del algoritmo cSPADE; el paquete *timetools* proporciona múltiples herramientas para manipular series temporales secuenciales y estacionales; y el paquete *eventInterval* proporciona funciones para el análisis del ratio de cambios en eventos secuenciales.

En la plataforma Github [3] también podemos encontrar numerosos repositorios que incluyen paquetes como *CRFSuite*, que es una implementación de *Conditional Random Fields* para problemas secuenciales de predicción, el paquete *Emotion Classification in Microblog Texts Using Class Sequential Rules*, que permite utilizar un clasificadora sociativo generado a partir de secuencias de eventos. También podemos encontrar una serie de paquetes de Python en el repositorio Pypi [4]. Por ejemplo, el paquete *Maximal Sequential Patterns Mining* es un *wrapper* para ejecutar todas las herramientas y métodos de SPMF desde código Python; y el paquete *prefixspan* proporciona varios métodos para la obtención de patrones secuenciales. También encontramos algunos métodos para la extracción de patrones secuenciales como parte de algunas de las plataformas software de código abierto de minería de datos, como puede ser Weka [WFH16], Knime [BCD+09] o Mahout [LP16].

Además podemos encontrar en repositorios de Internet múltiples BDs temporales que se encuentran disponibles como BDs open-source. Por ejemplo, la página web de SPMF proporciona un gran número de BDs en el formato propio de SPMF; y el paquete de Python *SecAlertSeqMining* permite la generación de BDs secuenciales en los formatos de entrada que utilizan los algoritmos de SPMF; e *IBM Generator* permite generar BDs secuenciales sintéticas.

2.6. Consideraciones y lineas de trabajo futuro

En los últimos años se han propuesto muchos algoritmos diferentes con el objetivo de extraer RATs. Cada publicación ha causado un gran interés en las novedades que presenta con respecto a las propuestas existentes anteriormente en la literatura. Pero, ¿cuales con los aspectos críticos de las publicaciones actuales? A continuación proporcionamos una serie de recomendaciones que deberían tenerse en cuenta en publicaciones futuras:

- Muchas de las propuestas que han sido publicadas para extracción de

[3] `https://github.com`

[4] `https://pypi.org`

RATs hacen uso o extienden algún método clásico para extraer RAs clásicas. Sin embargo, estos métodos suelen presentar problemas de complejidad (debido a la gran cantidad de reglas que son capaces de generar y al gran número de items que suelen involucrar en cada regla) y de escalabilidad (en términos de memoria y tiempo de computo necesario para su ejecución). En la literatura, podemos encontrar un buen número de propuestas recientes que tratan de realizar un proceso de extracción eficiente al mismo tiempo que tratan de optimizar diferentes medidas de interés que les permitan obtener conocimiento más conciso e interesante para el usuario [RKJ15, RAKJ19]. El uso de estas propuestas permitiría obtener RATs más interesantes, evitando al mismo tiempo los problemas de escalabilidad y complejidad.

- La publicación de nuevas propuestas requiere un análisis experimental comparando el comportamiento de la propuesta con respecto a los mejores algoritmos que fueron propuestos en publicaciones anteriores. Este tipo de análisis permite a los investigadores disponer de una base científica solida para justificar la efectividad de sus propuestas. Nosotros recomendamos realizar un estudio experimental comparando los resultados obtenidos con métodos clásicos ampliamente utilizados y conocidos junto con algunas de las propuestas más recientes de la literatura, haciendo uso de diferentes BDs de prueba y aplicando tests estadísticos siempre que sea posible. Sin embargo, los investigadores de este campo no siempre incluyen estos análisis en sus propuestas, o solo comparan con algoritmos clásicos que fueron superados hace muchos años. Esto hace realmente difícil para el resto de investigadores determinar cuales son las necesidades en este campo, y, por lo tanto, centrar sus esfuerzos de investigación para futuras propuestas significativas.

- Existen algunas herramientas de software libre y de código abierto para la extracción de reglas secuenciales, patrones secuenciales, etc., y para el manejo de BDs temporales, pero pocos investigadores comparten los códigos fuentes asociados a sus propuestas. Este problema, junto con la elevada complejidad de algunas propuestas, impide el uso generalizado de las RATs en aplicaciones reales y dificulta que los investigadores puedan comparar sus propuestas con otros algoritmos. Recomendamos a los autores que hagan público el código fuente de sus propuestas, ya que tendrá un impacto muy positivo para el desarrollo de nueva spropuestas futuras y para acercar este campo a nuevas áreas de aplicación del mundo real.

Desde nuestro punto de vista, aún hay mucho por hacer. Algunas de las posibles lineas de trabajo futuro dentro del campo de extracción de RATs son:

- La revolución tecnológica que se está produciendo en los últimos años está generando una gran cantidad de datos que acaban almacenándose en BDs, y ocultan un interesante conocimiento a la espera de ser descubierto. Por ejemplo, una de las principales fuentes de BDs temporales son las aplicaciones IoT, en las que se obtiene y registra información de una gran cantidad de sensores y dispositivos diferentes [ARKJ17]. Estas colecciones de datos temporales, en algunos casos conocidos como *Big Data*, suponen un reto para el proceso de extracción ya que la capacidad de procesamiento de la mayoría de las técnicas existentes no es suficiente para manejarlas. El uso de soluciones basadas en el paradigma MapReduce, en Deep Learning y Fog Computing entre otros, permiten a los investigadores tratar esta enorme cantidad de datos y aprovechar su dimensión temporal, aunque el uso de estas soluciones también introduce nuevos retos, como el mantenimiento de la seguridad y la privacidad de los datos. Aunque algunas propuestas de extracción de RATs han sido aplicadas a problemas Big Data aún hay mucho trabajo por hacer para lograr que existan propuestas eficientes y aplicables a este tipo de problemas.

- Algunas propuestas recientes también incorporan la información espacial en el proceso de extracción para obtener Reglas de Asociación Espacial-Temporal (RAET) [LZL+19], ya que la presencia de algunos eventos suele estar asociada a un periodo temporal y una ubicación concreta. En primer lugar, estas colecciones de datos también constan de atributos espaciales que determinan el ámbito de los objetos y su ubicación espacial, y las dependencias entre ellos no están representadas explícitamente en la base de datos. Por ello, es necesario realizar una etapa de preprocesamiento para transformar los datos espaciales en predicados y conectores espaciales (como *cercano a*, etc.), que permiten representar las relaciones topológicas. Este tipo de propuestas pueden suponer un problema debido al exceso de memoria que requieren para almacenar todos los datos temporales, por lo que se debe llenar a un punto medio entre pre-procesamiento y el proceso de extracción de reglas. Por otro lado, algunos autores han señalado que encuentran problemas de interpretación de las reglas espacio-temporales, debido a que los predicados espaciales como *cercano a* normalmente

dependen del propio usuario o de la información almacenada en la BD, ya que lo que para un usuario es cercano, para otro podría no serlo. Por este motivo, creemos que el desarrollo de técnicas que extraigan este tipo de reglas espacio-temporales es de gran interés, debido al aumento de BDs con información temporal y espacial que estamos viviendo en los últimos años.

- El descubrimiento de RATs de Alta Utilidad es otro área que está teniendo un cierto auge en los últimos años [ZFVW+15]. Muchos de los métodos publicados asumen que todos los items del conjunto de datos tienen la misma relevancia, sin embargo, en muchas de las aplicaciones reales cada item puede tener asociado una unidad de beneficio determinada, que indica cuál es la importancia de ese item en el problema (por ejemplo, el precio de un producto en una BD de ventas). El objetivo es extraer reglas que incluyan aquellos items con un alto valor de beneficio. Otro enfoque reciente consiste en considerar cada uno de los eventos de una secuencia como si tuviese un coste asociado para generar reglas interesantes (con valores elevados de las medidas confianza, Lift, etc) que incluyan items con un bajo coste para el usuario [DFVN17]. Este tipo de propuestas son realmente útiles para guiar el proceso de aprendizaje para proporcionar conocimiento mucho más interesante y útil al usuario, optimizando así los resultados y los beneficios de las posibles decisiones que puedan tomar asociadas a estos.

Como hemos podido apreciar a lo largo de este capítulo, resulta fundamental en muchos casos tener en cuenta la componente temporal de los datos para extraer reglas realmente útiles, especialmente en ciertos problemas en los que el conocimiento depende del intervalo o instante temporal en el que suceden los hechos. Este tipo de técnicas han sido aplicadas satisfactoriamente a una gran variedad de problemas reales, destacando la gran cantidad de propuestas en las áreas de la medicina y la biología, lo que muestra que estos son campos en los que este tipo de propuestas son de especial interés. Además, podemos ver como ligeramente hay una mayor proporción de propuestas dirigidas hacia la extracción de reglas secuenciales. También hemos propuesto algunas lineas de trabajo futuro, como es el descubrimiento de RATs de Alta Utilidad. Este enfoque es muy interesante ya que se introduce en el proceso de extracción una medida que determina la utilidad de los elementos de la BDs para el usuario, reduciendo el espacio de búsqueda al guiar la búsqueda hacia reglas de alta calidad que realmente son de utilidad para el usuario. En los próximos capítulos veremos una nueva propuesta pa-

ra la extracción de RATs de Alta Utilidad que permite abordar algunos de los problemas que actualmente están abiertos en este enfoque. Además, esta técnica ha sido aplicada sobre un problema real para extraer información útil sobre el problema de la obesidad en humanos.

Capítulo 3

Extracción de reglas secuenciales de alta utilidad media a partir de bases de datos de secuencias

Como hemos comentado, el descubrimiento de asociaciones es una de las técnicas de MD más utilizadas [HK06] para obtener información interesante y comprensible a partir de los datos, acorde a la Inteligencia Artificial eXplicable (denominada en inglés eXplainable Artificial Intelligence (XAI)) [BDRD$^+$20, FHC$^+$19], que ayude a los expertos a tomar buenas decisiones. Además, que la información obtenida sea comprensible es fundamental en muchas ámbitos como el de la medicina o la salud, ya que los investigadores necesitan comprender las asociaciones identificadas por las reglas para realizar un tratamiento personalizado y eficaz a cada paciente.

En muchas ocasiones, los datos de los que disponemos son secuencias de valores/eventos que se han ido almacenando a medida que se han ido produciendo. Esta información temporal es fundamental que sea considerada a la hora de identificar dependencias a partir de los datos para que las reglas obtenidas sean realmente útiles para los expertos. Por ejemplo, esta información temporal es fundamental que sea considerada para po-

der explicar los mecanismos de regulación de los procesos biológicos en las ciencias ómicas. Debido a ello, las reglas secuenciales han sido utilizas con éxito en una gran cantidad de problemas de diferentes áreas de aplicación [ARSDA+20, KKA+19, NLPV18], ya que permiten representar dependencias secuenciales entre valores/eventos (items) de BDs secuenciales. Estas reglas normalmente son evaluadas haciendo uso de diferentes medidas de interés (como suporte, confianza, lift, entre otras) para obtener y ordenar las reglas en base a su interés inherente al experto. Además, al filtrar las reglas en función de estas medidas evitamos las limitaciones para representar causalidad de los patrones secuenciales, haciéndolas más adecuadas para tareas de predicción y toma de decisiones.

Sin embargo, la mayoría de los algoritmos no prestan especial atención a la utilidad que tiene cada valor/evento del problema a la hora de extraer las reglas, considerando que todos tienen la misma utilidad para el usuario. No obstante, el valor de la utilidad permite centrar el proceso de búsqueda en las reglas secuenciales con items que son altamente relevantes para el experto en el problema de interés.

Por ejemplo, si aplicamos estos métodos sobre una base de BD de expresión génica, se consideraría que todos los genes tienen la misma importancia dentro del proceso biológico, sin embargo, los estudios clínicos han demostrado que algunos genes tienden a ser más significativos a la hora de explicar la causa de una enfermedad o son más eficaces para combatirla. Por ello, se han propuesto diferentes algoritmos para extraer reglas secuenciales intentando máximizar la utilidad de los valores/eventos que sean considerados en las reglas, dando lugar a lo que se conoce como Reglas Secuenciales de Alta Utilidad (RSAU) [ZFVW+15]. Esta medida de utilidad permite guiar el proceso de búsqueda, reduciendo el espacio de búsqueda al centrarse principalmente en los elementos de mayor utilidad. Sin embargo, puede dar lugar a un gran conjunto de reglas inútiles y poco comprensibles, ya que conjuntos de items con alta utilidad podrían ser ampliados con items con alto soporte pero con baja utilidad. Esto daría lugar a reglas que incluyen items de baja utilidad y de mayor longitud, por lo que son menos comprensibles para los expertos.

Los algoritmos genéticos (AGs) o, en general, Algoritmos Evolutivos (AEs) [ES03, Gol89] están considerados como uno de los métodos más interesantes para extraer conocimiento útil e interesante a partir BDs temporales [CM19, FHC+19, ZH19]. Estos métodos de búsqueda y optimización han sido aplicados a un amplio rango de problemas de extracción de conocimien-

to de forma efectiva. Estos métodos normalmente evolucionan un conjunto de soluciones (conocido como *población*) en función de un criterio de evaluación para aprender la mejor solución posible para el problema. En los últimos años, esta tarea también ha sido planteada como un problema multiobjetivo en el que varios objetivos son optimizados al mismo tiempo durante el proceso de aprendizaje, proporcionando un conjunto de soluciones con un buen equilibrio para los distintos objetivos optimizados [MGH13]. Así, los Algoritmos Evolutivos Multi-Objetivo (AEMOs; o MOEAs del inglés Multiple Objective Evolutionary Algorithms), nos proporciona una forma interesante de abordar esta tarea, ya que requiere que se genere un conjunto de reglas igualmente válidas con distintos valores para las medidas de interés en una sola ejecución, de forma que cada regla tiende a optimizar un objetivo en mayor medida que a otro [CVL07].

En este capítulo presentamos una nueva propuesta, llamada **HAUSrules**, para extraer reglas secuenciales interesantes, fáciles de comprender y de alta utilidad a partir de bases de datos secuenciales. Para ello hemos extendido el bien conocido AEMO NSGA-II [DAPM02] para realizar un aprendizaje evolutivo de las reglas maximizando dos objetivos (Utilidad e Interés). Para presentar esta propuesta, este capítulo ha sido organizado de la siguiente manera. En primer lugar veremos una introducción a los conceptos básicos de las Reglas Secuenciales de Alta Utilidad Media (RSAUM; o HAUSR del inglés High Average Utility Sequential Rules). Después se describen en detalle todos los componentes de esta propuesta. A continuación, se realiza un estudio experimental sobre una serie de bases de datos secuenciales, comparando los resultados estadísticamente con otros algoritmos de la literatura para valuar su comportamiento. Y finalmente, presentamos una serie de conclusión basadas en el estudio realizado.

3.1. Reglas secuenciales de alta utilidad media

Como hemos comentado, en muchas áreas los datos que se recogen están basados en secuencias de valores o eventos que se ordenan de forma lineal de acuerdo al instante de tiempo en el que ocurren. Esta información temporal es esencial para ser capaz de extraer conocimiento útil de los datos en un gran numero de problemas, por ejemplo, en problemas biológicos en los que permite obtener conocimiento que es capaz de explicar diferentes mecanismos que regulan diferentes procesos biológicos [ARSDA$^+$20, KKA$^+$19, NLPV18].

Las reglas secuencuales nos permiten representar dependencias secuen-

ciales entre valores o eventos (items) de una BD secuencial. Estas reglas son de la forma $X \rightarrow Y$, siendo X e Y conjuntos de items ordenados o no, y teniendo en cuenta que $X \cap Y = \oslash$, y que X debe ocurrir antes que Y en el tiempo basándonos en la ordenación secuencial que se realiza en base a la ocurrencia de los eventos [FWT+15]. Como ya se introdujo en el capitulo 2, para valorar el interés de estas reglas se hace uso de las medidas de calidad clásicas de RAs, el **soporte** y la **confianza**. En el caso de las reglas secuenciales, la definición de estas medidas se extiende de forma natural, donde el soporte de un itemset I ($Sop(I)$) se calcula como la división de la cantidad de secuencias que contienen I por el total de secuencias que conforman la BD. Por lo tanto, y recordando lo que se introdujo en el capítulo 2, podemos definir el soporte y la confianza como sigue:

$$soporte(X \rightarrow Y) : Sop(XY) \tag{3.1}$$

$$confianza(X \rightarrow Y) : \frac{Sop(XY)}{Sop(X)} \tag{3.2}$$

siendo $Sop(XY)$ el soporte del itemset XY, formado por la unión de los itemset X e Y, y teniendo en cuenta que al ser X el antecedente e Y el consecuente, debe X ocurrir antes de para que se tenga en cuenta la secuencia en el calculo del soporte.

