UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Uso de patrones secuenciales multivariantes para clasificación y extracción de conocimiento temporal. Estudio de supervivencia de pacientes en la Unidad de Quemados Críticos

D. Isidoro Jesús Casanova López

2023

UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Uso de patrones secuenciales multivariantes para clasificación y extracción de conocimiento temporal. Estudio de supervivencia de pacientes en la Unidad de Quemados Críticos

Autor: D. Isidoro Jesús Casanova López

Directores: D. Manuel Campos Martínez, D. José Manuel Juárez Herrero, D. José Ángel Lorente Balanza

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD
DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR
Aprobado por la Comisión General de Doctorado el 19-10-2022

D. **Isidoro Jesús Casanova López**

doctorando del Programa de Doctorado en

Informática *(Inteligencia Artificial: Fundamentos y Aplicaciones en Ciencias de la Vida e Ingeniería)*

de la Escuela Internacional de Doctorado de la Universidad Murcia, como autor de la tesis presentada para la obtención del título de Doctor y titulada:

Uso de patrones secuenciales multivariantes para clasificación y extracción de conocimiento temporal. Estudio de supervivencia de pacientes en la Unidad de Quemados Críticos

y dirigida por,

D. **Manuel Campos Martínez**

D. **José Manuel Juárez Herrero**

D. **José Ángel Lorente Balanza**

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Si la tesis hubiera sido autorizada como tesis por compendio de publicaciones o incluyese 1 o 2 publicaciones (como prevé el artículo 29.8 del reglamento), declarar que cuenta con:

- *La aceptación por escrito de los coautores de las publicaciones de que el doctorando las presente como parte de la tesis.*
- *En su caso, la renuncia por escrito de los coautores no doctores de dichos trabajos a presentarlos como parte de otras tesis doctorales en la Universidad de Murcia o en cualquier otra universidad.*

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

En Murcia, a 11 de noviembre de 2022

Fdo.: Isidoro Jesús Casanova López

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la primera página de la tesis presentada para la obtención del título de Doctor.

Información básica sobre protección de sus datos personales aportados	
Responsable:	Universidad de Murcia. Avenida teniente Flomesta, 5. Edificio de la Convalecencia. 30003; Murcia. Delegado de Protección de Datos: dpd@um.es
Legitimación:	La Universidad de Murcia se encuentra legitimada para el tratamiento de sus datos por ser necesario para el cumplimiento de una obligación legal aplicable al responsable del tratamiento. art. 6.1.c) del Reglamento General de Protección de Datos
Finalidad:	Gestionar su declaración de autoría y originalidad
Destinatarios:	No se prevén comunicaciones de datos
Derechos:	Los interesados pueden ejercer sus derechos de acceso, rectificación, cancelación, oposición, limitación del tratamiento, olvido y portabilidad a través del procedimiento establecido a tal efecto en el Registro Electrónico o mediante la presentación de la correspondiente solicitud en las Oficinas de Asistencia en Materia de Registro de la Universidad de Murcia

Agradecimientos

No puede impedirse el viento, pero pueden construirse molinos[1]

Realizar esta investigación y compaginarla con mi trabajo diario ha supuesto un gran reto en mi vida. Se necesita mucha paciencia, y todavía más constancia, perseverancia y resiliencia, en un largo y continuo aprendizaje cíclico. Por todo ello, esta tesis se la dedico a los pilares fundamentales que componen mi vida, mis padres y mi mujer Josefina.

[1] Los patrones que se repiten frecuentemente en los que se basa esta tesis se asemejan a los *proverbios*, ya que son muchos, y suelen describir situaciones interesantes que se repiten a lo largo del tiempo de las que no solemos percatarnos. Al estudiar detalladamente los proverbios veremos que solo existen unos pocos que podamos considerar verdaderamente interesantes, y son minoritarios aquellos que realmente nos proporcionan un nuevo conocimiento práctico de utilidad en nuestra vida. En esta tesis buscaremos aquellos patrones minoritarios que nos aporten realmente un nuevo conocimiento médico, con la esperanza de poder convertirlos en nuevos proverbios.

Resumen

Objetivos: Esta tesis centra su investigación en los datos provenientes de pacientes con grandes quemados que han ingresado en la UCI y cuya evolución ha sido registrada diariamente. Algunos parámetros clínicos disponibles desde la llegada del paciente, como la edad o la extensión del quemado, permiten realizar una primera valoración de la severidad y ayudan a pronosticar la supervivencia estimada en la admisión. Sin embargo, el estudio de la evolución de otros parámetros clínicos registrados durante los primeros 5 días de estancia en la UCI (pH, diuresis, exceso de base, ...) puede ayudar a definir objetivos y a valorar la evolución y respuesta al tratamiento. En esta tesis se plantea la generación de potencial conocimiento observando la evolución en el tiempo de estas variables, de forma que se podría llegar a pronosticar la supervivencia de un paciente o sugerir nuevas ideas a los médicos acerca del comportamiento de estas variables.

Metodología: Se define inicialmente un proceso de descubrimiento de conocimiento con los siguientes 4 pasos: 1) discretización de los atributos temporales, 2) minería de patrones secuenciales multivariantes, 3) posprocesamiento, filtrando como patrones interesantes aquellos que sean discriminatorios, y aplicando posteriormente una representación comprimida de estos patrones, 4) clasificación de la supervivencia de los pacientes con modelos interpretables. Posteriormente compararemos como diferentes discretizaciones afectan a la clasificación e intentaremos reducir el número de patrones secuenciales usados como predictores en los clasificadores, realizando una evaluación de su consistencia. Además, proponemos el uso de un indicador estadístico ampliamente utilizado en estudios epidemiológicos, la Razón de Probabilidades Diagnóstica (DOR), como medida alternativa de interés respecto a la frecuencia, para realizar la selección de los patrones interesantes. Por último, presentamos un original método para obtener un subconjunto reducido de patrones secuenciales novedosos que representan la evolución temporal sorpresiva del estado clínico del paciente, a los que llamaremos como *Jumping Diagnostic Odds Ratio Sequential Patterns (JDORSP)*. Utilizaremos el DOR para seleccionar aquellos patrones secuenciales que representen un cambio drástico en la evolución del paciente, es decir, patrones que se convierten en un factor de protección cuando extendemos un patrón que era un factor de riesgo, o viceversa.

Resultados: Los resultados de las pruebas de clasificación muestran que nuestro enfoque supera a las puntuaciones de gravedad de quemaduras utilizadas actualmente por los médicos siguiendo la puntuación de Brier, y hasta donde sabemos, este sería el primer trabajo donde patrones secuenciales multivariantes se utilizan como predictores de mortalidad en la UCI. Respecto a la utilización de diferentes discretizaciones, que conozcamos, ningún estudio previo ha realizado esta comparación utilizando patrones secuenciales específicamente. El mejor rendimiento con la clasificación se ha obtenido con la discretización automática UCPD. También obtenemos un resultado aceptable con la discretización experta, superando a muchos algoritmos de discretización automática. Mediante la evaluación de la consistencia, hemos reducido aún más el número de patrones secuenciales, encontrado patrones uniformemente distribuidos por toda la base de datos de pacientes. Respecto a la utilización de una métrica estadística, el DOR, como medida de interés para reducir el número de patrones secuenciales y seleccionar sólo los más discriminatorios, con la discretización experta, la mayor especificidad se alcanza utilizando directamente el valor del DOR para seleccionar los patrones. Esta es, hasta donde sabemos, la primera vez que algunos de estos enfoques han sido propuestos y comparados en la literatura científica. Por último, respecto a los novedosos patrones *JDORSP* propuestos, que conozcamos, esta es la primera vez en la que el DOR y los patrones secuenciales se utilizan de esta manera. Destacamos la drástica reducción de patrones secuenciales con respecto al estado actual de la técnica, permitiendo realizar una revisión manual por expertos médicos de la sorpresividad y relevancia de los patrones descubiertos. Así, el hecho más interesante encontrado es la alta sorpresividad en los patrones secuenciales que inicialmente tienen un factor de riesgo, y sus extensiones se convierten en un factor de protección, es decir, pacientes que se recuperan a los pocos días de estar en alto riesgo de morir.

Palabras clave: Inteligencia artificial, descubrimiento de conocimiento en bases de datos, minería de datos, patrones secuenciales, clasificación de supervivencia, patrones emergentes, patrones discriminatorios, calidad e interés de los patrones, razón de probabilidades, unidad de quemados críticos

Índice general

Capítulo 1

Introducción

1.1. Antecedentes y motivación

Nuestra propuesta de investigación parte de los pacientes que han pasado por la Unidad de Grandes Quemados de un Servicio de Medicina Intensiva y cuya evolución ha sido registrada. Algunos parámetros clínicos disponibles desde la llegada del paciente, como la edad, la presencia de lesiones en la inhalación o la extensión y profundidad del quemado, permiten una primera valoración de la severidad [104] [108]. Otras variables, como la necesidad temprana de ventilación mecánica o el tipo de herida del quemado, ayudan a pronosticar mejor la supervivencia estimada en la admisión. Sin embargo, la evolución en el tiempo de otros parámetros clínicos durante la fase de resucitación (primeros 2 días) y durante la fase de estabilización (siguientes 3 días) puede ser también importante.

La evaluación inicial y resucitación de pacientes con importantes quemados que requieren cuidados intensivos solo puede ser guiada por fórmulas y reglas. La inherente inexactitud de las fórmulas requiere la continua reevaluación y ajuste del tratamiento basado en el objetivo de la resucitación. La entrada de líquidos, diuresis, balance de fluidos o el balance de ácido base (pH, bicarbonato, exceso de base) pueden ayudar a definir los objetivos y valorar la evolución y respuesta del tratamiento.

En esta tesis se plantea la generación de potencial conocimiento que pueda sugerir nuevas ideas a los médicos acerca del comportamiento de estas variables, de manera que se podría llegar a pronosticar la supervivencia de un paciente, observando la evolución en el tiempo de estas variables.