Tal y como ya introdujimos en el capitulo 2, las medidas de calidad clásicas de RAs presentan una serie de problemas por los cuales se debe ser cauto a la hora de basarse en ellas para determinar si el conjunto de reglas obtenido es interesante o no.

La Tabla 2.1 muestra algunas de las medidas de interés más utilizadas para seleccionar las reglas en función de su interés para el usuario. La definición de estas medidas puede ser directamente extendida para reglas secuenciales haciendo uso de la definición de soporte para secuenciales. Una buena revisión bibliográfica sobre las distintas medidas de interés que hay en la literatura y sus propiedades puede encontrarse en [GH06].

La mayor parte de los algoritmos de extracción de RAs secuenciales no suelen prestar atención durante el proceso de aprendizaje a la utilidad final de las reglas en base al problema real que se está tratando. Sin embargo, tener un valor de referencia que indique la utilidad del item permite que el proceso de búsqueda se centre en los items o conjuntos de items que tienen una mayor utilidad para el experto, reduciendo el espacio de búsqueda y el conjunto de reglas generado. La utilidad de un itemset I en una secuencia

s en la que I ocurre ($util(I, s)$) se calcula como la suma de la utilidad en la secuencia s de cada item i que forma parte del itemset I (conocida como utilidad interna) multiplicada por su importancia o peso en el problema (conocida como utilidad externa). Es decir, cada item debe presentar dos valores de utilidad que indican el interés o su utilidad para el experto en el problema que se afronta, teniendo así la utilidad externa (utilidad global del item en el problema) y la utilidad interna (utilidad específica del item en una secuencia de la BD). Por lo tanto, si consideramos una regla de la forma $X \rightarrow Y$, la utilidad puede definirse como sigue:

$$utilidad(X \rightarrow Y) = \sum_{s \in D} util(XY, s) \tag{3.3}$$

donde D es una BD secuencial. Algunas de las propuestas tratan de generar conjuntos de reglas secuenciales de alta utilidad en las que el valor de utilidad de la regla sea mayor que un valor mínimo de utilidad ($minUtil$) indicado por el usuario. En otras ocasiones el objetivo consiste en extraer un top con las K reglas con mayor valor de utilidad, obteniendo así el conjunto de K reglas más prometedoras para el experto. Sin embargo, la longitud de los itemsets (el número de items que forman el itemset) influye directamente en su valor de utilidad y puede dar lugar a reglas inútiles para el usuario. Items con alto soporte y baja utilidad pueden ser añadidos a itemsets que tengan una alta utilidad, permitiendo que se generen a partir de ellos reglas que contienen items poco útiles para el usuario. Por este motivo, la utilidad suele dividirse por el número de items que conforman el itemset, obteniendo así la utilidad media con el objetivo de generar a partir de ellos RSAUM ($utilidadMedia(r)$) [HLW11, WLPF18]. Por ejemplo, consideremos la BD secuencial de la Tabla 3.1 con información sobre los cambios de expresión génica de 3 genes de 5 sujetos diferentes considerando los resultados de 2 análisis que se le han realizado a lo largo de un tratamiento determinado. Además, la Tabla 3.2 muestra la importancia que tiene la información aportada por cada gen para los expertos para comprender mejor la respuesta de los pacientes y ajustar el tratamiento. Teniendo en cuenta esta BD de ejemplo, el valor de utilidad media para la regla $r : (g_1 \uparrow) \rightarrow (g_3 \downarrow)$, se calcula como sigue:

- $utilidadMedia(r) = \frac{3 \cdot ((3 \cdot 8) + (3 \cdot 9))}{2} = 76,5$

Además, estas medidas de utilidad pueden combinarse con las medidas de interés, considerando de esta forma la información proporcionada por ambos

Tabla 3.1: Una BD con información sobre los cambios en determinados genes de 5 sujetos que reciben un mismo tratamiento.

Sujetos	Secuencias - Expresión de eventos de cambio en los genes en dos instantes de tiempo
s_1	$\langle\{(g_1 \uparrow, 3)\ (g_2 \uparrow, 1)\ (g_3-, 0)\}, \{(g_1-, 0)\ (g_2-, 0)\ (g_3 \downarrow, 3)\}\rangle$
s_2	$\langle\{(g_1-, 0)\ (g_2 \downarrow, 1)\ (g_3-, 0)\}, \{(g_1 \downarrow, 1)\ (g_2-, 0)\ (g_3 \uparrow, 1)\}\rangle$
s_3	$\langle\{(g_1-, 0)\ (g_2 \downarrow, 1)\ (g_3 \downarrow, 2)\}, \{(g_1 \downarrow, 1)\ (g_2 \uparrow, 1)\ (g_3-, 0)\}\rangle$
s_4	$\langle\{(g_1 \uparrow, 3)\ (g_2 \downarrow, 1)\ (g_3-, 0)\}, \{(g_1-, 0)\ (g_2-, 0)\ (g_3 \downarrow, 3)\}\rangle$
s_5	$\langle\{(g_1 \uparrow, 3)\ (g_2-, 0)\ (g_3-, 0)\}, \{(g_1-, 0)\ (g_2 \downarrow, 1)\ (g_3 \downarrow, 3)\}\rangle$

Tabla 3.2: Utilidad externa para cada uno de los genes indicado por un experto.

Gen	g_1	g_2	g_3
Utilidad	8	2	9

tipos de medidas a la hora de valorar las reglas obtenidas [DDG+20]. Por ejemplo, las reglas obtenidas pueden ser seleccionadas en función del valor que obtienen del producto entre las medidas factor de certeza y la utilidad media ($FC * utilidadMedia$). Las reglas secuenciales de alta utilidad media obtenidas durante la experimentación de nuestra investigación se evaluarán en posteriores secciones utilizando las medidas que se han introducido a lo largo de esta sección.

3.2. Algoritmo evolutivo multiobjetivo para la extracción de Reglas Secuenciales de Alta Utilidad Media: HAUS-rules

En nuestro trabajo de investigación hemos desarrollado, además de una propuesta de taxonomía, un nuevo método de extracción de reglas secuenciales de alta utilidad media, completando de esta forma otro de los objetivos de esta tesis. Para el desarrollo de este nuevo algoritmo hemos trabajado en la extensión de un AEMO bien conocido, como es el algoritmo NSGA-II [DAPM02]. Nuestra extensión, denominada HAUS-rules, persigue el ob-

jetivo de extraer un conjunto final de RSAUMs, con elevado interés para el experto. Además, se busca que las reglas obtenidas sean fácilmente interpretables, lo que permite que el usuario final pueda aprovechar al máximo el conocimiento obtenido. Para ello, HAUS-rules maximiza dos objetivos (Utilidad e Interés) e incluye dos componentes nuevos al algoritmo original: población externa y métodos de reinicialización de la población.

A continuación vamos a ver cada una de las características principales de nuestra propuesta. Veremos el esquema de representación de soluciones utilizado para codificar las soluciones en forma de cromosomas de la población, el proceso de generación de la población inicial, las funciones objetivo a maximizar, los operadores genéticos utilizados para el cruce y mutación de soluciones, y, finalmente, las novedades introducidas relacionadas con la población externa y los métodos de reinicio de la población.

3.2.1. Codificación del Cromosoma y Generación de Población Inicial

Un cromosoma viene representado por una sucesión de genes que representan los items envueltos en una regla secuencial. Cada gen esta formado por dos partes:

- La primera parte (ev) representa el item presente en la regla.
- La segunda parte (ac) indica si el item es parte del antecedente o del consecuente de la regla. Cuando esta parte tienen un valor '0', el item es parte del antecedente, mientras que si el valor es '1', indica que el item forma parte del consecuente de la regla.

Por lo tanto, un cromosoma C se representa como sigue, donde n representa el número de items que forman la regla representada por el cromosoma:

$$Gen_i = (ev_i, ac_i), \;\; i = 1, \ldots, n \;\; ,$$
$$C = Gen_1 Gen_2 \ldots Gen_n$$

La población inicialmente consiste en un conjunto de reglas secuenciales formadas únicamente por un item en el consecuente, de forma que se generen reglas fácilmente entendibles por el experto durante el proceso de aprendizaje, aunque el esquema de representación implementado permitiría incluir múltiples genes en el consecuente. En cuanto a la generación de la

población inicial, el proceso desarrollado selecciona de forma aleatoria una de las secuencias no marcadas de la BD, y a continuación, se seleccionan los items de esta secuencia que formarán parte de una de las reglas de la población inicial. Para seleccionar estos items se lleva a cabo una selección por torneo, en la que tendrán mayor probabilidad de ser elegidos aquellos eventos con un mayor valor de utilidad. Durante esta selección por torneo el antecedente de la regla podría crecer demasiado si no se introduce algún tipo de limitación, produciendo reglas demasiado grandes que dificultan la interpretabilidad de las mismas. Por ello, se limita el máximo de items que pueden formar parte del antecedente a $maxEvent$, un parámetro de nuestra propuesta y que debe ser determinado por el experto (por defecto es recomendable utilizar un valor de 5).

A continuación, todas las secuencias de la BD en las que ocurre la regla generada se marcan, de forma que no pueden volverse a utilizar para generar futuras reglas. Este procedimiento de generación de la población inicial se aplica de forma iterativa sobre las secuencias no marcadas de la BD hasta que se generan el número de reglas necesario para generar esta población inicial. El marcado de secuencias de la BD evita que pueda generarse siempre la misma regla o reglas muy similares, proporcionando diversidad al generar las reglas en la población inicial. Puede darse el caso, sobretodo en BDs con pocos ejemplo, en el que todas las secuencias de la BD hayan sido marcadas sin que se haya completado la población inicial. Para resolver este problema, una vez que se detecta esta situación, todas las secuencias de la BD se desmarcan, permitiendo así que se termine de generar la población inicial volviendo a utilizar todas las posibles secuencias para seguir generando reglas.

3.2.2. Objetivos

En este método maximizamos dos objetivos para la obtención del conjunto final de reglas: Utilidad e Interés. El objetivo de utilidad nos permite centrar los esfuerzos del proceso de búsqueda en la información que el experto considera más útil. Con este objetivo en mente, utilizaremos la utilidad media como medida ($UM(r)$), que tiene en cuenta la longitud de la regla cuando evalúa la utilidad de la misma. Esta medida nos permite evitar los problemas que suelen aparecer cuando se tiene en cuenta únicamente la utilidad, y que tienen que ver con la generación de reglas demasiado largas, lo que dificulta la interpretabilidad de la regla por parte del usuario [DDG$^+$20]. Tal y como ya se ha señalado por parte de diferentes investigadores en oca-

siones anteriores, cuanto mayor sea la cantidad de items que forma la regla, menos comprensible será para el experto [BDRD$^+$20, FHC$^+$19]. La medida de utilidad toma valores en el intervalo [0, ∞], por lo que las reglas que obtienen un mayor valor debería ser más útiles para el usuario, tal y como se ha introducido en secciones anteriores (ver sección 3.1).

El objetivo de interés se utiliza para tener en consideración el potencial interés de la regla para el usuario. Para ello, hemos elegido la medida de interés *Factor de Certeza* (FC) [SB75]. Como vimos en la Tabla 2.1 proporciona valores dentro del intervalo [-1, 1], en el que los valores negativos representan una correlación negativa entre los items, cero representa independencia y valores positivos representan correlación positiva entre los items. Además, en esta propuesta solo se generarán reglas *strong/fuertes* [BBSV02]. Una regla r tal que $X \rightarrow Y$ se considera *strong* o fuerte si cumple las siguientes condiciones:

- $Sop(r) >$ minSop
- $\neg Sop(r) >$ (1 - minSop)
- $FC(r) > 0$

Cada regla de la población es evaluada haciendo uso de un criterio de no-dominancia, a partir del cual una regla r_1 domina a otra regla r_2 si para todos los objetivos definidos la primera de ellas, r_1, tiene un valor mayor o igual a los valores de r_2, y en al menos uno de los valores el valor es mejor al valor de r_2. De este modo, la población actual se evalúa por completo siguiendo este criterio. En primer lugar, todas las reglas no dominadas se ordenan en un ranking y se marcan (reglas del Pareto). Después, todas las reglas no dominidas de la población actual, considerando solamente las reglas no marcadas, se ordenan formando un segundo ranking y se marcan. Este proceso se realiza de forma iterativa hasta que todas las reglas de la población actual son ordenadas. Finalmente, las reglas que comparten posición, por no ser mejores que otras, se ordenan entre ellas teniendo en cuenta la medida (*crowding*) (para una explicación más detallada consúltese [DAPM02]).

3.2.3. Operadores genéticos

Los operadores genéticos utilizados en nuestro algoritmo son los habituales operadores de cruce y mutación. El cruce de cromosomas se realiza

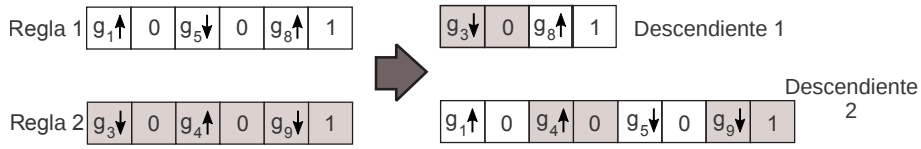

Figura 3.1: Operador de cruce.

llevando a cabo una selección aleatoria de dos cromosomas a través de un torneo binario. A continuación, se generan 2 descendientes a partir de los dos cromosomas elegidos intercambiando los genes de los padres de forma aleatoria. En la Figura 3.1 se puede apreciar un ejemplo gráfico sencillo de este operador genético de cruce a partir del cual dos cromosomas de la población actual permiten generar otros dos descendientes.

En cuanto al operador de mutación, éste se lleva a cabo seleccionando aleatoriamente un gen del cromosoma y modificando sus dos partes (*ev* y *ac*). El valor de *ev* se actualiza seleccionando al azar un items de la BD secuencial. Para la parte *ac* se selecciona un valor al azar del conjunto de valores {0, 1}.

Adicionalmente, se incluye un operador de reparación que se aplica después de que se haya realizado el cruce y la mutación de los descendientes. Para realizar la reparación, se comprueba si el antecedente y consecuente de la regla son consistentes, es decir, si al realizar el cruce o la mutación se ha generado una regla que no sea correcta. Si una regla no tiene ni antecedente ni consecuente, se selecciona un item al azar entre los items que no estaban incluidos en la regla. Para los consecuentes con dos o más genes, el consecuente se forma con uno de los genes seleccionados al azar y los restantes se reubican en el antecedente.

3.2.4. Población Externa y Mecanismo de Reinicialización

En esta propuesta hemos mejorado el modelo evolutivo propuesto en NSGA-II gracias a la incorporación de un mecanismo de reeinicialización de la población y a la introducción de una población externa que trata de mantener un nivel aceptable de diversidad a lo largo del proceso de búsqueda, al mismo tiempo que se mantienen todas las reglas obtenidas. La población externa actua a modo de repositorio de soluciones que permite mantener todas las reglas secuenciales no dominadas que se han obtenido. El tamaño de esta población no esta limitado, lo que permite obtener un mayor número

de reglas a partir del frente del Pareto, independientemente del tamaño de la población de nuestro modelo evolutivo. Además, esto permite reducir el tamaño de la población del modelo con el objetivo de mejorar la convergencia del proceso de búsqueda y utilizar un tamaño fijo en todas las BDs (puesto que proporciona independencia con respecto a la BD). Una vez que finaliza el proceso, la población externa normalmente contiene un número reducido de reglas secuenciales debido al hecho de que solamente se almacenan reglas del frente del Pareto, eliminando todas aquellas reglas que son redundantes.

El mecanismo de reeinicialización de la población permite introducir una mayor diversidad a la población, lo que nos permite evitar una convergencia prematura y asegurar que tendremos una diversidad suficiente para lograr un exploración mayor del espacio de búsqueda. Este mecanismo se ejecuta cuando el número de soluciones nuevas en la población actual es menor que un $\alpha\,\%$ de la población completa (siendo α un parámetro de entrada para el algoritmo cuyo valor recomendado es $5\,\%$). En ese caso, el algoritmo actualizará la población externa, marcará las secuencias de la BD en las que ocurren las soluciones y la población se reiniciará.

3.2.5. Modelo Evolutivo Multi-Objetivo

La población se inicializa haciendo uso del proceso de inicialización indicado anteriormente en la subsección 3.2.1, consideerando de esta forma reglas que proporcionan información de diversas zonas del espacio de búsqueda. A continuación, se construye una nueva población candidata a partir de una selección de cromosomas de la población actual, para que tras utilizar los operadores genéticos (cruce y mutación), se obtengan los nuevos individuos que forman la nueva población candidata. Finalmente, se obtiene la siguiente generación de la población a partir de la población candidata, aplicando una selección elitista de las mejores reglas de la población actual y de la población candidata. Después de esto, aquellas reglas no dominadas de la población actual se añaden a la población externa, eliminando aquellas reglas redundantes. Por último, si el porcentaje de reglas nuevas de la población actual es menor que $\alpha\,\%$, se aplica el mecanismo de reinicialización de la población, actualizando así la población externa y reiniciando la población actual. Este proceso se repite hasta que se satisface alguna condición de parada, que habitualmente es un número máximo de evaluaciones de las funciones objetivo.