Desde hace décadas se está trabajando en el descubrimiento de patrones presentes en bases de datos. Uno de los tipos de patrones de interés son los patrones frecuentes. Encontrar patrones frecuentes juega un papel esencial en la extracción de asociaciones, correlaciones y

muchas otras relaciones interesantes entre los datos. Los patrones frecuentes son conjuntos de elementos o subestructuras que aparecen en un conjunto de datos con una frecuencia no inferior a un umbral especificado por el usuario. Por ejemplo, un conjunto de elementos, como tener los síntomas de una gripe, con fiebre, tos y dolor de garganta, que aparecen frecuentemente juntos en el historial clínico de los pacientes de un centro sanitario en invierno, es un conjunto de elementos frecuentes.

La complejidad de los patrones puede ser mayor si tenemos en consideración la dimensión temporal. De ahí surgen los patrones secuenciales, donde importa el orden que ocupan los elementos, que pueden ser registrados con o sin una noción concreta de tiempo. Así, por ejemplo, normalmente en un proceso gripal, el primer síntoma que se siente es la irritación en la garganta, para a continuación aparecer tos acompañada de fiebre, y finalmente la fiebre y los demás síntomas tienden a desaparecer.

Un orden temporal es bastante natural en muchos escenarios, por lo que se recopilan diariamente enormes cantidades de datos secuenciales en diferentes aplicaciones, como por ejemplo en entornos médicos, biológicos o financieros. De forma que las secuencias se pueden usar para estudiar la evolución de una enfermedad en un paciente registrada en su historia clínica, o para capturar cómo se comportan los humanos a través de su historial de actividad, tal y como podrían ser los clics realizados por cada cliente en una página web o su historial de compras. Las secuencias también se pueden usar para describir cómo se comportan las empresas a través de sus históricos de ventas, como las ventas totales de varios artículos a lo largo del tiempo en un supermercado.

A los patrones se les pueden aplicar técnicas adicionales de descubrimiento del conocimiento para construir modelos predictivos, en los que se estiman valores futuros o desconocidos de variables de interés usando otras variables. Una de estas técnicas es la clasificación, que consiste en categorizar los datos en distintas clases. Esta categorización puede ser un etiquetado de los datos (aprendizaje supervisado), la división de los datos en clases (aprendizaje no supervisado), la selección de las características más significativas de los datos (selección de características) o una combinación de más de una de estas tareas.

Diseñar algoritmos que sean capaces de aprender patrones y modelos de clasificación a partir de estos datos es uno de los temas más desafiantes en la investigación de minería de datos [11]. Un enfoque por el cual lidiar con este problema es el de descubrir patrones que se utilizan como predictores en algoritmos de clasificación [17].

El primer trabajo en el que se realizan clasificaciones utilizando patrones secuenciales fue de Lesh, Zaki y Ogihara [68], para posteriormente darle una utilidad práctica proporcionando un algoritmo que predice los fallos de los planes de ejecución en bases de datos [133]. Sin embargo, es una línea aún por explotar debido a los pocos trabajos realizados.

Uno de los problemas abiertos en el descubrimiento de conocimiento es que el número de patrones generados inicialmente suele ser muy grande, donde es probable que solo unos pocos de estos patrones sean de interés para el experto del dominio que analiza los datos. Hay varias razones para esto: muchos de los patrones son irrelevantes u obvios, no proporcionan nuevo conocimiento con respecto al dominio, o son similares o están incluidos en otros. Este problema tiene una repercusión especial cuando queremos usar los patrones como variables predictoras para mejorar los clasificadores. Por ejemplo, en un dominio médico, podríamos utilizar no sólo los datos de ingreso para determinar la supervivencia del paciente, sino también incluir los patrones que contengan la evolución inmediata del paciente.

Por lo tanto, frente a la obtención inicial de un alto número de patrones, se requieren medidas sobre el nivel de interés que nos ayuden a ordenar o a reducir este número de patrones, aumentando así el aprovechamiento, utilidad y relevancia de los patrones descubiertos [40]. Tampoco podemos olvidarnos de que, en un entorno clínico, lo que le preocupa al médico no es únicamente lo sofisticado que puede llegar a ser el método para minar los datos, sino también cómo de comprensibles son los resultados [71].

El descubrimiento de características discriminatorias y diferencias humanamente interpretables entre conjuntos de datos con etiquetas de clase es uno de los objetivos más importantes en la minería de datos. Esta desafiante tarea comprende un grupo de técnicas de minería de patrones diseñadas para descubrir un conjunto de patrones significativos que se producen con frecuencias dispares en diferentes conjuntos de datos con etiqueta de clase [54].

Los patrones discriminatorios pueden capturar las diferencias entre conjuntos de datos con varias etiquetas de clase, lo que contribuye a crear clasificadores eficaces y a la caracterización de estas clases. Además, estos patrones tienen un valor considerable en una amplia gama de aplicaciones, como la detección del grupo de riesgo del paciente en la medicina o la identificación de características distintivas en la gestión de relaciones con el cliente [93]. Por lo tanto, la identificación de tales patrones es un trabajo valorado en cuestiones prácticas.

La investigación sobre patrones discriminatorios evoluciona rápidamente bajo varias definiciones, como conjuntos de contraste, patrones emergentes o subgrupos. Sin embargo, estas definiciones son prácticamente equivalentes ya que sus patrones objetivo se pueden utilizar indistintamente con la misma capacidad de capturar las diferencias entre clases distintas [54].

En general, los algoritmos existentes para la minería de patrones discriminatorios se pueden dividir en dos categorías de acuerdo con sus estrategias: algoritmos con umbrales especificados por el usuario y algoritmos basados en la significación estadística [54].

La mayoría de los estudios de exploración de patrones discriminatorios adoptan un pro-

cedimiento de dos pasos: primero generar un conjunto de patrones candidatos (es decir, patrones frecuentes en una clase); y entonces, realizar una prueba de significación estadística para evaluar su capacidad discriminatoria y podar patrones insignificantes [25, 54].

En el primer paso, para generar un conjunto de patrones candidatos, la frecuencia de un patrón se puede calcular mediante su soporte, que se define como el porcentaje de individuos de la base de datos que contiene este patrón. Un *patrón es frecuente* si su valor de soporte no es menor que un umbral determinado. Para el segundo paso, la significancia estadística de los patrones discriminatorios puede medirse mediante varias pruebas estadísticas como la razón de probabilidades (odds ratio), la ganancia de información [25] o el test chi-cuadrado [13] entre otras. Se define que un patrón es significativo si su valor de significación generado a partir de una determinada medida estadística puede cumplir algunas condiciones definidas por el usuario, por ejemplo, no más (o menos) que un umbral determinado.

1.2. Caso de estudio: Unidad de Grandes Quemados

Esta tesis centra su investigación en los datos provenientes de pacientes que han ingresado en la Unidad de Grandes Quemados de un Servicio de Medicina Intensiva y cuya evolución ha sido registrada. A continuación, se describen las principales cuestiones clínicas más utilizadas en el ámbito médico, incluyendo la valoración de gravedad del paciente.

1.2.1. Quemaduras

Las quemaduras son lesiones de la piel y el tejido adyacente causadas por un agente físico, químico o biológico. Existen varias clasificaciones de las quemaduras, aunque las más utilizadas son las siguientes (véase [38]):

- *Según la extensión*: de acuerdo con el área de la Superficie Corporal Total (SCT) quemada, en inglés Total Burn Surface Area (TBSA), que se puede calcular con la regla de los nueves de Wallace o si se quiere mayor precisión se utiliza la carta de Lund-Browder.

- *Según la profundidad*: para calcular la profundidad se utilizan tres grados; sin embargo, en la evaluación inicial del servicio de urgencias basta con clasificar las quemaduras en superficiales o profundas.

- *Según la severidad*: de acuerdo con los criterios de la Sociedad Americana de Quemaduras, existen las siguientes categorías: a) Quemaduras críticas b) Quemaduras moderadas y c) Quemaduras menores.

Cuando sobreviene una quemadura mayor, se produce una cascada de cambios fisiológicos, los cuales forman el escenario clínico del paciente quemado. Según [31], estos trastornos incluyen:

- *Desbalance hidroelectrolítico*: La herida por quemadura se edematiza rápidamente debido a los cambios microvasculares, inducidos en forma directa por el calor e indirectamente por la liberación de mediadores químicos de respuesta inflamatoria en la zona de lesión. Esto resulta en pérdida intravascular sistémica de agua, sodio, albúmina y glóbulos rojos. En este escenario, el desarrollo del shock hipovolémico es inminente al menos que no se restaure el volumen desplazado hacia los espacios intersticiales en forma rápida y adecuada.

- *Trastornos metabólicos*: Éstos se evidencian por el aumento del consumo de oxígeno en reposo (hipermetabolismo), pérdida excesiva de nitrógeno (catabolismo) y pérdida pronunciada de masa corporal (desnutrición).

- *Contaminación bacteriana de tejidos*: Los tegumentos lesionados facilitan una zona extensa para la infección superficial o invasión de microorganismos; los pacientes quemados desarrollan compromiso en casi todos los aspectos del sistema inmune, aumentando los riesgos de shock séptico.

Actualmente es más probable la supervivencia tras quemaduras extensas, gracias a los avances en la comprensión de la fisiopatología de la quemadura y el tratamiento más agresivo de ésta. Para ello se requiere de un tratamiento prehospitalario efectivo, transporte, reanimación, sostén de funciones vitales y reparación de la cubierta cutánea [88].

1.2.2. Reanimación y estabilización del paciente quemado

Uno de los mayores avances del siglo XX en el tratamiento de los pacientes quemados ha sido el desarrollo y la adopción de guías clínicas para la reanimación de los pacientes quemados en la Unidad de Quemados Críticos [33]. La fase de reanimación hemodinámica, que se prefiere denominar por los anglosajones como resucitación del paciente quemado (ver Figura 1.1), tiene como objetivo restituir las pérdidas de fluido originadas por el secuestro y la evaporación de líquidos [9]. La reanimación mediante fluidoterapia se debe iniciar lo antes posible en los pacientes en los que las quemaduras cubren más del 15 % de su SCT; en caso contrario, el paciente puede experimentar un shock hipovolémico [30]. También se debe colocar un catéter urinario permanente para vigilar estrechamente la eliminación de orina.