3.3. Estudio Experimental

Para evaluar el buen comportamiento del algoritmo propuesto se ha realizado un estudio experimental sobre una serie de BDs secuenciales, comparando los resultados obtenidos estadísticamente con otro métodos de la literatura. De esta forma, se trata de demostrar la eficacia de la nueva propuesta frente a las ya existentes, mostrando así sus puntos fuertes y el interés del desarrollo de ésta.

3.3.1. Experimentos

En este estudio hemos utilizado 10 BDs secuenciales que incluye valores de utilidad para cada uno de los items que las conforman. La cantidad de secuencias que forman dichas bases de datos va desde 730 secuencias, de la más pequeña, hasta las 77.512 secuencias de la más grande. En cuanto a la cantidad de items, tenemos BDs que van desde 250 items, para la más pequeña, hasta los 13.905 de la más grande. En la Tabla 3.3 podemos observar la cantidad de secuencias, la cantidad de items o eventos diferentes y la longitud media de cada una de las secuencias para cada una de las BDs utilizados.

Las primeras 5 BDs (Bible, BMS, FIFA, Kosarak y SIGN) han sido obtenidas del repositorio que la librería SPMF [FVGG$^+$14] tiene disponible en su página web [1]. Las otras 5 BDs (Sintética 1-5) han sido generadas de forma sintética utilizando las herramientas disponibles en SPMF para la generación de BDs secuenciales. Para introducir la utilidad de los items que forman parte de las BDs, y que es necesaria para la comparativa, se ha utilizado una distribución normal logarítmica [TSWY13]. Para cada una de las BDs, nuestra propuesta ha sido ejecutada con 5 semillas diferentes, por lo tanto, los valores que se muestran en las tablas son los resultados medios de esas ejecuciones.

Nuestra propuesta ha sido comparada con métodos bien conocidos para extraer reglas secuenciales. Como son **CMRules** [FVFNN12] y **HUSRM** [ZFVW$^+$15], los cuales están disponibles en la librería de código abierto SPMF [FVGG$^+$14].
En el caso de CMRules se trata de un algoritmo basado en el algoritmo clásico de extracción de RAs Apriori [SA96], y que permite extraer reglas secuenciales (sin restricciones temporales entre eventos que forman parte del antecedente o del consecuente) que ocurren de forma frecuente en varias

[1] Web de SPMF: `https://www.philippe-fournier-viger.com/spmf/`

Tabla 3.3: Resumen de las principales características de las BDs.

BD	*Secuencias*	*Items*	*Longitud Media*
Bible	36.369	13.905	21,6
BMS	77.512	3.340	4,62
FIFA	20.450	2.990	34,74
Kosarak	10.000	10.094	8,14
SIGN	730	267	51,997
Sintética 1	1000	250	45,6
Sintética 2	2000	250	49,8
Sintética 3	5000	300	50,2
Sintética 4	7000	300	49,7
Sintética 5	10000	300	52,01

secuencias de la BD. Por otro lado, HUSRM es un algoritmo que permite generar reglas secuenciales de alta utilidad en una sola pasada. Para ello hace uso de una nueva estructura de datos a la que llaman *tabla de utilidad* (utility-table), y de varios métodos de optimización que le permiten extraer las reglas de alta utilidad de una forma eficiente.

En la Tabla 3.4 se muestran los valores que han sido utilizados para los parámetros de cada uno de los métodos. Estos valores son los propuestos por los propios autores en los artículos de las propuestas originales, que han mostrado funcionar correctamente en la mayor parte de los casos. Señalar que hemos adaptado el valor de utilidad mínima ($minUtil$) para cada uno de las BDs en los algoritmos HUSRM y CMRules, puesto que estos algoritmos no son capaces de ejecutarse en todas las BDs con un valor fijo para este umbral. Además, en el caso de CMRules, hemos utilizado 0,005 como valor de $minSop$ en la BD Kosarak, ya que con valores superiores no ha sido posible extraer reglas, y en el caso de SIGN y todas las BDs sintéticas hemos utilizado un valor de 0,4 para $minSop$, al presentar problemas de escalabilidad al ejecutarse.

3.3.2. Comparación con otros enfoques para extraer RSAUMs

La Tabla 3.5 muestra la media de los resultados obtenidos por cada uno de los métodos analizados en las BDs, indicando entre paréntesis la

Tabla 3.4: Parámetros utilizados para la comparación.

Algoritmo	*Parámetros*
CMRules	*minConf = 0,7, minSop = 0,01, minUtil = Adaptada*
HUSRM	*minConf = 0,7, maxLongitud = 15, minUtil = Adaptada*
HAUS-rules	*TamañoPoblación = 200,* N_{eval}*=150.000,* $Prob_{mut}$*= 0,1,* $\alpha = 5$

Tabla 3.5: Resultados obtenidos para las medidas de interés en cada BD.

Algoritmo	$\#R(\sigma)$	$\#Items(\sigma)$	$Sop(\sigma)$	$Conf(\sigma)$	$Lift(\sigma)$	$FC(\sigma)$	$Yule'sQ(\sigma)$
CMRules	348,22 (349,54)	**3,27** (0,51)	**0,38** (0,27)	0,83 (0,03)	3,52 (4,31)	0,20 (0,38)	-0,24 (0,79)
HUSRM	**10164,30** (16161,56)	9,15 (2,34)	0,07 (0,13)	0,82 (0,08)	8,74 (19,35)	0,59 (0,27)	0,42 (0,48)
HAUS-rules	15,00 (10,02)	3,32 (0,41)	0,22 (0,14)	**0,89** (0,10)	**8,84** (16,43)	**0,68** (0,12)	**0,73** (0,20)

desviación estándar (σ). En esta tabla podemos ver la media de la cantidad de reglas obtenida por cada algoritmo en cada una de las ejecuciones ($\#R$), y la media de la cantidad de items que forman parte de las reglas ($\#Items$), y la media de las diferentes medidas de interés: soporte (Sop), confianza ($Conf$), Lift ($Lift$), Factor de Certeza (FC) y Yule'sQ ($Yule'sQ$).

Por otro lado, la Tabla 3.6 muestra los resultados obtenidos para las medidas de utilidad, donde UM y $FC * UM$ son, respectivamente, la el valor medio de la medida UM y del producto de FC y UM de las reglas que se han obtenido (ver sección 3.1).

A simple vista, podemos observar en las Tablas 3.5 y 3.6, que nuestra propuesta obtiene un conjunto reducido de reglas de alta utilidad media, presentando valores medios similares o incluso mejores que los obtenidos por el resto de métodos, tanto para las medidas de interés como para la medida $FC * UM$. Esto pone de manifiesto su capacidad para obtener reglas con un buen equilibrio entre interés y utilidad para el usuario. Además, las reglas obtenidas por nuestro método están formadas por una cantidad reducida de items o eventos, lo que significa que éstas deberían ser más sencillas de interpretar y, por tanto, de comprender.

Para comparar los resultados por los métodos analizados hemos aplicado tests no-paramétricos para la comparación múltiple [GMLH09][2] sobre los resultados medios obtenidos para las medidas de confianza, FC, Yules'Q y

[2] `http://sci2s.ugr.es/sicidm/`

Tabla 3.6: Resultados obtenidos para las medidas de utilidad en cada BD.

Algoritmo	$UM(\sigma)$	$FC * UM(\sigma)$
CMRules	**28.874,54** (31541,28)	649,88 (3626,57)
HUSRM	10.233,82 (26.248,40)	5.737,47 (15.863,02)
HAUS-rules	23.986,97 (22.207,98)	**17.126,37** (15.853,04)

$FC * UM$. Estos valores han sido previamente normalizados al intervalo [0, 1], ya que es necesario que los valores se encuentren dentro de dicho intervalo para la aplicación de los tests. No se ha tenido en cuenta la medida de interés Lift en este análisis, puesto que esta medida representa dependencia positiva cuando su valor es > 1 (que es el caso en todos los métodos comparados), y valores mayores para esta medida no quieren decir que necesariamente las reglas obtenidas tengan un mayor interés para el usuario. Además, se ha considerado la medida de utilidad $FC * UM$ para considerar la utilidad de las reglas obtenidas.

Analizando los resultados mostrados en la Tabla 3.7, podemos ver como el test de Friedman [Fri37] rechaza la hipotesis de igualdad, proporcionando el mejor ranking a HAUS-rules. A continuación, ha sido aplicado el test de Shaffer [Sha86], con el objetivo de comparar por parejas cada una de las propuestas. En la Tabla 3.7 podemos observar el valor de p ajustado (VPA), para el cual se rechaza la hipótesis de igualdad de nuestra propuesta con respecto al resto de métodos analizados.

Dados los VPA, se puede observar que existen diferencias significativas entre los resultados obtenidos por cada propuesta, con un nivel de significancia de al menos 0,1. Como consecuencia de esto, teniendo en cuenta los resultados, podemos deducir que nuestra propuesta ha sido la que mejor rendimiento ha mostrado frente al resto, obteniendo unos resultados significativamente mejores.

3.4. Sumario

En este capítulo hemos propuesto **HAUS-rules**, un nuevo AEMO diseñado para la extracción de RSAUMs a partir de BDs secuenciales. Esta propuesta extiende el ya ampliamente conocido algoritmo NSGA-II, para llevar a cabo un aprendizaje evolutivo de las reglas secuenciales, introduciendo una población externa y un mecanismo de reinicialización de la población en

Tabla 3.7: Resultados de los tests estadísticos para las medidas de confianza, FC, Yules'Q y $FC * UM$.

Algoritmo	Ranking $Conf$	Ranking FC	Ranking $Yule'sQ$	Ranking $FC * UM$	VPA $Conf$	VPA FC	VPA $Yule'sQ$	VPA $FC * UM$
CMRules	2,25	2,75	2,55	2,57	0,1	0,005	0,015	0,012
HUSRM	2,35	1,9	2,13	2,15	0,095	0,057	0,059	0,057
HAUS-rules	**1,4**	**1,35**	**1,3**	**1,3**	-	-	-	-

el modelo evolutivo para mejorar la diversidad de las soluciones evaluadas en el proceso de búsqueda y preservar todas las reglas no-dominadas. Además, esta propuesta trata de maximizar dos objetivos (utilidad e interés), con la idea final de obtener reglas que sean interesante, fáciles de entender y de una gran utilidad para el usuario.

Los resultados que se han obtenido a partir de las diez BDs utilizadas para la comparación muestran que nuestra propuesta permite obtener un conjunto reducido de reglas secuenciales de alta utilidad media, con unos valores altos de interés y utilidad en todas las BDs. Además, estas reglas están formadas por un reducido número de items, lo que facilita la comprensión de las reglas por el usuario final. Tal y como hemos podido observar, no solo los resultados son buenos a simple vista, sino que tras realizar una serie de tests estadísticos se ha podido probar que existen diferencias significativas con los resultados obtenidos por los métodos analizados. Tras quedar probada su eficacia, en el siguiente capítulo se aplicará la propuesta a un problema real, lo que permitirá determinar si es posible obtener información realmente útil para la resolución de problemas del mundo real.

Capítulo 4

Análisis temporal sobre un estudio longitudinal in vivo de la expresión genética en tejido adiposo humano

Tal y como ya se ha podido observar en anteriores capítulos, la extracción de reglas secuenciales es un campo de gran interés, especialmente en el área de la medicina y la biología, donde hemos podido observar en la sección 2.5 que se trataba de una de las áreas con mayor cantidad de contribuciones. Es por ello que se ha aplicado nuestra nueva propuesta a un problema real de este ámbito. Una de las principales novedades del estudio es que se realiza sobre la expresión genética **in vivo** en tejido adiposo humano, mientras que buena parte de este tipo de estudios se realizan **in vitro**, tal y como ocurre en el trabajo realizado por Nam y otros en [MH09], en el que se tratan de identificar RATs a un conjunto de microarrays genéticos. Para la realización de este estudio se ha contado con la participación de varios expertos del Instituto de Nutrición y Tecnología de los Alimentos "José Mataix Verdú".

Las reacciones bioquímicas que se producen en los humanos son sistemas complejos que están sujetos a interacciones regulatorias entre los genes. En estos sistemas regulatorios, existe un cierto desfase temporal en el modo en

el que los genes se relacionan los unos con los otros. Es decir, debe pasar un periodo de tiempo entre el momento en el que una proteína de un gen se comienza a traducir hasta que se produce la regulación final [LK17]. El paso del tiempo necesario para que se produzca la regulación de un gen puede apreciarse especialmente en los ensayos longitudinales que se realizan a seres humanos. En esos casos, las modificaciones identificadas en los patrones de expresión de los genes de un tejido genético podrían representar la acción molecular de agentes externos, y provocar modificaciones posteriores en los niveles de otros genes del ácido ribonucleico mensajero (ARNm).

Las plataformas de *microarrays* genéticos han permitido la creación de repositorios genómicos, que proporcionan un gran número de perfiles diferentes de expresión del genoma [BWL+12]. El incremento de experimentos con enormes microarrays temporales abre un nuevo mundo de posibilidades, como puede ser el descubrimiento de interesantes retrasos temporales en las relaciones entre pares de genes. Si a esto le añadimos que el descubrimiento de asociaciones es una de las técnicas de MD más comunes e interesantes para obtener conocimiento interesante e interpretable, tenemos los ingredientes necesarios para aplicar este tipo de técnicas a estos problemas.

Concretamente, y como ya se ha introducido anteriormente, las técnicas de extracción de patrones secuenciales han sido ampliamente utilizados con éxito, demostrando su eficacia, en problemas en los que se buscaba extraer asociaciones secuenciales que se produjesen de forma frecuente en genes a lo largo del tiempo [FVLK+17, LXL+17, NSET20]. Sin embargo, la mayor parte de las contribuciones que se han realizado a lo largo de los últimos años, asumen que todos los genes que intervienen tienen la misma importancia dentro del proceso biológico. Esto no es así, y se ha demostrado en múltiples estudios clínicos que algunos genes tienden a ser más importantes en ciertos procesos genéticos, dando lugar a enfermedades concretas o siendo más efectivos que otros para luchar contra éstas. Lo que demuestra que tener en cuenta únicamente la frecuencia en la que se producen ciertos datos o en la que actúan ciertos genes no es necesaria para poder identificar las secuencias más interesantes que tienen lugar en los procesos genéticos. Es por ello que tener en cuenta la utilidad o importancia de ciertos genes en el estudio de los procesos genéticos puede permitir que se obtengan relaciones más interesantes para los expertos, pudiendo así obtener patrones secuenciales con una alta utilidad o interés a partir de una BD de microarrays temporales [FVLN+19, ZDA17]. No obstante, es importante tener en cuenta la longitud de los patrones, ya que cuanto mayor longitud mayor será el valor de utilidad del mismo. Sin embargo, a mayor

longitud menor será la capacidad de interpretar el patrón y el conocimiento que éste representa, siendo necesario obtener patrones con una alta utilidad media [DDG+20, HLW11, WLPF18].

Queda claro que uno de los ámbitos en los que resulta más interesante la aplicación de técnicas que permitan detectar relaciones secuenciales de alta utilidad entre items o eventos es la biología y la medicina, que es precisamente una de las principales áreas en la que encontramos contribuciones de este tipo. Por ello, hemos aplicado el nuevo algoritmo de extracción de reglas secuenciales de alta utilidad media a un problema real biológico en el que se trata de determinar cuales son los genes que tienen una mayor importancia en los procesos genéticos que se producen tras llevar a cabo un tratamiento medico concreto. En este problema, nuestro algoritmo permitirá obtener reglas con una gran utilidad, aunque al guiarse para maximizar la utilidad media de la regla y no el valor directo de la medida de utilidad, se obtendrán reglas con una longitud menor, y por tanto, más fácilmente comprensibles por el experto, lo que puede permitir determinar si dichos tratamientos son realmente eficaces.

A lo largo de este capítulo estudiaremos la aplicación de nuestra nueva propuesta a dicho problema, con el objetivo de permitir un análisis temporal sobre los datos obtenidos mediante un estudio longitudinal in vivo de la expresión genética de tejido adiposo en humanos. Este problema se expondrá en la primera sección. A continuación, se introducirán una serie de medidas de calidad biológicas, que se calcularán a las reglas, y que permiten al experto determinar como de interesantes son realmente estas reglas desde el punto de vista biológico. Finalmente, se exponen los experimentos realizados, junto a los resultados y conclusiones obtenidas.

4.1. Microarray genético temporal sobre tejido adiposo en seres humanos

Se va a aplicar la propuesta a un problema real en el que disponemos de un microarray temporal de genes obtenido de seres humanos vivos que han estado bajo dos tipos de tratamientos distintos. Estos tratamientos son dietas alimenticias diferentes que se han aplicado a 57 pacientes con obesidad durante un largo periodo de tiempo [VRA+16]. A estos 57 pacientes se les ha asignado una de las dos dietas en estudio de forma aleatoria. Las dietas a estudiar son las siguientes:

- Dieta baja en calorias (*low-calorie diet*; LCD): Se trata de una dieta de 12 semanas de duración en la que los individuos deben alimentarse con un máximo de 1.250 kilocalorias al día.

- Dieta muy baja en calorias (*very-low-calorie diet*; VLCD): Se trata de una dieta de 5 semanas de duración en la que los individuos deben alimentarse con un máximo de 500 kilocalorias al día.

La idea del estudio es determinar si, más allá de las restricciones de alimentación y tiempo de cada una de las dietas, alguna de ellas tiene efectos significativamente distintos y mejores (de forma que el experto pueda determinar cual de ellas puede tener un mejor resultado) y qué genes intervienen en los procesos genéticos derivados en cada una de las dietas. Para la obtención de los microarrays genéticos, se realizan biopsias en 3 instantes de tiempo diferentes a lo largo del tratamiento (al inicio, después de la perdida de peso y en la etapa final de estabilización de peso) [VRF+17]. La BD está disponible para su uso en el repositorio de expresión genética Omnibus (Gene Expression Omnibus; GEO)(ID:GSE77962) [BWL+12].

Los datos de intensidad fluorescente obtenidos a través de las intervenciones a los pacientes se transformaron a una matriz numérica de valores relativos a la abundancia de ARNm, en la que las filas se corresponden con los pacientes y las columnas se corresponden con las sondas de los genes analizados. De entre todos los individuos tratados con las dietas, 22 de los individuos tratados con la dieta LCD y 24 de los tratados con la dieta VLCD presentaron datos de expresión genética válidos y medidas de fluorescencia válidos para más de 33.000 sondas. Todos los datos recogidos a lo largo de los tratamientos se han combinado en una única matriz, en la que cada individuo presenta entradas consecutivas que se corresponden a las tres recogidas de datos de microarrays genéticos (un microarray por individuo e instante temporal). Por tanto, el número total de filas que se incluyen en la matriz es de 138.

Sobre la matriz de datos obtenida se ha aplicado un procesamiento inicial para reducir la cantidad de sondas, seleccionando únicamente aquellas que mostraban diferencias en su expresión entre los diferentes instantes de tiempo en los que se toman los datos. Para identificar las sondas cuya expresión es diferente en distintos instantes de tiempo se comprueba si la expresión del gen cambia entre los instantes de tiempo analizados, generando dos periodos de tiempo a partir de los tres instantes de tiempo iniciales: periodo de perdida de peso (periodo de la primera recogida de datos a la segunda)

y periodo de estabilización (periodo entre la segunda recogida y la tercera). Si en estos periodos existe un cambio de expresión en algún gen (cambio de expresión entre una recogida y otra), entonces se entiende que existen diferencias en la expresión, por lo que dicha sonda se tiene en cuenta para el posterior proceso de extracción de conocimiento. Sin embargo, si no existe cambio de expresión en ninguno de los periodos, dicha sonda se descarta. De esta forma es posible reducir la cantidad de sondas, que al inicio es de 33.000, lo que genera un espacio de búsqueda enorme realmente difícil de afrontar con la capacidad computacional disponible para el estudio.

Para determinar si existe un cambio significativo en la expresión de los genes, se aplican tests estadísticos. Concretamente, si las sondas evaluadas presentan un valor p ajustado de Bonferroni menor de 0,05 y un ratio de señal logarítmica $\geqslant 1$ o $\leqslant -1$ (*Log2(FoldChange)*) en un $t-test$ por parejas con una corrección Bayesiana [She03], entonces se entiende que se produce un cambio significativo en la expresión del gen. De esta forma, la selección final de sondas es de 431, que se corresponden con un total de 398 genes únicos, lo que reduce en gran medida las más de 30.000 sondas iniciales.

Tras la selección de sondas a partir del estudio de cambios en la expresión de genes, se lleva a cabo un proceso de normalización de los datos, para lo que se aplica un algoritmo de medias robustas multichip o multiarray (Robust Multiarray Average; RMA) [IBC$^+$03]. Esta discretización permite principalmente transformar los datos de entrada para hacerlos más comprensibles para los expertos y para poder hacer un uso más sencillo de los mismos, además de permitir una homogenización más sencilla de las diferentes BDs [GCC$^+$16]. De esta forma, tras aplicar este algoritmo de discretización, se obtiene una matriz con un total de 431 sondas por cada grupo de individuos (dieta). Cada una de esas sondas presenta tres posibles valores: 0 (ausencia de cambio), 1 (regulación negativa, menor expresión) y 2 (regulación positiva, mayor expresión).

Finalmente, los datos de la matriz se utilizan para generar una BD secuencial para cada uno de los grupos de la dieta, correspondiendo cada una de las secuencias de la BD a un individuo. Cada uno de los eventos de la secuencia nos indica si durante dicho periodo temporal se produce o no un cambio de expresión, y si dicho cambio es positivo o negativo. Cada uno de los items se representa con un número de 4 cifras, en el que cada una de las cifras representa información sobre los datos genéticos:

- Primera cifra: indica si el cambio en la expresión del gen es positiva (2), negativa (1) o si no se produce cambio (0).

- Cifras Segunda, Tercera y Cuarta: representan el número identificador de la sonda evaluada.

De esta forma, conseguimos transformar la matriz inicial con el microarray que contiene la información genética de los pacientes a una BD secuencial a la que se puede aplicar nuestra propuesta de extracción de reglas de alta utilidad media. Sin embargo, para la extracción de reglas secuenciales de alta utilidad media es necesario definir cual es el valor de utilidad para cada uno de los genes. Para la utilidad interna de cada item o gen, se calcula la cantidad de ocurrencias de dicho gen en la base de datos *Phenopedia* [YCKG09], que es una fuente online y actualizada que recopila todas las asociaciones genéticas en humanos que se han reportado en el Centro Nacional para la Información Biotecnológica (National Center for Biotechnology Information; NCBI) hasta la fecha. Para la utilidad externa de cada uno de los genes, se utiliza el valor de cambio (Fold Change) de expresión del gen en cada periodo de tiempo. Gracias a estos dos criterios sugeridos por el experto, es posible obtener reglas guiando el algoritmo durante el proceso de extracción para proporcionar conocimiento de mayor utilidad para el experto.

4.2. Medidas de Calidad Biológica para las reglas

Para determinar los valores de utilidad, tal y como se ha introducido en la sección anterior, se aprovecha el conocimiento del experto y los estudios que se han realizado hasta ahora sobre relaciones e interacciones genéticas. Sin embargo, Además de las medidas de interés propuestas por el experto para cada gen/item (que permiten determinar la utilidad de estos dentro del proceso biológico), se han diseñado cinco nuevas métricas biológicas con la intención de determinar qué reglas tienen una mayor relevancia a la hora de utilizar el conocimiento obtenido.

Estas cinco nuevas medidas biológicas propuestas por el grupo de expertos permiten, además de calcular el interés desde el punto de vista genético, realizar un ranking a partir del conjunto de reglas obtenido, lo que permite su posterior explotación para la toma de decisiones. Las cinco medidas se basan en el calculo de las coincidencias que se producen entre los items/genes del antecedente de las reglas y los de sus correspondientes consecuentes según las bases de datos de conocimiento y relaciones genéticas, como son las del proyecto de Oncología Genética (en inglés Genetic Onco-

logy; GO)[ABB+00] y las del proyecto de la Enciclopedia de Genes y Genomas de Kyoto (KEGG)[KGS+11]. A partir de la información sobre funciones genéticas, localización de proteinas e interacciones moleculares disponibles en estos dos proyectos, los expertos pueden calcular nuevas métricas a las reglas obtenidas, ya que cada coincidencia de las relaciones expresadas por el conjunto de reglas con la información que contienen las BDs de los proyectos anteriores, reforzará el hecho de que las reglas obtenidas representen realmente conocimiento verdadero.

En el caso de la BD del proyecto GO, ésta incluye información estructurada y utiliza un vocabulario específico para describir los resultados que se producen en el organismo humano a partir de las relaciones de parejas de genes. Ésta se estructura de acuerdo a tres categorías u ontologías diferentes: Componente celular, función molecular y proceso biológico. Estas tres categorías permiten definir tres de las cinco medidas de calidad biológica. En el caso de la BD del proyecto KEGG, encontramos recursos relacionados con la bioinformática, entre los que se integra el conocimiento disponible en la actualidad acerca de redes de interacción molecular, caminos celulares e información funcional de los genes, lo que permite definir la cuarta medida biológica. Ambas BDs han sido utilizadas de forma satisfactoria en estudios previos, lo que ha ayudado a explicar biológicamente las asociaciones de genes obtenidas [ARBAR10]. Por tanto, a partir de estas dos BDs de conocimiento biológico, se definen cuatro de las cinco medidas biológicas: Función Molecular (FM), Proceso Biológico (PB), Camino de Señalización (CS) y Comportamiento Celular (CC). Los valores posibles para estas medidas varían entre 6 y 1, ya que el calculo final de éstas las sitúa en estos valores para poder ordenar las reglas de forma sencilla en base a las mismas. De esta forma, aquellas reglas con valores cercanos a 1 serán mejores, mientras que aquellas con valores cercanos a 6 serán las menos relevantes. Estos valores no solamente permiten ordenar las reglas por interés, sino que los valores de las mismas indican la categoría de la regla, existiendo cinco categorías de calidad, tal y como se puede apreciar en la Figura 4.1

Por último, la quinta medida propuesta por los expertos es el Factor de Transcripción (TF), que viene dada a partir del conocimiento biológico recopilado en la BD TRRUST [HCL+17]. Tal y como el propio nombre de la medida indica, esta BD contiene información sobre los factores de transcripción de los genes que se producen en los procesos regulatorios biológicos. Los posibles valores para esta medida son: 0, 1, 2 y 3. Estos valores se obtienen en función de una serie de condiciones que puede o no cumplir la regla evaluada, cuantas más condiciones supere la regla, mayor será el valor y,

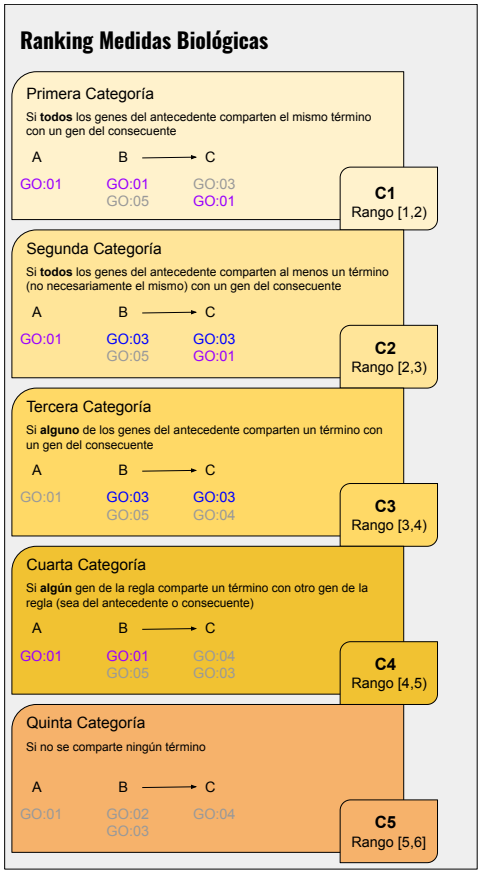

Figura 4.1: Categorías de interés para las medidas biológicas FM, PB, CS y CC.

por tanto, mayor será la relevancia de la misma. Las condiciones y valores asignados por cada regla para la medida TF son:

- **Condición 1**: Si al menos uno de los genes que forma parte del antecedente de la regla está reportado como un factor de transcripción en la BD TRRUST se proporciona un valor inicial de 1 para esta medida. En caso contrario, la regla tendrá un valor de 0.
- **Condición 2**: Si además de la condición 1, también se cumple que el gen del consecuente de la regla esta confirmado como un gen objetivo o diana de la regulación del gen o genes anteriormente reportados en la BD TRRUST, entonces se asigna un valor 2 a la regla.

Tabla 4.1: Parámetros utilizados para los experimentos de las BDs de dietas.

Algoritmo	*Parámetros*
HAUS-rules	*TamañoPoblación = 200, N_{eval}=150.000, $Prob_{mut}$= 0,1, α = 5*

- **Condición 3**: Si además de las condiciones 1 y 2, también se cumple que la regulación indicada por la regla (regulación positiva, regulación negativa o desconocida) coincide con la información acerca de esta regulación en la BD TRRUST, entonces se le asigna un valor de 3 a la regla.

Como podemos observar, estas condiciones tienen como finalidad determinar si las reglas representan correctamente el conocimiento ya probado en estudios anteriores, lo que al mismo tiempo significaría que el nuevo conocimiento descubierto a partir de las mismas es también correcto y debe tenerse en cuenta.

4.3. Experimentos

Tras introducir el problema biológico al que hemos aplicado nuestra nueva propuesta, incluyendo las modificaciones sobre los datos en bruto de los microarrays que se obtienen tras realizar las biopsias, en esta sección se introduce la experimientación realizada sobre la BD secuencial finalmente obtenida, incluyendo los resultados que se han obtenido y un breve análisis sobre los mismos. Para realizar las ejecuciones sobre cada una de las BDs secuenciales de cada una de las dietas, se han utilizado los parámetros por defecto, que coinciden con los parámetros utilizados en la experimentación realizada para la comparación con otras propuestas existentes en la literatura (ver sección 3.3.1). En la Tabla 4.1 podemos observar los parámetros utilizados para los experimentos.

A partir de la aplicación de nuestra propuesta a las BDs biológicas con las diferentes dietas con las que se han tratado los pacientes, se han obtenido 6 reglas secuenciales de alta utilidad media para la BD LCD (normal diet) y 9 para la VLCD (agressive diet). En las Tablas 4.2 y 4.3 se pueden observar las reglas extraídas tras la experimentación, incluyendo los resultados de las medidas de calidad biológica propuestas. Para ambas dietas, las reglas

Tabla 4.2: Selección de reglas secuenciales con interés biológico para la BD LCD.

Regla	*FM*	*PB*	*CS*	*CC*	*FT*
R_1 : (8062821/ $\downarrow$) $\rightarrow$ (7896214/ $\downarrow$)	6,00	6,00	1,20	6,00	0
R_2 : (7935776/ $\downarrow$) $\rightarrow$ (8019392/$FASN$ $\uparrow$)	6,00	6,00	1,20	6,00	0
R_3 : (8062821/ $\downarrow$) $\rightarrow$ (7894918/ $\downarrow$)	6,00	6,00	1,20	6,00	0
R_4 : (8105084/$C7$ $\uparrow$) $\rightarrow$ (7940762/$LGALS12$ $\uparrow$)	6,00	1,03	6,00	1,03	0
R_5 : (7935776/ $\downarrow$) $\rightarrow$ (8071036/$S100B$ $\uparrow$)	6,00	6,00	1,20	6,00	0
R_6 : (7935776/ $\downarrow$) y (7912692/$HSPB7$ $\downarrow$) $\rightarrow$ (8019392/$FASN$ $\uparrow$)	6,00	6,00	1,20	6,00	0
Media (Desviación Estándar)	6,00 (0,0)	5,17 (2,03)	2,00 (1,96)	5,17 (2,03)	0 (0,0)

que se han obtenido representan interacciones entre genes que participan en procesos moleculares que han sido previamente presentados en estudios conocidos como parte de una respuesta a la perdida de peso por parte del tejido adiposo [VRA+16].

De acuerdo a los valores medios de la medidas de calidad de las reglas, los mejores valores se obtienen en las reglas descubiertas a partir del grupo de individuos tratado por la dieta más agresiva, la VLCD. Que la dieta VLCD obtenga unos mejores valores que la dieta LCD se debe al mayor impacto molecular sobre el tejido adiposo provocado por la dieta. Aún así, todas las reglas obtenidas presentan buenos valores en al menos una de las medidas de calidad biológicas, excepto para la medida FT. El hecho de que para esta medida todas las reglas presenten un valor 0 podría explicarse por el hecho de que en los patrones de expresión de genes identificados no se producen factores de transcripción. Sin embargo, esto no significa que las reglas secuenciales descubiertas no sean interesantes desde el punto de vista biológico.