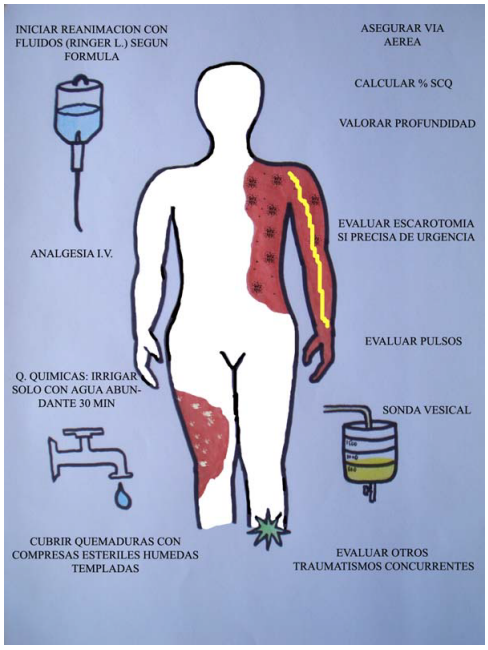

Figura 1.1: Resucitación del gran quemado ([100]).

Antes de los años cuarenta, los pacientes con quemaduras moderadas y graves, desarrollaban shock hipovolémico, fracaso renal y muerte. A partir de entonces comienzan a desarrollarse múltiples fórmulas para calcular las soluciones hidroelectrolíticas necesarias en la reanimación de los pacientes quemados. Difieren en el tipo de cristaloide, la proporción de coloide y el momento de su administración.

La aplicación de estas fórmulas ha logrado disminuir considerablemente la mortalidad en las primeras 24-48 horas tras la quemadura debido a una mayor comprensión de los desplazamientos masivos de fluidos desde el espacio intravascular al espacio extravascular (intracelular e intersticial) durante la fase de shock de la quemadura [82].

El objetivo primario de la reanimación mediante fluidoterapia es restituir el fluido perdido como consecuencia de la quemadura. A juzgar por las múltiples fórmulas de resucitación propuestas a lo largo de los años, no existe ninguna pauta que pueda considerarse universal [82]. Debe ponerse énfasis en que, sea cual sea la forma utilizada, no es más que una guía y que la cantidad administrada en concreto debe individualizarse de acuerdo con la respuesta clínica. En principio, debe administrarse la menor cantidad de fluido necesario para mantener la perfusión de los órganos, y el fluido administrado debe contener sodio para reemplazar la perdida de sodio extracelular en el tejido quemado y en interior de la célula. Las diferentes fórmulas están basadas en la experiencia clínica; así la extensión de la quemadura, su profundidad y el peso corporal del enfermo son los factores de los que depende el ritmo y composición del fluido a infundir.

Dados los volúmenes masivos de líquidos administrados por vía intravenosa a los pacientes que presentan quemaduras (son frecuentes las tasas de 1.000ml/h), es imprescindible una vigilancia estrecha del estado hemodinámico para prevenir una sobrecarga hídrica [30]. Los signos y síntomas del «exceso de líquido», es decir, de la sueroterapia de reanimación excesiva en función de lo que predice la fórmula de Parkland, son los siguientes: síndrome compartimental abdominal, síndrome compartimental en las extremidades y síndrome de dificultad respiratoria aguda.

Una de las carencias más importantes en esta fase inicial es la falta de un objetivo válido y mensurable para una óptima reanimación de los pacientes. Esta dificultad se debe a la compleja fisiopatología de la quemadura y a los múltiples factores de confusión que rodean al daño térmico. La diuresis sigue siendo piedra angular en la monitorización de los quemados, a pesar de la reciente aparición de parámetros más sofisticados (como el exceso de base, considerado un marcador de la mala perfusión tisular relacionado con la morbilidad) [33].

Tras la primera fase de reanimación (primeros 2 días), entraríamos en la fase de estabilización (3 días siguientes), donde el objetivo de la fluidoterapia es el mantenimiento del equilibrio interno, el ajuste de los electrolitos y las alteraciones en el equilibrio ácido-básico. En esta nueva fase existe una mayor estabilidad cardiovascular y se puede hacer un mayor énfasis en el manejo de las quemaduras.

La inexactitud inherente de las fórmulas requiere una reevaluación y ajuste continuos de las infusiones basadas en objetivos de reanimación. Los líquidos administrados, la diuresis, el equilibrio de fluidos, o el equilibrio ácido-base (pH, bicarbonato, exceso de base) entre otros, ayudan a definir objetivos y a evaluar la evolución y la respuesta al tratamiento.

1.2.3. Predicción de la mortalidad y sistemas de puntuación de la gravedad

Los avances de la estadística a finales del siglo XX con el análisis multifactorial y la regresión logística permitieron analizar simultáneamente varios factores que influyen en la mortalidad y calcular su importancia de forma independiente.