Una vez obtenido el conjunto de reglas para cada una de las dietas evaluadas, y con el objetivo de determinar la utilidad biológica de las reglas extraídas, una serie de expertos en el área de la investigación en obesidad han estudiado el conjunto de reglas secuenciales, contrastando con otros estudios de la literatura científica y analizando las reglas visualmente a través de dos representaciones gráficas que podemos observar en las Figuras 4.2 y 4.3, correspondientes a la dieta LCD y VLCD respectivamente. En estas figuras, podemos observar que cada nodo representa una *sonda/gen* (mostrando su identificador), las flechas que conectan los nodos representan las reglas del tipo $X \rightarrow Y$, mientras que el diámetro de cada uno de los vérti-

Tabla 4.3: Selección de reglas secuenciales con interés biológico para la BD VLCD.

Regla	*FM*	*PB*	*CS*	*CC*	*FT*
R_1 : (7896700/ ↑) → (7902623/$DNASE2B$ ↓)	6,00	6,00	6,00	1,32	0
R_2 : (8091723/$RARRES1$ ↑) → (8093053/$TFRC$ ↓)	1,62	1,62	1,20	1,62	0
R_3 : (8034304/$ACP5$ ↑) → (8093053/$TFRC$ ↓)	1,43	1,43	6,00	1,43	0
R_4 : (8046124/$DHRS9$ ↑) → (8046124/$DHRS9$ ↓)	1,00	1,00	1,20	1,00	0
R_5 : (8091723/$RARRES1$ ↑) → (7923562/$CHIT1$ ↓)	1,81	1,81	1,20	1,81	0
R_6 : (7935776/ ↓) *y* (7929816/SCD ↓) → (8165684/ ↑)	6,00	6,00	1,20	6,00	0
R_7 : (7936322/$GPAM$ ↓) *y* (8046124/$DHRS9$ ↑) → (8046124/$DHRS9$ ↓)	1,62	1,62	1,20	1,62	0
R_8 : (7935776/ ↓) *y* (7929816/SCD ↓) *y* (8033818/$OLFM2$ ↓) → (8165684/ ↑)	6,00	6,00	1,20	6,00	0
R_9 : (7936322/$GPAM$ ↓) *y* (8092177/$NCEH1$ ↑) *y* (8126784/$PLA2G7$ ↑) → (8000184/$IGSF6$ ↓)	1,93	1,93	1,20	1,93	0
Media (Desviación Estándar)	3,05 (2,23)	3,05 (2,23)	2,27 (2,12)	2,52 (1,99)	0 (0,0)

ces representa el valor de la medida de calidad FC de la regla, el color del vértice representa el tipo de cambio de expresión del gen (azul oscuro regulaciones positivas y azul claro regulaciones negativas), el color de la flecha representa el valor de la medida de calidad biológica PB (los mejores valores se representan con colores más oscuros), y finalmente, el ancho de la flecha representa el valor de la medida CS (mejores valores se representan con bordes más anchos).

Tras un análisis del conjunto de reglas, para la BD de la dieta LCD, debemos remarcar la regla R_4, para la cuál se obtiene el mejor valor de la medida de calidad biológica PB. Ésto indica que ambos genes participan en los mismos procesos biológicos. Si profundizamos en este patrón podemos observar como una regulación positiva del gen *C7*, que esta relacionada con una respuesta inmune inducida por lectinas, está seguida por una posterior regulación positiva de la expresión del gen *LGAL12*. Curiosamente, *LGAL12* (conocido como galectin-12) es una proteína humana lectina, expresada en el tejido adiposo y envuelta en la regulación de la degradación de lípidos (lipólisis) [YHL12]. Por sus funciones, *LGAL12* ha sido sugerido anteriormente como una diana muy útil para el tratamiento de la obesidad producida por causas de metabolismo, como la resistencia a la insulina, síndrome metabólico, y la diabetes de tipo 2 [YHL12]. Aunque aún es necesaria una validación de está interacción, esta regla revela un posible mecanismo por el cuál las lectinas provocan efectos negativos en la obesidad y el sistema inmunitario.

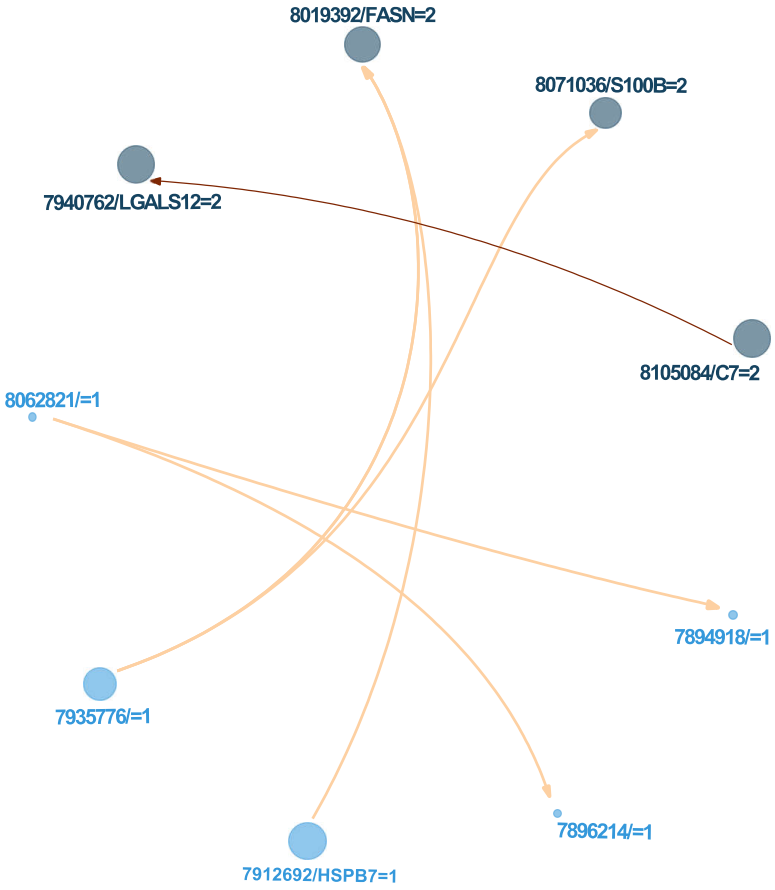

Figura 4.2: Representación jerárquica de la agrupación de aristas para las reglas secuenciales de la dieta LCD.

En cuanto al conjunto de reglas obtenido para la BD de la dieta VLCD, la regla R_4 destaca frente al resto por tener el mejor valor de la medida de calidad biológica PB. El patrón de interacción que representa esta regla nos permite observar como, aunque ciertos genes experimentan una regulación positiva como respuesta a la perdida de peso, sus niveles de expresión vuelven al valor de referencia tan pronto como se recupera una dieta de calorías normales durante la fase de estabilización del peso. Resulta curioso como este fenómeno puede explicar la existencia de una serie de rápidos cambios moleculares dinámicos y transitorios en el tejido humano como respuesta a las intervenciones. Otro patrón interesante que ha podido ser identificado en el grupo al que se le trata con la dieta VLCD es el identificado entre los

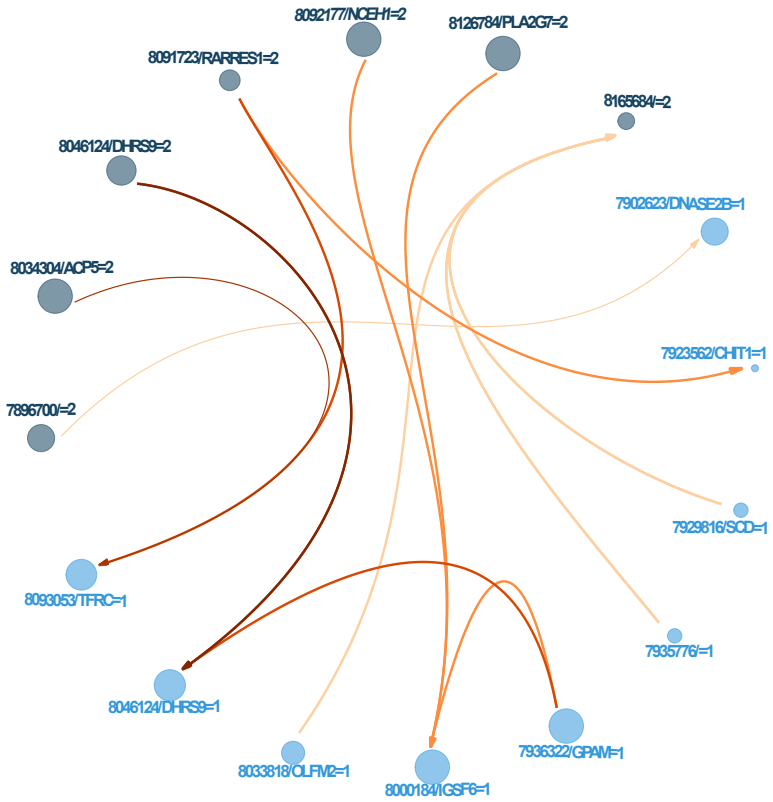

Figura 4.3: Representación jerárquica de la agrupación de aristas para las reglas secuenciales de la dieta VLCD.

genes *ACP5* y *TFRC* (regla R_3), cuya regulación orquestada y participación conjunta en diferentes redes moleculares se ha descrito en estudios previos sobre la obesidad [MGMA+11].

Considerando todo lo que se ha descubierto a través de la aplicación de la nueva propuesta a este problema, podemos afirmar que el algoritmo ha sido capaz de generar reglas comprensibles con un importante significado biológico y un gran interés para los expertos biólogos. Por tanto, la aplicación de nuestro método podría extenderse a otros problemas de intervenciones longitudinales a humanos en las que se cuente con información sobre la expresión de genes, además de otros muchos problemas del mundo real en los que contemos con BDs con información temporal secuencial.

4.4. Sumario

A lo largo de este capítulo se ha presentado la aplicación a un problema real del nuevo algoritmo de extracción de reglas secuenciales de alta utilidad media que se ha diseñado y desarrollado para cumplir con el objetivo principal de esta tesis. A lo largo de la sección 4.1 se ha introducido el problema biológico que se ha abordado con nuestro algoritmo. Tal y como se ha introducido, se trataba de un problema en el que se trataban 57 individuos con dos dietas diferentes. Una de ellas se trataba de una dieta más agresiva en la que los pacientes estaban limitados a ingerir una cantidad muy reducida de calorias (VLCD), mientras que en la otra los pacientes también se encontraban con una restricción en la cantidad de calorías a ingerir, aunque esta restricción era menos agresiva. Durante el tratamiento se realizan tres biopsias a través de las cuales se obtienen los datos genéticos de los pacientes en forma de microarrays, que tras una fase de preprocesado inicial se transforman en las BDs secuenciales finales que se han utilizado para la ejecución de la propuesta.

Tras realizar y describir el preprocesamiento y transformación de los datos, se han definido una serie de medidas de calidad biológicas que permitan a los expertos biólogos determinar el interés y la calidad de las reglas obtenidas. En la sección 4.2, se introducen estas medidas biológicas, incluyendo una breve explicación de su significado, por qué son interesantes para los expertos, como se calculan y el dominio de valores que pueden tomar.

Finalmente, en la sección 4.3, se han presentado los experimentos que se han llevado a cabo, poniendo especial interés en mostrar y analizar los resultados finales que se han obtenido. Tal y como se ha podido observar, se ha obtenido un conjunto de reglas altamente comprensibles para los expertos y con un conocimiento útil, que puede verificarse a través de estudios científicos previos, lo que permite afirmar que nuestra nueva propuesta cumple con su cometido, siendo capaz de extraer conocimiento útil y fácilmente comprensible al mismo tiempo. Ésto junto a las comparativas con otros algoritmos disponibles en la literatura (véase sección 3.3) permite verificar la eficacia de nuestro algoritmo.

Comentarios Finales

A Resumen y Conclusiones

En esta memoria hemos propuesto una nueva taxonomía para las RATs mediante un estudio en profundidad del estado del arte del área. Una vez identificados los distintos enfoques existentes en la literatura y algunas de las líneas de investigación más actuales, hemos propuesto un nuevo AEMO que permite extraer RATs de alta utilidad media a partir de BDs secuenciales. El comportamiento de esta propuesta ha sido contrastado sobre varias BDs de prueba con el objetivo de comparar su eficacia frente a algunas de las propuestas existentes en la literatura especializada. Además, el método propuesto para extraer RATs de alta utilidad ha sido aplicado para realizar un análisis temporal sobre un estudio longitudinal in vivo de la expresión génica en tejido adiposo humano. Los siguientes apartados resumen brevemente los resultados obtenidos y presentan algunas conclusiones sobre ellos.

A.1 Taxonomía - Estudio del estado del arte RATs

Para cumplir con una de las lineas del propósito general de esta tesis, que consiste en el estudio, diseño y desarrollo de nuevas técnicas de extracción de RATs, hemos propuesto una nueva clasificación del estado del arte de las RATs para proporcionar un marco de trabajo bien definido a los investigadores de ámbito.

Uno de los principales problemas de este campo de investigación es la falta de visibilidad de muchos de sus trabajos debido a que no existe una terminología estándar para referirse a él, lo que dificulta la búsqueda y comparación de propuestas y estudios en este campo. Ésto ha provocado que

sea complicado el desarrollo de propuestas novedosas que extiendan modelos recientes o proporcionen soluciones a problemas abiertos en la actualidad, y que incluyan un análisis experimental y estadístico con otras propuestas sobre distintas medidas de calidad para resaltar la importancia de las novedades presentadas en las propuestas. Además, es fundamental que todas las aportaciones sigan un enfoque open/free source debido al impacto tan positivo que tiene tanto en el desarrollo de mejores algoritmos como en su aplicabilidad a áreas muy diversas y en el ámbito empresarial. De hecho, ya hay algunas editoriales en las que es obligatorio proporcionar como open/free source los códigos que hayan sido desarrollados en las propuestas publicadas en sus revistas de investigación. Todo esto ralentiza el crecimiento de esta área de investigación, que aun así, ha tenido una gran cantidad de publicaciones en los últimos años.

Por ello, hemos propuesto una taxonomía a través de la cual se proporcione un marco de trabajo bien definido para que los investigadores puedan encontrar más fácilmente propuestas existentes y desarrollar nuevas bien fundamentadas. Esta taxonomía proporciona una instantánea clara del estado del arte, de los diferentes tipos de propuestas y sus diferencias. Esto no solo ayuda a desarrollar nuevas propuestas a investigadores de este campo, sino también a que investigadores y desarrolladores de otras áreas puedan ver en la aplicación de estas técnicas una solución a sus problemas.

En esta taxonomía nos hemos centrado en la forma de tratar la componente temporal por parte de las diferentes técnicas existentes, encontrando un primer nivel de la taxonomía que se divide en dos tipos de propuestas: las que consideran el tiempo como una componente implícita y las que consideran el tiempo como una componente explícita o integral. En el caso de las propuestas del primer tipo, las que consideran el tiempo como una componente implícita, la variable temporal se considera como una relación de orden entre los datos que forman parte del conjunto y/o como una serie de restricciones temporales entre estos datos. Mientras que esta primera categoría se subdivide a su vez en otras dos subcategorías en función del tipo de BD temporal: secuencial e inter-transaccional. Por otro lado, las propuestas que consideran el tiempo como una componente explícita o integral introducen la componente temporal como una variable o atributo mas dentro del proceso de aprendizaje. A su vez, esta categoría se subdivide en otras cinco subcategorías en función de como se incluye la información temporal en el proceso de aprendizaje: periódicas, de intervalos temporales, de tiempo de vida, de cambios en reglas e incrementales.

Tras analizar los trabajos de la literatura podemos observar que este tipo de técnicas han sido aplicadas satisfactoriamente en una gran cantidad áreas. Si nos fijamos en detalle, podemos encontrar aplicaciones en la industria, seguridad, medicina y atención sanitaria, entre muchas otras áreas, destacando su uso para resolver problemas relacionados con las áreas de la medicina y la salud. Esto demuestra el gran potencial que tienen las RATs y que su uso ha ido incrementándose en los últimos años, proporcionando una mayor evidencia de la necesidad y del interés de estas técnicas de extracción de RATs para la resolución de problemas del mundo real.

A pesar de ello, existen muchas lineas de trabajo futuro dentro del área, tales como el desarrollo de nuevas propuestas que permitan manejar eficientemente datos masivos generados en distintos ámbitos (por ejemplo, IoT) teniendo en cuenta distintos tipos de problemas derivados de los datos (por ejemplo, seguridad, privacidad, etc), o combinar la componente temporal con otras dimensiones (por ejemplo, espacio) para desarrollar modelos multidimensionales, o hacer uso de nuevas medidas que permitan dirigir la búsqueda de los métodos hacia reglas más interesantes y útiles para los usuarios.

A.2 Diseño y desarrollo del nuevo algoritmo de extracción de RSAUMs

En muchas áreas de aplicación, los datos recogidos quedan estructurados como secuencias de valores/eventos que se ordenan linealmente según el momento en que se produjeron. Esta información temporal es esencial para poder extraer conocimiento útil de los datos y es la razón por la que las reglas secuenciales se han utilizado con éxito para resolver satisfactoriamente una gran variedad de problemas de distintas áreas.

La mayoría de los algoritmos suelen extraer las reglas secuenciales sin prestar especial atención a la utilidad que tiene cada elemento de la BD para el usuario, sin embargo, su utilidad dependerá del problema concreto con el que estemos tratando. Por ejemplo, si aplicamos estos métodos sobre una base de datos de expresión génica, se consideraría que todos los genes tienen la misma importancia dentro del proceso biológico, sin embargo, los estudios clínicos han demostrado que algunos genes tienden a ser más significativos a la hora de explicar la causa de una enfermedad o son más eficaces para combatirla.

Nosotros hemos propuesto HAUS-rules, un AEMO que nos permite realizar un aprendizaje evolutivo de reglas secuenciales interesantes, fáciles de

comprender y de alta utilidad a partir de bases de datos secuenciales haciendo uso de la capacidad de búsqueda y optimización de estos algoritmos. El proceso de extracción es por lo tanto considerado como un problema multiobjetivo en el que varias medidas son optimizadas conjuntamente, permitiendo no sólo obtener reglas interesantes según diversas medidas de interés disponibles en la literatura (como soporte, confianza, lift, factor de certeza, etc) sino también por la utilidad que tienen los distintos elementos de la BD para el usuario. Además, un aspecto fundamental es que las reglas generadas sean comprensibles por el usuario, por lo que la búsqueda desarrollada e implementada en HAUS-rules es guiada para generar reglas que incluyan pocos elementos y que sean de un gran interés y utilidad para el usuario.

Del estudio realizado podemos concluir que el método propuesto consigue generar conjuntos reducidos de reglas interesantes, fáciles de comprender y de alta utilidad, mejorando estadísticamente los resultados obtenidos por otras propuestas de la literatura para diversas medidas de interés y de utilidad. Ésto pone de manifiesto el gran potencial de HAUS-rules para explorar el espacio de búsqueda optimizando varios objetivos de forma conjunta. Además, las medidas de utilidad media (considerando la cantidad de elementos incluidos en la regla para calcular su utilidad) permiten reducir el espacio de búsqueda, permitiendo dirigir la búsqueda hacia el conjunto de reglas que realmente deseamos extraer.

El método propuesto y desarrollado en esta tesis doctoral abre las puertas a la aplicación de este tipo de enfoques para la resolución de una gran cantidad de problemas reales en distintos ámbitos de aplicación, así como al desarrollo de nuevos enfoques que permitan resolver otros problemas abiertos en el área.

A.3 Aplicación de la propuesta a un problema real: Estudio longitudinal in vivo de la expresión genética en tejido adiposo humano

Por último, en este trabajo de investigación hemos aplicado todo el conocimiento adquirido durante el estudio del estado del arte y el posterior desarrollo de nuestro algoritmo a un problema bio-sanitario real, en el que se realiza un estudio longitudinal in vivo de la expresión genética en tejido adiposo humano. Para ello, hemos trabajado con dos BDs de expresión genética relativas a dos dietas de perdida de peso diferentes, una de ellas más agresiva (VLCD), con una cantidad de calorías a ingerir por los pacientes muy

baja, mientras que la otra se trataba de una dieta menos agresiva (LCD). La información de expresión genética almacenada en las BDs se corresponde con diferentes instantes temporales en los que se realizan las biopsias a los pacientes para obtener los microarrays genéticos. El objetivo de este problema era tratar de encontrar qué efectos a nivel de expresión génica producen ambas dietas y a qué cambios de expresión se podrían deber, qué diferencias reales existían entre ambos tratamientos, y si alguno de ellos podría tener algunos beneficios frente al otro.

A partir de los datos en bruto, de los microarrays genéticos con información de miles de sondas y genes, se ha aplicado una fase inicial de pre-procesamiento a partir de la cual se han obtenido las BDs secuenciales que pueden utilizarse como entrada para nuestra propuesta. En esta etapa de pre-procesamiento no sólo se ha transformado la BD, sino que también se han reducido la cantidad de genes a tener en cuenta, ya que la cantidad inicial era demasiado grande, y la mayoría de ellos no tenía una importancia real en los procesos biológicos, puesto que no veían afectada su expresión.

Tras aplicar nuestra propuesta a este problema, hemos obtenido un conjunto de reglas reducido y con un buen tamaño para que estas puedan ser fácilmente interpretadas por los expertos. La obtención de un conjunto de reglas reducido y que las mismas presenten un tamaño adecuado son dos características importantes cuando se aplican técnicas de extracción de RAs a un problema real, ya que se desea comprender el conocimiento que representan y que este conocimiento pueda ser manejable. Es decir, es altamente recomendable que el número de reglas no sea excesivo, evitando las reglas redundantes o que no aporten conocimiento realmente interesante.

Junto a investigadores de primer nivel se han desarrollado varias medidas de calidad que permiten constatar el interés desde el punto biológico de las reglas obtenidas. Estas medidas se han desarrollado a partir de las anotaciones disponibles en diferentes repositorios de BDs biológicas de renombre (GO y KEGG). Gracias a todo ello, ha sido posible observar que las reglas obtenidas son realmente interesantes y útiles, además de que representan fenómenos biológicos reales, que se habían observado previamente en estudios ya publicados. Todo esto demuestra la eficacia de nuestra propuesta al aplicarla a un problema real, en el que ha dado unos buenos resultados, obteniendo conocimiento útil, interesante e interpretable.

B Publicaciones Asociadas a la Tesis

A continuación se presenta la lista completa de publicaciones en diferentes revistas y congresos que se han realizado como consecuencia directa del trabajo desarrollado en esta tesis. En concreto, se han publicado 2 artículos en revistas de investigación internacionales JCR Q1 (uno publicado y otro sometido), 1 artículo en un congreso internacional y 1 artículo de divulgación. Por otro lado, también se ha colaborado con otros investigadores de diferentes ámbitos en temas relacionados con la tesis, dando lugar a la publicación de 1 artículo en una revista de investigación internacional JCR Q1 y 1 artículo en un congreso internacional.

- Publicaciones en revistas internacionales:
  - Segura-Delgado, A., Gacto, M. J., Alcalá, R., & Alcalá-Fdez, J. (2020). Temporal association rule mining: An overview considering the time variable as an integral or implied component. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(4), e1367. (JCR Q1) [SDGAAF20]
  - Segura-Delgado, A., Anguita-Ruiz A., Alcalá, R., & Alcalá-Fdez, J. (2021). Mining High Average-Utility Sequential Rules to Identify High-Utility Gene Expression Sequences in Longitudinal Human Studies. Expert Systems With Applications. (JCR Q1) *Sometido (en segunda vuelta de revisión).*
- Publicaciones en congresos internacionales:
  - Segura A., Pérez-Pérez R. y Alcalá-Fdez J. (2016) New open source modules in KEEL to analyze and export fuzzy association rules. En IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2016), páginas 225-232. Vancouver, Canada [SPPAF16]
- Publicaciones de divulgación:
  - Alberto Segura-Delgado, María José Gacto, Rafael Alcalá, and Jesús Alcalá-Fdez. Solving real-life problems with temporal association rules. Advanced Science News - Computer Science 1-5. Disponible en [1] desde el 20 de Mayo de 2020.

[1] `https://www.advancedsciencenews.com/solving-real-life-problems-with-temporal-association-rules/`

- Colaboraciones en temáticas relacionadas con la tesis:
  - Anguita-Ruiz A., Segura-Delgado A., Alcalá R., Aguilera C. M. y Alcalá-Fdez J. (2020) explainable artificial intelligence (xai) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. PLOS Computational Biology 16(4): 1-34. (JCR Q1) [ARSDA$^+$20]
  - Anguita-Ruiz A., Segura-Delgado A., Alcala R., Aguilera C. M. y Alcala-Fernandez J. (2019) Describing sequential association patterns from longitudinal microarray data sets in humans. En Rojas I., Valenzuela O., Rojas F. y Ortuño F. Bioinformatics and Biomedical Engineering, páginas 318-329. Springer International Publishing. [ARSDA$^+$19]

C Líneas de investigación Futuras

A continuación, se presentan algunas de las lineas de investigación futuras que, como conclusión de nuestra investigación, consideramos que son interesantes para tener en cuenta en el futuro.

C.1 Extensión del método propuesto con nuevos modelos de representación de reglas

Existen varios enfoques para extender nuestra propuesta para extraer RSAUMs. Sin embargo, consideramos que una de las más importantes puede ser la posibilidad de introducir nuevos modelos de representación de reglas, lo que aumentaría las posibilidades de representación de conocimiento, permitiendo extraer conocimiento interesante y útil que no es posible representar con la estructura clásica regla y presentar el conocimiento actualmente extraído de una forma más condensada.

Entre los modelos de representación de reglas más interesantes que podrían introducirse, destacamos la posibilidad de introducir la representación de reglas negadas, lo que permitiría no sólo representar el conocimiento que se obtendría con las reglas habituales, sino también detectar y representar aquellas relaciones en las que la no ocurrencia de un item puede suponer la ocurrencia de otro. Este tipo de relaciones se producen en el mundo real

y poder representarlas en las RSAUMs sería una forma muy interesante de ampliar las posibilidades de nuestra propuesta, que podría obtener conocimiento más útil e interesante.

Otra opción es considerar un proceso de extracción de reglas difusas. Este modelo de regla ha sido ampliamente utilizado para extraer conocimiento interesante y comprensible para un humano a partir de grandes BDs, haciendo uso de las ventajas proporcionadas por la teoría de conjuntos difusos, que ha sido frecuentemente utilizada en sistemas inteligentes debido a su simplicidad y cercanía al razonamiento humano. Además, la lógica difusa permite evitar los límites no naturales mediante la partición de los dominios de atributos y ampliar el tipo de asociaciones que pueden representarse.

Definitivamente, las mejoras relacionadas con la representación de las reglas secuenciales son realmente prometedoras de cara a mejorar nuestra propuesta, con el objetivo de aumentar aún más la capacidad de representar el conocimiento y, por tanto, la capacidad de obtener un conocimiento interesante y útil para el usuario.

C.2 Aplicación del método propuesto a otros problemas reales para extraer relaciones a partir de datos de distintas ómicas

Tal y como hemos observado en la sección 2.5, buena parte de las aplicaciones reales de las propuestas para extracción de RATs están relacionadas con problemas biológicos en las áreas de medicina y/o salud. Durante nuestra investigación hemos aplicado nuestra propuesta a un problema biológico en el que se aplican dos tipos de dietas a los pacientes para obtener información sobre los efectos de cada uno de estos tratamientos sobre los pacientes. Se ha demostrado que este tipo de propuestas son realmente interesantes para generar información útil que permite a los científicos comprender mejor los procesos biológicos asociados a los tratamientos y generar nuevas hipótesis para abordar el problema. Los buenos resultados obtenidos nos hacen pensar que nuestra propuesta puede ser de utilidad si es adaptada para resolver otros problemas similares.

Por ello, una de las líneas de trabajo futuro que nos planteamos es su adaptación a otros problemas ómicos relacionados con salud, sobretodo teniendo en cuenta la situación tan difícil que nos ha tocado vivir a todos en estos últimos años debido a la pandemia. Creemos que sería realmente interesante estudiar, por ejemplo, las reacciones de diferentes tratamientos y/o vacunas frente al COVID-19 con el paso del tiempo y las reacciones a nivel

genético mediante la aplicación de algoritmos de extracción de RSAUMs, con las que los expertos puedan comprender mejor como funcionan este tipo de enfermedades y los tratamientos a las mismas.

C.3 Adaptación de los modelos para generar nuevos métodos de clasificación

Una de las líneas de trabajo futuro planificadas es la extensión de HAUS-rules para su aplicación en problemas de clasificación. Ésto permitiría utilizar RSAUMs en nuevos problemas y retos, en los que sea necesario generar un clasificador cuyas reglas sean comprensibles y que incluyan elementos que sean importantes para el experto.

De este modo, podría aplicarse un método de clasificación que utilice las RSAUMs obtenidas a un problema en el que se trate de clasificar software malicioso frente a software no malicioso. Modelando la ejecución del software como una sucesión de eventos, relacionados con las llamadas a ciertas funciones del sistema, es posible obtener relaciones secuenciales entre las llamadas a ciertas funciones para tratar de determinar de forma temprana si un software tendrá en las siguientes llamadas un comportamiento malicioso.

Como podemos observar, estos nuevos métodos de clasificación, si se diseñan correctamente, podrían utilizarse no sólo para resolver problemas de clasificación clásicos, sino también para problemas de clasificación en tiempo real, en los que se proporciona como entrada un flujo de datos y el sistema debe proporcionar como salida una categoría lo antes posible. Además, con el conocimiento de los expertos que se introduce a través de la medida de utilidad, estos sistemas podrían obtener mejores resultados que los métodos de clasificación típicos.

Bibliografía

[ABB+00] Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nature genetics* 25(1): 25–29.

[ABT+17] Arnaud M., Begaud B., Thurin N., Moore N., Pariente A. y Salvo F. (2017) Methods for safety signal detection in healthcare databases: a literature review. *Expert Opinion on Drug Safety* 16(6): 721–732.

[AC02] Au W.-H. y Chan K. C. (2002) Fuzzy data mining for discovering changes in association rules over time. En *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2002)*, páginas 890–895. Honolulu, HI, USA.

[AIS93] Agrawal R., Imielinski T. y Swami A. (1993) Mining association rules between sets of items in large databases. En *SIGMOD*, páginas 207–216. Washington D.C., USA.

[All83] Allen J. F. (1983) Maintaining knowledge about temporal intervals. *Communications of the ACM* 26(11): 832–843.

[ALW+18] Ao X., Luo P., Wang J., Zhuang F. y He Q. (2018) Mining precise-positioning episode rules from event sequences. *IEEE Transactions on Knowledge and Data Engineering* 30(3): 530–543.

[AR00] Ale J. y Rossi G. (2000) An approach to discovering temporal association rules. En *ACM Symposium on Applied Computing (SAC 2000)*, volumen 1, páginas 294–300. Como, Italy.

[ARBAR10] Alves R., Rodriguez-Baena D. S. y Aguilar-Ruiz J. S. (2010) Gene association analysis: a survey of frequent pattern mining from gene expression data. *Briefings in Bioinformatics* 11(2): 210–224.

[ARKJ17] Aljawarneh S. A., Radhakrishna V., Kumar P. V. y Janaki V. (2017) G-SPAMINE: an approach to discover temporal association patterns and trends in internet of things. *Future Generation Computer Systems* 74: 430–443.

[ARSDA+19] Anguita-Ruiz A., Segura-Delgado A., Alcala R., Aguilera C. M. y Alcala-Fernandez J. (2019) Describing sequential association patterns from longitudinal microarray data sets in humans. En Rojas I., Valenzuela O., Rojas F. y Ortuño F. (Eds.) *Bioinformatics and Biomedical Engineering*, páginas 318–329. Springer International Publishing, Cham.

[ARSDA+20] Anguita-Ruiz A., Segura-Delgado A., Alcalá R., Aguilera C. M. y Alcalá-Fdez J. (2020) explainable artificial intelligence (xai) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. *PLOS Computational Biology* 16(4): 1–34.

[AS94] Agrawal R. y Srikant R. (1994) Fast algorithms for mining association rules. En *International Conference Large Data Bases*, páginas 487–499. Santiago de Chile, Chile.

[AS95] Agrawal R. y Srikant R. (1995) Mining sequential patterns. En *11th International Conference on Data Engineering (ICDE 1995)*, páginas 3–14. Washington, DC, USA.

[BBJ98] Bohlen M., Busatto R. y Jensen C. (1998) Point-versus interval-based temporal data models. En *Proceedings 14th International Conference on Data Engineering*, páginas 192–200. Orlando, USA.

[BBSV02] Berzal F., Blanco I., Sánchez D. y Vila M. A. (2002) Measuring the accuracy and interest of association rules: A new framework. *Intelligent Data Analysis* 6(3): 221–235.