Uno de los grandes avances en el tratamiento de los quemados durante los años setenta, fue la popularización del desbridamiento tangencial propuesto por Janzekovic [130]. Por primera vez se identifican así factores de riesgo adquiridos o intrahospitalarios, como el tipo y extensión de desbridamiento, el balance hídrico positivo en las primeras 48 horas de reanimación y la bacteriemia. Por otra parte, otros factores, como la edad, el tanto por ciento de SCT quemada y el componente de tercer grado están íntimamente relacionados con la mortalidad [33].

La predicción temprana de la mortalidad después de la admisión es esencial antes de que una terapia agresiva o conservadora se pueda recomendar, adecuando el esfuerzo terapéutico a cada paciente [33]. Las puntuaciones de la gravedad son herramientas simples pero útiles para los médicos que permiten evaluar el estado del paciente [107]. Los sistemas de puntuación tienen como objetivo utilizar los factores premórbidos y de lesiones más predictivos para calcular una probabilidad esperada de muerte para un paciente determinado.

Las puntuaciones de Baux y el Índice de Quemaduras de Pronóstico (Prognostic Burn Index, PBI, por sus siglas en inglés) proporcionan una tasa de mortalidad al sumar la edad y el porcentaje de SCT quemada.

En general se considera que los pronósticos basados en la edad y la SCT quemada tienen suficiente especificidad para la predicción clínica, aunque su baja sensibilidad exija la adición del análisis de otros factores de riesgo para incrementar su exactitud pronóstica [60].

Tobiasen et al. [113] utilizaron la regresión logística multivariante para realizar el índice abreviado de gravedad de las quemaduras (ABSI, "Abbreviated Burn Severity Index"), en el que también se considera el género, la profundidad y la presencia o no de lesión pulmonar por inhalación.

1.2.4. Base de datos y discretización aplicada

La base de datos en la que se basa esta tesis tiene 480 registros de pacientes ingresados entre 1992 y 2002 recogidos en el Hospital Universitario de Getafe, uno de los 7 centros nacionales de referencia en quemados críticos.

En el primer artículo que realizamos [20], recogido en el Capítulo 4, decidimos eliminar todos los pacientes con datos faltantes sobre las variables seleccionadas para este estudio, quedando únicamente 379 pacientes. Entre otros motivos, se eliminaron tantos pacientes para poder comparar los resultados con las puntuaciones de gravedad de la quemadura. En los restantes experimentos realizados, recogidos en los siguientes capítulos, de la base de datos únicamente se eliminaron aquellos pacientes que fallecieron en el transcurso del estudio o aquellos en los que no se pudiera estimar las horas de estancia del primer día, quedando un total de 465 pacientes, de los que el 81.29 % (378/87) son supervivientes, el 69.68 % (324/141) son hombres y el 43.23 % (201/264) tienen daños por inhalación. La Tabla 1.1 muestra un resumen de los atributos estáticos.

Los atributos temporales que permiten el seguimiento y la evaluación de la respuesta al tratamiento de los pacientes son registrados durante cinco días. Todos los atributos son variables continuas y representan el valor acumulado durante 24 horas. Las variables registradas son:

Atributo	Mín.	Máx.	Media	Desviación estándar
Edad (edad)	9	95	46.42	20.34
Peso (kg)	25	120	71.05	10.77
Duración de la estancia (días)	3	162	25.02	24.24
Superficie Corporal Total quemada (%)	1	90	31.28	20.16
Superficie quemada profunda (%)	0	90	17.01	17.41
Escala de severidad SAPS II	6	58	20.67	9.49

Tabla 1.1: Resumen atributos estáticos (465 pacientes)

- Total líquidos entrantes administrados (incomings): *INC* (medido en cc).
- Diuresis: *DIU* (en cc).
- Balance de fluidos: *BAL* (en cc).
- Bicarbonato: *BIC* (en mmol/L).
- Escala de acidez de la piel: *pH* ([0...14]).
- Exceso de base: *BE* (en mEq/L).

Nótese que el balance de fluidos no es la diferencia entre los ingresos y la diuresis, sino que se consideran todas las posibles eliminaciones de fluidos. Se presenta a continuación una discusión de cómo se ha procedido discretizar por un experto estas diferentes variables, ya que este tipo de discretización experta se utilizará en todos los experimentos de los diferentes capítulos. Además, se puede encontrar una discusión sobre el uso de otras posibles discretizaciones automáticas en el Capítulo 5.

1.2.4.1. Discretización experta aplicada para cada variable

La discretización con rangos de referencia para cada una de las variables ha sido realizada por un experto, y se ha determinado a partir de una variedad de fuentes.

- Los entrantes (Incomings, INC) son los fluidos administrados medidos originalmente en cc, aunque decidimos hacerlos uniformes al peso del paciente y al porcentaje de SCT quemada de acuerdo con la fórmula de reanimación de Parkland (la más utilizada). Hemos utilizado cuartiles de pacientes con todos los valores para hacer cuatro intervalos: [<, 2.3), [2.3, 3.66), [3.66, 5.78), [5.78, >]. Nótese que uno de los objetivos de la fase de reanimación es la restauración de fluidos.