[BCD+09] Berthold M. R., Cebron N., Dill F., Gabriel T. R., Kötter T., Meinl T., Ohl P., Thiel K. y Wiswedel B. (2009) KNIME–the konstanz information miner: Version 2.0 and beyond. *ACM SIGKDD Explorations Newsletter* 11(1): 26–31.

[BDRD+20] Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R. y Herrera F. (2020) Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58: 82 – 115.

[Boe11] Boettcher M. (2011) Contrast and change mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1(3): 215–230.

[BRBHEHA19] Bou Rjeily C., Badr G., Hajjarm El Hassani A. y Andres E. (2019) *Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field*, páginas 71–99. Springer International Publishing, Cham.

[BVG+19] Buxton E., Vohra S., Guo Y., Fogleman A. y Patel R. (2019) Pediatric population health analysis of southern and central illinois region: A cross sectional retrospective study using association rule mining and multiple logistic regression. *Computer Methods and Programs in Biomedicine* 178: 145–153.

[BWJ98] Bettini C., Wang X. S. y Jajodia S. (1998) Mining temporal relationships with multiple granularities in time sequences. *IEEE Data Engineering Bulletin* 21: 32–38.

[BWL+12] Barrett T., Wilhite S. E., Ledoux P., Evangelista C., Kim I. F., Tomashevsky M., Marshall K. A., Phillippy K. H., Sherman P. M., Holko M., Yefanov A., Lee H., Zhang N., Robertson C. L., Serova N., Davis S. y Soboleva A. (11 2012) NCBI GEO: archive for functional genomics data setsupdate. *Nucleic Acids Research* 41(D1): D991–D995.

[Car14] Cariñena P. (2014) Fuzzy temporal association rules: combining temporal and quantitative data to increase rule expressiveness. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 4(1): 64–70.

[CLHL16] Chen C.-H., Lan G.-C., Hong T.-P. y Lin S.-B. (2016) Mining fuzzy temporal association rules by item lifespans. *Applied Soft Computing* 41: 265–274.

[CM19] Chamazi M. y Motameni H. (2019) Finding suitable membership functions for fuzzy temporal mining problems using fuzzy temporal bees method. *Soft Computing* 23(10): 3501–3518.

[CP99] Chen X. y Petrounias I. (1999) Mining temporal features in association rules. En *European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 1999)*, páginas 295–300. Prague, Czech Republic.

[CPH98] Chen X., Petrounias I. y Heathfield H. (1998) Discovering temporal association rules in temporal databases. En *International workshop on issues and applications of database technology (IADT 1998)*, páginas 312–319. Berlin, Germany.

[CSC+11] Concaro S., Sacchi L., Cerra C., Fratino P. y Bellazzi R. (2011) Mining health care administrative data with temporal association rules on hybrid events. *Methods of Information in Medicine* 50(2): 166–179.

[CVL07] Coello C., Veldhuizen D. y Lamont G. (2007) *Evolutionary Algorithms for solving multi-objective problems.* Kluwer Academic Publishers, 2 edition.

[DAG19] Das A., Ahmed M. y Ghasemzadeh A. (2019) Using trajectory-level SHRP2 naturalistic driving data for investigating driver lane-keeping ability in fog: An association rules mining approach. *Accident Analysis and Prevention* 129: 250–262.

[DAPM02] Deb K., Agrawal S., Pratab A. y Meyarivan T. (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions Evolutionary Computation* 6(2): 182–197.

[DDG+20] Diop L., Diop C. T., Giacometti A., Li D. y Soulet A. (2020) Sequential pattern sampling with norm-based utility. *Knowledge and Information Systems* 62(5): 2029–2065.

[DFVN17] Dalmas B., Fournier-Viger P. y Norre S. (2017) Twincle : A constrained sequential rule mining algorithm for event logs. *Procedia Computer Science* 112: 205–214.

[DJ05] Deogun J. y Jiang L. (2005) Prediction mining – an approach to mining association rules for prediction. En Slezak D., Yao J., Peters J. F., Ziarko W. y Hu X. (Eds.) *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, páginas 98–108. Springer Berlin Heidelberg, Berlin, Heidelberg.

[DLM+98] Das G., Lin K.-L., Mannila H., Renganathan G. y Smyth P. (1998) Rule discovery from time series. En *International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 1998)*, páginas 16–22. ACM Pres, New York, USA.

[Do18] Degirmenci T. y ozbakir L. (2018) Differentiating households to analyze consumption patterns: A data mining study on official household budget data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8(1): 1–15.

[ES03] Eiben A. y Smith J. (2003) *Introduction to Evolutionary Computing*. Springer-Verlag, Berlin, Germany.

[FDL01] Feng L., Dillon T. y Liu J. (2001) Inter-transactional association rules for multidimensional contexts for prediction and their applications to studying meteological data. *Data and Knowledge Engineering* 37: 85–115.

[FDPD19] Funde N. A., Dhabu M. M., Paramasivam A. y Deshpande P. S. (2019) Motif-based association rule mining and clustering technique for determining energy usage patterns for smart meter data. *Sustainable Cities and Society* 46: 101415.

[FHC+19] Fernandez A., Herrera F., Cordon O., Jose del Jesus M. y Marcelloni F. (2019) Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? *IEEE Computational Intelligence Magazine* 14(1): 69–81.

[FHHP12] Febrer-Hernández J. K. y Hernández-Palancar J. (2012) Sequential pattern mining algorithms review. *Intelligent Data Analysis* 16(3): 451–466.

[FM17] Fouad M. y Mostafa M. (2017) IndxTAR: An efficient algorithm for indexed mining of incremental temporal association rules. *International Journal of Computer Information Systems and Industrial Management Applications* 9: 103–113.

[Fri37] Friedman M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Statist. Assoc.* 32(200): 675–701.

[FVFNN12] Fournier-Viger P., Faghihi U., Nkambou R. y Nguifo E. M. (2012) Cmrules: Mining sequential rules common to several sequences. *Knowledge-Based Systems* 25(1): 63–76.

[FVGG+14] Fournier-Viger P., Gomariz A., Gueniche T., Soltani A., Wu C.-W. y Tseng V. S. (2014) SPMF: A java open-source pattern mining library. *Journal of Machine Learning Research* 15: 3569–3573.

[FVGZT14] Fournier-Viger P., Gueniche T., Zida S. y Tseng V. S. (2014) ERMiner: Sequential rule mining using equivalence classes. En Blockeel H., van Leeuwen M. y Vinciotti V. (Eds.) *Advances in Intelligent Data Analysis XIII*, páginas 108–119. Springer International Publishing, Cham.

[FVLK+17] Fournier-Viger P., Lin J.-W., Kiran R., Koh Y. y Thomas R. (2017) A survey of sequential pattern mining. *Data Science and Pattern Recognition* 1(1): 54–77.

[FVLN+19] Fournier-Viger P., Lin J. C.-W., Nkambou R., Vo B. y Tseng V. S. (2019) *High-Utility Pattern Mining: Theory, Algorithms and Applications*. Springer, 1st edition.

[FVLV+17] Fournier-Viger P., Lin J.-W., Vo B., Chi T., Zhang J. y Le H. (2017) A survey of itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7(4): 1–18.

[FVWTN12] Fournier-Viger P., Wu C.-W., Tseng V. S. y Nkambou R. (2012) Mining sequential rules common to several sequences with the window size constraint. En Kosseim L. y Inkpen D. (Eds.) *Advances in Artificial Intelligence*, páginas 299–304. Springer Berlin Heidelberg, Berlin, Heidelberg.

[FVYL+19] Fournier-Viger P., Yang P., Lin J., Duong Q., Dam T.-L., Frnda J., Sevcik L. y Voznak M. (2019) Discovering periodic itemsets using novel periodicity measures. *Advances in Electrical and Electronic Engineering* 17(1): 33–44.

[FWT+15] Fournier-Viger P., Wu C., Tseng V. S., Cao L. y Nkambou R. (2015) Mining partially-ordered sequential rules common to multiple sequences. *IEEE Transactions on Knowledge and Data Engineering* 27(8): 2203–2216.

[GBDW+17] Guillame-Bert M., Dubrawski A., Wang D., Hravnak M., Clermont G. y Pinsky M. (2017) Learning temporal rules to forecast instability in continuously monitored patients. *Journal of the American Medical Informatics Association* 24(1): 47–53.

[GCC+16] Gallo C. A., Cecchini R. L., Carballido J. A., Micheletto S. y Ponzoni I. (2016) Discretization of gene expression data revised. *Briefings in Bioinformatics* 17(5): 758–770.

[GGB20] Ghosh S., Ghosh S. K. y Buyya R. (2020) Mario: A spatio-temporal data mining framework on google cloud to explore mobility dynamics from taxi trajectories. *Journal of Network and Computer Applications* 164: 102692.

[GH06] Geng L. y Hamilton H. J. (2006) Interestingness measures for data mining: A survey. *ACM Computing Surveys* 38(3): 1–32.

[GKM19] Giannoulis M., Kondylakis H. y Marakakis E. (2019) Designing and implementing a collaborative health knowledge system. *Expert Systems with Applications* 126: 277–294.

[GMLH09] Garcia S., Molina D., Lozano M. y Herrera F. (2009) A study on the use of non-parametric tests for analyzing the evolutionary algorithmsbehaviour: A case study on the

ceć2005 special session on real parameter optimization. *Journal Heuristics* 15: 617–644.

[GNTA10] Gharib T. F., Nassar H., Taha M. y Abraham A. (2010) An efficient algorithm for incremental mining of temporal association rules. *Data and Knowledge Engineering* 69(8): 800–815.

[Gol89] Goldberg D. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley/ Longman, USA.

[HCL+17] Han H., Cho J.-W., Lee S., Yun A., Kim H., Bae D., Yang S., Kim C. Y., Lee M., Kim E., Lee S., Kang B., Jeong D., Kim Y., Jeon H.-N., Jung H., Nam S., Chung M., Kim J.-H. y Lee I. (2017) TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research* 46(D1): D380–D386.

[HD04] Harms S. y Deogun J. (2004) Sequential association rule mining with time lags. *Journal of Intelligent Information Systems* 22(1): 7–22.

[HDC07] Huang J.-W., Dai B.-R. y Chen M.-S. (2007) Twain: Two-end association miner with precise frequent exhibition periods. *ACM Transactions on Knowledge Discovery from Data* 1(2): 1–33.

[HDY99] Han J., Dong G. y Yin Y. (1999) Efficient mining of partial periodic patterns in time series database. En *International Conference on Data Engineering (ICDE 1999)*, páginas 106–115. Sydney, Australia.

[HGY98] Han J., Gong W. y Yin Y. (1998) Mining segment-wise periodic patterns in time-related databases. En *International Conference on Knowledge Discovery and Data Mining (KDDM 1998)*, páginas 214–218. New York, USA.

[HHC16] Huang T.-K., Huang C.-H. y Chuang Y.-T. (2016) Change discovery of learning performance in dynamic educational environments. *Telematics and Informatics* 33(3): 773–792.

[HK01] Höppner F. y Klawonn F. (2001) Finding informative rules in interval sequences. En *Advances in Intelligent Data Analysis. Lecture notes in Computer Science*, volumen 2189, páginas 125–134. Cascais, Portugal.

[HK05] Huang Y.-P. y Kao L.-J. (2005) A novel approach to mining inter-transaction fuzzy association rules from stock price variation data. En *14th IEEE International Conference on Fuzzy Systems*, páginas 791–796. Reno, USA.

[HK06] Han J. y Kamber M. (2006) *Data Mining: Concepts and Techniques*. Morgan Kaufmann, USA, 2 edition.

[HKS07] Huang Y.-P., Kao L.-J. y Sandnes F. (2007) Data mining and fuzzy inference based salinity and temperature variation prediction. En *IEEE International Conference on Systems, Man and Cybernetics*, páginas 2074–2079. Montreal, Canada.

[HLS+16] Hong T.-P., Lan G.-C., Su J.-H., Wu P.-S. y Wang S.-L. (2016) Discovery of temporal association rules with hierarchical granular framework. *Applied Computing and Informatics* 12(2): 134–141.

[HLW11] Hong T.-P., Lee C.-H. y Wang S.-L. (2011) Effective utility mining with the measure of average utility. *Expert Systems with Applications* 38(7): 8259 – 8265.

[Höp01] Höppner F. (2001) Discovery of temporal patterns. En *European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2001). Lecture notes in Computer Science*, volumen 2168, páginas 192–203. Springer Berlin Heidelberg, Freiburg, Germany.

[HPY00] Han J., Pei J. y Yin Y. (2000) Mining frequent patterns without candidate generation. En *ACM Sigmod Record*, volumen 29, páginas 1–12. ACM.

[HT19] Heydari J. y Tajer A. (2019) Quickest search and learning over correlated sequences: Theory and application. *IEEE Transactions on Signal Processing* 67(3): 638–651.

[HYWC16] Hsieh Y.-L., Yang D.-L., Wu J. y Chen Y.-C. (2016) Efficient mining of profit rules from closed inter-transaction itemsets. *Journal of Information Science and Engineering* 32(3): 575–595.

[HYYZ20] Han Y., Yu D., Yin C. y Zhao Q. (2020) Temporal association rule mining and updating and their application to blast furnace in the steel industry. *Computational Intelligence and Neuroscience* 2020.

[HZLH12] Huang F., Zou Z., Liu X. y He J. (2012) Association rules mining for academic cooperation based on time extension and duration accumulation. En *International Conference on Computer Science and Service System*, páginas 2007–2012. Nanjing, China.

[IBC+03] Irizarry R., Bolstad B., Collin F., Cope L. y Hobbs B. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* 31(4): e15.

[INN04] Ishibuchi H., Nakashima T. y Nii M. (2004) *Classification and Modeling with Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining*. Springer-Verlag, Berlin.

[JYT+16] Ji Y., Ying H., Tran J., Dews P., Lau S.-Y. y Massanari R. (2016) A functional temporal association mining approach for screening potential drug-drug interactions from electronic patient databases. *Informatics for Health and Social Care* 41(4): 387–404.

[KGS+11] Kanehisa M., Goto S., Sato Y., Furumichi M. y Tanabe M. (2011) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* 40(D1): D109–D114.

[KKA+19] Kilgore P. C. S. R., Korneeva N., Arnold T. C., Trutschl M. y Cvek U. (2019) Gatewaynet: a form of sequential rule mining. *BMC Medical Informatics and Decision Making* 19:87(1): 1–13.

[KP18] Khan S. y Parkinson S. (2018) Eliciting and utilising knowledge for security event log analysis: An association rule

mining and automated planning approach. *Expert Systems with Applications* 113: 116–127.

[LCL03] Lee C.-H., Chen M.-S. y Lin C.-R. (2003) Progressive partition miner: an efficient algorithm for mining general temporal association rules. *IEEE Transactions on Knowledge and Data Engineering* 15(4): 1004–1017.

[LD05] Li D. y Deogun J. S. (2005) Discovering partial periodic sequential association rules with time lag in multiple sequences for prediction. En *International Symposium on Methodologies for Intelligent Systems*, páginas 332–341. NY, USA.

[LFH00] Lu H., Feng L. y Han J. (2000) Beyond intratransaction association analysis: Mining multidimensional intertransaction association rules. *ACM Transactions on Information Systems* 18(4): 423–454.

[LFW05] Li Q., Feng L. y Wong A. (2005) From intra-transaction to generalized inter-transaction: Landscaping multidimensional contexts in association rule mining. *Information Sciences* 172(3): 361–395.

[LHTD02] Liu H., Hussain F., Tan C. y Dash M. (2002) Discretization: An enabling technique. *Data Mining and Knowledge Discovery* 6(4): 393–423.

[LJL08] Lee W.-J., Jiang J.-Y. y Lee S.-J. (2008) Mining fuzzy periodic association rules. *Data and Knowledge Engineering* 65(3): 442–462.

[LK17] Liang Y. y Kelemen A. (2017) Dynamic modeling and network approaches for omics time course data: Overview of computational approaches and applications. *Briefings in Bioinformatics* 19(5): 1051–1068.

[LKW09] Lo D., Khoo S.-C. y Wong L. (2009) Non-redundant sequential rules - theory and algorithm. *Information Systems* 34(4-5): 438–453.

[LL02] Lin M.-Y. y Lee S.-Y. (2002) Fast discovery of sequential patterns by memory indexing. En Kambayashi Y., Winiwarter W. y Arikawa M. (Eds.) *Data Warehousing and Know-*

ledge Discovery, páginas 150–160. Springer Berlin Heidelberg, Berlin, Heidelberg.

[LL04] Lee W.-J. y Lee S.-J. (2004) Discovery of fuzzy temporal association rules. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 34(6): 2330–2342.

[LLC01] Lee C.-H., Lin C.-R. y Chen M.-S. (2001) Sliding-window filtering: An efficient algorithm for incremental mining. En *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM 2001)*, páginas 263–270. Atlanta, Georgia, USA.