- La unidad habitual con significado para diuresis (DIU) es cc/kg/h, por lo que hemos dividido todos los valores entre peso y 24 (cada valor es registro diario con acumulado 24 horas). Los términos clínicos generalmente utilizados en adultos para la diuresis son oliguria por debajo de 0.5 cc/kg/h, diuresis normal dentro de 0.5 y 1 cc/kg/h, y diuresis aumentada de más de 1 cc/kg/h. Los valores por encima de 1 indican un funcionamiento normal, pero hemos utilizado un tercer cuartil para diferenciar un valor aumentado respecto a valores realmente altos. Los intervalos definidos son [0, 0.5), [0.5, 1), [1, 1.9), [1.9, >].

- El balance de fluidos (BAL) mide la diferencia entre los entrantes y la salida total de líquido (no solo la diuresis). Un valor deseado sería un equilibrio ligeramente positivo. En este caso, el objetivo terapéutico se cumple todos los días, y hemos utilizado la mediana de los días consecutivos como una forma de mejora. Definimos los intervalos [<, -2), [-2, 10.5), [10.5, 20.4), [20.4, 52.22), [52.22, >], siendo 10.5, 20.4 y 52.22 las medianas del cuarto, tercer y segundo día respectivamente.

- Para bicarbonato, pH y exceso de base, no existe un criterio estándar para realizar una discretización cualitativa. Los niveles normales de bicarbonato (BIC) están dentro de [21,25] mmol/L, y hemos definido los siguientes intervalos utilizando la diferencia intercuartil Q3-Q1 para crear una referencia alrededor de los valores normales: [<, 17), [17,21), [21,25), [25,29), [29, >].

- Una posible abstracción del pH es la distinción de los valores en acidosis grave [<,7.20), acidosis moderada [7.20, 7.30), acidosis leve [7.30, 7.35), normal [7.35, 7.45), alcalosis leve [7.45, 7.50), alcalosis moderada [7.50, 7.60] y alcalosis grave [7.6, >).

- La abstracción cualitativa del exceso de base (base excess, BE) se realiza con respecto al pH y manteniendo el pCO2 en 40 mmHg. Los valores normales de BE están entre -2 y 2 mEq/L. Hemos utilizado 4 intervalos para diferentes niveles de acidosis y alcalosis: [<,-4), [-4,-2), [-2,2), [2,4), [4,>].

En la Tabla 1.2 se muestra un resumen de los diferentes intervalos de discretización aplicados. Para una mejor comprensión de la representación de la discretización en los patrones, explicamos a modo de ejemplo el siguiente patrón usando la discretización experta $INC_0 < DIUR_1 < PH_2$, de forma que $_i$ marca el intervalo de discretización i donde $i = 0$ es el intervalo más bajo. Esta secuencia temporal comienza con entrantes de fluidos administrados menores de 2.3 (INC_0), le sigue la diuresis entre 0.5 y 1 ($DIUR_1$), y más tarde el pH se encuentra entre 7.3 y 7.35 (PH_2).

Núm. intervalo	INC	DIUR	BAL	BIC	pH	BE
0	[0, 2.30)	[0, 0.5)	[-, -2.0)	[-, 17)	[-, 7.20)	[-, -4)
1	[2.30, 3.66)	[0.5, 1.0)	[-2.0, 10.5)	[17, 21)	[7.20, 7.30)	[-4, -2)
2	[3.66, 5.78)	[1.0, 1.9)	[10.5, 20.4)	[21, 25)	[7.30, 7.35)	[-2, 2)
3	[5.78, -)	[1.9, -)	[20.4, 52.22)	[25, 29)	[7.35, 7.45)	[2, 4)
4			[52.22, -)	[29, -)	[7.45, 7.50)	[4, -)
5					[7.50, 7.60)	
6					[7.60, -)	

Tabla 1.2: Intervalos de discretización de cada atributo mediante discretización experta.

Capítulo 2

Objetivos propuestos, organización y contribuciones

2.1. Hipótesis y objetivos

La **HIPÓTESIS** de partida de esta tesis es que los patrones secuenciales proporcionan nuevo conocimiento clínico relevante acerca de la evolución de los pacientes y además aportan un alto valor en la construcción de modelos de clasificación en problemas clínicos, ya que permiten obtener modelos con buena precisión.

Planteamos una serie de objetivos que permiten responder a diferentes aspectos de la hipótesis de partida:

OBJETIVO 1: Establecer si un modelo de clasificación basado en patrones secuenciales es comparable a las puntuaciones de gravedad de quemaduras utilizadas actualmente por los médicos, como son los sistemas de puntuación clínicos Baux, R-Baux, ABSI o SAPSII. Además, estas puntuaciones se basan en variables recogidas en un momento puntual que se corresponde con la evaluación inicial del paciente, y no incorporan la evolución temporal.