[LLC+09] Lee Y. J., Lee J. W., Chai D. J., Hwang B. H. y Ryu K. H. (2009) Mining temporal interval relational rules from temporal data. *Journal of Systems and Software* 82(1): 155–167.

[LNWJ03] Li Y., Ning P., Wang X. S. y Jajodia S. (2003) Discovering calendar-based temporal association rules. *Data and Knowledge Engineering* 44(2): 193–218.

[LOW02] Lin W., Orgun M. y Williams G. (2002) An overview of temporal data mining. En *Australian data mining workshop (ADM 2002)*, páginas 83–90. Canberra, Australia.

[LP16] Lyubimov D. y Palumbo A. (2016) *Apache Mahout: Beyond MapReduce*. CreateSpace Independent Publishing Platform, USA, 1st edition.

[LRRV12] Lo D., Ramalingam G., Ranganath V.-P. y Vaswani K. (2012) Mining quantified temporal rules: Formalism, algorithms, and evaluation. *Science of Computer Programming* 77(6): 743–759.

[LSU05] Laxman S., Sastry P. S. y Unnikrishnan K. P. (2005) Discovering frequent episodes and learning hidden Markov models: A formal connection. *IEEE Transactions on Knowledge and Data Engineering* 17(11): 1505–1517.

[LSU07] Laxman S., Sastry P. y Unnikrishnan K. (2007) Discovering frequent generalized episodes when events persist for different durations. *IEEE Transactions on Knowledge and Data Engineering* 19(9): 1188–1201.

[LUS02] Laxman S., Unnikrishnan K. P. y Sastry P. S. (2002) Generalized frequent episodes in event sequences. En *International Conference on Knowledge Discovery and Data Mining, Workshop on Temporal Data Mining (ACM SIGKDD 2002)*, páginas 1–7. Edmonton, Alberta, Canada.

[LWP+16] Liu L., Wang S., Peng Y., Huang Z., Liu M. y Hu B. (2016) Mining intricate temporal rules for recognizing complex activities of daily living under uncertainty. *Pattern Recognition* 60: 1015–1028.

[LXL+17] Liu Z., Xue Y., Li M., Ma B., Zhang M., Chen X. y Hu X. (2017) Discovery of deep order-preserving submatrix in dna microarray data based on sequential pattern mining. *International Journal of Data Mining and Bioinformatics* 17(3): 217–237.

[LZL+19] Liu C., Zhang Q., Luo H., Qi S., Tao S., Xu H. y Yao Y. (2019) An efficient approach to capture continuous impervious surface dynamics using spatial-temporal rules and dense landsat time series stacks. *Remote Sensing of Environment* 229: 114–132.

[MCC18] Miholca D.-L., Czibula G. y Czibula I. (2018) A novel approach for software defect prediction through hybridizing gradual relational association rules with artificial neural networks. *Information Sciences* 441: 152–170.

[MCCC20] Moodley R., Chiclana F., Caraffini F. y Carter J. (2020) A product-centric data mining algorithm for targeted promotions. *Journal of Retailing and Consumer Services* 54: 101940.

[MER08] Moskovitch R., Elovici Y. y Rokach L. (2008) Detection of unknown computer worms based on behavioral classification of the host. *Computational Statistics & Data Analysis* 52(9): 4544–4566.

[MGH11] Matthews S. G., Gongora M. A. y Hopgood A. A. (2011) Evolving temporal association rules with genetic algorithms. En Bramer M., Petridis M. y Hopgood A. (Eds.) *Research and Development in Intelligent Systems XXVII*, páginas 107–120. Springer, London.

[MGH13] Matthews S. G., Gongora M. A. y Hopgood A. A. (2013) Evolutionary algorithms and fuzzy sets for discovering temporal rules. *International Journal of Applied Mathematics and Computer Science* 23(4): 855–868.

[MGMA+11] Marrades M., González-Muniesa P., Arteta D., Alfredo Martínez J. y Moreno-Aliaga M. (2011) Galectin-12: A protein associated with lipid droplets that regulates lipid metabolism and energy balance. *Journal of Physiology and Biochemistry* 67: 15–26.

[MH09] Ma S. y Huang J. (2009) Regularized gene selection in cancer microarray meta-analysis. *BMC bioinformatics* 10(1): 1–12.

[MR13] Mooney C. H. y Roddick J. F. (2013) Sequential pattern mining – approaches and algorithms. *ACM Computing Surveys* 45(2): 1–39.

[MS15] Moskovitch R. y Shahar Y. (2015) Fast time intervals mining using the transitivity of temporal relations. *Knowledge and Information Systems* 42(1): 21–48.

[MTV97] Mannila H., Toivonen H. y Verkamo A. I. (1997) Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery* 1(3): 259–289.

[NKD09] Nam H., KiYoung L. y Doheon L. (2009) Identification of temporal association rules from time-series microarray data sets. *BMC Bioinformatics* 10(3): 1–9.

[NLPV18] Nguyen D., Luo W., Phung D. y Venkatesh S. (2018) LTARM: A novel temporal association rule mining method to understand toxicities in a routine cancer treatment. *Knowledge-Based Systems* 161: 313–328.

[NRC11] Nazerfard E., Rashidi P. y Cook D. (2011) Using association rule mining to discover temporal relations of daily activities. En *International Conference on Smart Homes and Health Telematics (ICOST 2011)*, páginas 49–56. Montreal, Canada.

[NSET20] Nasu M., Shimamura K., Esumi S. y Tamamaki N. (2020) Sequential pattern of sublayer formation in the paleocortex and neocortex. *Medical Molecular Morphology* 53: 168–176.

[ODS+18] Orphanou K., Dagliati A., Sacchi L., Stassopoulou A., Keravnou E. y Bellazzi R. (2018) Incorporating repeating temporal association rules in naive bayes classifiers for coronary heart disease diagnosis. *Journal of Biomedical Informatics* 81: 74–82.

[ORS98] Ozden B., Ramaswamy S. y Silberschatz A. (1998) Cyclic association rules. En *International Conference on Data Engineering (ICDE 1998)*, páginas 412–421. Orlando, Florida, USA.

[osi98] (1998) Open Source Initative. 1998. `http://www.opensource.org/docs/osd`.

[PKSG09] Papapetrou P., Kollios G., Sclaroff S. y Gunopulos D. (2009) Mining frequent arrangements of temporal intervals. *Knowledge and Information Systems* 21(2): 133–171.

[PP18] Panchal M. C. y Prajapati G. I. (2018) Calendric association rule mining from time series database. En Saeed K., Chaki N., Pati B., Bakshi S. y Mohapatra D. P. (Eds.) *Progress in Advanced Computing and Intelligent Engineering*, páginas 283–293. Springer Singapore.

[RAKJ19] Radhakrishna V., Aljawarneh S. A., Kumar P. V. y Janaki V. (2019) ASTRA - A novel interest measure for unearthing latent temporal associations and trends through extending basic gaussian membership function. *Multimedia Tools and Applications* 78(4): 4217–4265.

[RKJ15] Radhakrishna V., Kumar P. y Janaki V. (2015) An approach for mining similarity profiled temporal association patterns using gaussian based dissimilarity measure. En *Proceedings of the The International Conference on Engineering & MIS (ICEMIS 2015)*, páginas 1–6. Istanbul, Turkey.

[RM05] Roddick J. y Mooney C. (2005) Linear temporal sequences and their interpretation using midpoint relationships. *IEEE*

Transactions on Knowledge and Data Engineering 17(1): 133–135.

[RMS98a] Ramaswamy S., Mahajan S. y Silberschatz A. (1998) On the discovery of interesting patterns in association rules. En *24rd International Conference on Very Large Data Bases*, páginas 368–379. California, USA.

[RMS98b] Ramaswamy S., Mahajan S. y Silberschatz A. (1998) On the discovery of interesting patterns in association rules. En *International Conference on Very Large Data Bases (VLDB 1998)*, páginas 368–379. San Francisco, CA, USA.

[RS02] Roddick J. y Spiliopoulou M. (2002) A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering* 14(4): 750–767.

[SA96] Srikant R. y Agrawal R. (1996) Mining quantitative association rules in large relational tables. En *International Conference on Management of data (ACM SIGMOD 1996)*, páginas 1–12.

[SB75] Shortliffe E. y Buchanan B. (1975) A model of inexact reasoning in medicine. *Mathematical Biosciences* 23: 351–379.

[SDGAAF20] Segura-Delgado A., Gacto M. J., Alcalá R. y Alcalá-Fdez J. (2020) Temporal association rule mining: An overview considering the time variable as an integral or implied component. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10(4): e1367.

[Sha86] Shaffer J. (1986) Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association* 81(395): 826–831.

[She03] Sheskin D. (2003) *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, London, U.K.

[SK20] Sarker I. H. y Kayes A. (2020) Abc-ruleminer: User behavioral rule-based machine learning method for context-aware intelligent services. *Journal of Network and Computer Applications* 168: 102762.

[SKK01] Song H., Kim J. y Kim S. (2001) Mining the change of customer behavior in an internet shopping mall. *Expert Systems with Applications* 21(3): 157–168.

[SM11] Saleh B. y Masseglia F. (2011) Discovering frequent behaviors: time is an essential element of the context. *Knowledge and Information Systems* 28(2): 311–331.

[SPPAF16] Segura A., Pérez-Pérez R. y Alcalá-Fdez J. (2016) New open source modules in KEEL to analyze and export fuzzy association rules. En *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2016)*, páginas 225–232. Vancouver, Canada.

[Sud05] Sudkamp T. (2005) Discovery of fuzzy temporal associations in multiple data streams. En *Soft Computing: Methodologies and Applications*, páginas 3–13. Springer.

[SY18] Setiawan F. y Yahya B. N. (2018) Improved behavior model based on sequential rule mining. *Applied Soft Computing* 68: 944–960.

[TAKG15] Tebourski W., Abdessalem Kar W. y Ghezala H. (2015) From concrete to inferred knowledge: Enhanced mining constraint-based cyclic association rules from medical social network. *International Journal of Knowledge-Based and Intelligent Engineering Systems* 19(2): 109–116.

[TCL90] Teng H. S., Chen K. y Lu S. C. (1990) Adaptive real-time anomaly detection using inductively generated sequential patterns. En *IEEE Computer Society Symposium on Research in Security and Privacy*, páginas 278–284. CA, USA.

[Thi03] Thimbleby H. (2003) Explaining code for publication. *Software - Practice and Experience* 33: 975–1001.

[TK12] Tebourski W. y Karaa W. B. A. (2012) Cyclic association rules mining under constraints. *International Journal of Computer Applications* 49(20): 30–37.

[TKS02] Tan P., Kumar V. y Srivastava J. (2002) Selecting the right interestingness measure for association patterns. En *8th International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, páginas 32–41. Edmonton, Canada.

[TLHF03] Tung A. K. H., Lu H., Han J. y Feng L. (2003) Efficient mining of intertransaction association rules. *IEEE Transactions on Knowledge and Data Engineering* 15(1): 43–56.

[TLVH16] Tran M.-T., Le B., Vo B. y Hong T.-P. (2016) Mining non-redundant sequential rules with dynamic bit vectors and pruning techniques. *Applied Intelligence* 45(2): 333–342.

[TSWY13] Tseng V. S., Shie B., Wu C. y Yu P. S. (2013) Efficient algorithms for mining high utility itemsets from transactional databases. *IEEE Transactions on Knowledge and Data Engineering* 25(8): 1772–1786.

[TTT12] Thuan N. D., Toan N. G. y Tuan N. L. V. (2012) An approach mining cyclic association rules in e-commerce. En *International Conference on Network-Based Information Systems*, páginas 408–411. Melbourne, Australia.

[VJT+17] Vandromme M., Jacques J., Taillard J., Hansske A., Jourdan L. y Dhaenens C. (2017) Extraction and optimization of classification rules for temporal sequences: Application to hospital data. *Knowledge-Based Systems* 122: 148–158.

[VLLZ18] Vu H., Li G., Law R. y Zhang Y. (2018) Travel diaries analysis by sequential rule mining. *Journal of Travel Research* 57(3): 399–413.

[VRA+16] Vink R. G., Roumans N. J. T., Arkenbosch L. A. J., Mariman E. C. M. y van Baak M. A. (2016) The effect of rate of weight loss on long-term weight regain in adults with overweight and obesity. *Obesity* 24(2): 321–327.

[VRF+17] Vink R., Roumans N., Fazelzadeh P., Tareen S., Boekschoten M., van Baak M. y Mariman E. (2017) Adipose tissue gene expression is differentially regulated with different rates of weight loss in overweight and obese humans. *International journal of obesity* 41: 309–316.

[VV05] Verma K. y Vyas O. P. (2005) Efficient calendar based temporal association rule. *ACM SIGMOD Record* 34(3): 63–70.

[VVV05] Verma K., Vyas O. P. y Vyas R. (2005) Temporal approach to association rule mining using t-tree and p-tree. En *Inter-*

national Workshop on Machine Learning and Data Mining in Pattern Recognition, páginas 651–659. Leipzig, Germany.

[Wan15] Wang C.-S. (2015) Mining non-redundant inter-transaction rules. *Journal of Information Science and Engineering* 31(6): 1849–1865.

[WFH16] Witten I., Frank E. y Hall M. (2016) *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann $fourthedition$.

[WHSZ21] Wang L., Huang R., Shi W. y Zhang C. (2021) Domino effect in marine accidents: Evidence from temporal association rules. *Transport Policy* 103: 236–244.

[WLPF18] Wu J. M., Lin J. C., Pirouz M. y Fournier-Viger P. (2018) Tub-haupm: Tighter upper bound for mining high average-utility patterns. *IEEE Access* 6: 18655–18669.

[WMXP18] Wang L., Meng J., Xu P. y Peng K. (2018) Mining temporal association rules with frequent itemsets tree. *Applied Soft Computing Journal* 62: 817–829.

[WR07] Winarko E. y Roddick J. (2007) Armada - an algorithm for discovering richer relative temporal association rules from interval-based data. *Data and Knowledge Engineering* 63(1): 76–90.

[Wu10] Wu R. (2010) Mining generalized fuzzy association rules from web logs. En *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010)*, páginas 2474–2477. Yantai, China.

[WZS+19] Wen F., Zhang G., Sun L., Wang X. y Xu X. (2019) A hybrid temporal association rules mining method for traffic congestion prediction. *Computers & Industrial Engineering* 130: 779–787.

[XLT+19] Xiao Q., Li C., Tang Y., Li L. y Li L. (2019) A knowledge-driven method of adaptively optimizing process parameters for energy efficient turning. *Energy* páginas 142–156.

[XTZ14] Xiao Y., Tian Y. y Zhao Q. (2014) Optimizing frequent time-window selection for association rules mining in a temporal database using a variable neighbourhood search. *Computers & Operations Research* 52: 241–250.

[XWZ20] Xie D., Wang M.-H. y Zhao X.-M. (2020) A Spatiotemporal Apriori Approach to Capture Dynamic Associations of Regional Traffic Congestion. *IEEE Access* 8: 3695–3709.

[YCKG09] Yu W., Clyne M., Khoury M. J. y Gwinn M. (2009) Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics* 26(1): 145–146.

[YHL12] Yang R., Havel P. y Liu F. (2012) Galectin-12: A protein associated with lipid droplets that regulates lipid metabolism and energy balance. *Adipocyte* 1(2): 96–100.

[YNY21] Yang D., Nie Z. T. y Yang F. (2021) Time-aware cf and temporal association rule-based personalized hybrid recommender system. *Journal of Organizational and End User Computing* 33: 19–34.

[ZDA17] Zihayat M., Davoudi H. y An A. (2017) Mining significant high utility gene regulation sequential patterns. *BMC Systems Biology* 11(Suppl 6): 1–88.

[ZFVW+15] Zida S., Fournier-Viger P., Wu C.-W., Lin J. C.-W. y Tseng V. S. (2015) Efficient mining of high-utility sequential rules. En Perner P. (Ed.) *Machine Learning and Data Mining in Pattern Recognition*, páginas 157–171. Springer International Publishing, Cham.

[ZH17] Zhou H. y Hirasawa K. (2017) Evolving temporal association rules in recommender system. *Neural Computing and Applications* páginas 1–15.

[ZH19] Zhou H. y Hirasawa K. (2019) Evolving temporal association rules in recommender system. *Neural Computing and Applications* 31(7): 2605–2619.

[Zim14] Zimmermann A. (2014) Understanding episode mining techniques: Benchmarking on diverse, realistic, artificial data. *Intelligent Data Analysis* 18(5): 761–791.

[ZZ02] Zhang C. y Zhang S. (2002) *Association Rule Mining: Models and Algorithms*. Springer.