OBJETIVO 2: Establecer qué propiedades de los patrones secuenciales y qué posprocesamiento reducen significativamente el número de patrones usados como variables predictoras, a la vez que se mejoran los resultados de la clasificación. Este objetivo trata de paliar la sobredimensionalidad del problema debida al enorme número de patrones descubiertos por los algoritmos de minería de patrones. Para ello estudiaremos las representaciones comprimidas de los patrones frecuentes, como son los patrones cerrados o los patrones maximales. Además se seleccionarán aquellos patrones discrimi-

natorios que revelan una gran diferencia en la distribución subyacente de las diferentes clases. Todas estas propiedades podrían incluso incorporarse en los propios algoritmos de minería de patrones.

OBJETIVO 3: Determinar el mejor preprocesamiento en los datos para obtener patrones más relevantes desde el punto de vista de la clasificación y para maximizar la interpretación y utilidad desde el punto de vista médico. Centraremos este preprocesamiento en la establecer la influencia de la discretización previa a la obtención de los patrones secuenciales, comparando métodos automáticos de discretización con métodos basados en criterios clínicos.

OBJETIVO 4: Establecer medidas de interés para seleccionar patrones secuenciales y encontrar aquellos patrones con más poder discriminatorio. Los algoritmos de minería de patrones se basan en medidas de frecuencia, pero creemos que el uso de medidas de interés basadas en métricas estadísticas, como la Razón de Probabilidades Diagnóstica (Diagnostic Odds Ratio, DOR), puede ser usada también de distintas maneras para seleccionar un conjunto de patrones discriminatorios.

OBJETIVO 5: Evaluar la capacidad de extracción de conocimiento clínico utilizable a partir de los patrones secuenciales usados en la clasificación o seleccionados directamente en el posprocesamiento. De esta manera, estableceremos si se cumple uno de los objetivos primordiales de la minería de datos que es encontrar patrones significativos, útiles y novedosos.

2.2. Organización de la memoria

A continuación se detalla el contenido del resto de la tesis que permite ver los antecedentes y como se han logrado los objetivos propuestos.

Se repasa el estado del arte en el **Capítulo 3**, en donde se muestran conceptos de minería de datos, y los pasos necesarios para realizar la minería de patrones secuenciales, y cómo se pueden utilizar estos patrones en un clasificador. También se indica cómo se puede mejorar la calidad de los patrones descubiertos y reducir su número.

Para lograr el primer y el segundo objetivo, en el **Capítulo 4**, basado en el trabajo presentado en la 15^{th} *Conference on Artificial Intelligence in Medicine* [20], se predice la supervivencia de los pacientes en el contexto de las Unidades de Quemados Críticos. Para ello se define un proceso de descubrimiento de conocimiento de 4 pasos consistente en:

- Paso 0: Discretización de los atributos temporales (se utilizará la realizada por un experto).
- Paso 1: Minería de patrones secuenciales multivariantes
- Paso 2: Posprocesamiento, filtrando como patrones interesantes aquellos que sean discriminatorios mediante la utilización de la frecuencia como marco de trabajo, y aplicando posteriormente una representación comprimida de estos patrones.
- Paso 3: Clasificación con modelos interpretables

Hasta donde sabemos, este es el primer trabajo donde patrones secuenciales multivariantes se utilizan como predictores de mortalidad en una Unidad de Quemados Críticos o en una UCI.

Para alcanzar el tercer objetivo, en el **Capítulo 5**, basado en un artículo publicado en *Progress in Artificial Intelligence* [22], que amplía asimismo el artículo presentado en el *XVII Congreso de la Asociación Española para la Inteligencia Artificial* [21], analizamos y comparamos la discretización realizada por un experto contra la discretización automática. En particular, estudiamos el impacto de la discretización para predecir la supervivencia de los pacientes quemados estudiados en el primer capítulo. Que conozcamos, este es también el primer estudio donde se realiza una comparación de la discretización utilizando patrones secuenciales.

En los siguientes capítulos se intentará mejorar la calidad de los patrones descubiertos, reduciendo su número, para quedarnos solamente con aquellos que realmente puedan tener mayor relevancia, consistencia o poder discriminatorio, utilizando varias técnicas.

Con el fin de seguir mejorando el segundo objetivo, en el **Capítulo 6**, basado en al artículo presentado en el *XVI Congreso de la Asociación Española para la Inteligencia Artificial* [19], se plantea reducir el número de patrones usados como predictores en los clasificadores de la supervivencia de pacientes en la Unidad de Quemados Críticos, realizando una evaluación de la consistencia de los patrones secuenciales multivariantes. Para ello se ha introducido un proceso general de validación de k-particiones estratificadas, quedándonos solamente con aquellos patrones que tengan la mayor relevancia médica posible y se encuentren uniformemente distribuidos por toda la base de datos de pacientes.

Para lograr el cuarto objetivo, en el **Capítulo 7**, basado en el artículo publicado en *JMIR Medical Informatics* [24], se propone el uso de un indicador estadístico, la Razón de Probabilidades Diagnóstica (DOR), como medida alternativa de interés respecto a la frecuencia, para realizar la selección de los patrones interesantes. La razón de probabilidades —en inglés, odds ratio (OR)— es una medida estadística utilizada en estudios epidemiológicos y

por lo tanto conocida en entornos médicos. Realizamos la propuesta y evaluación de cuatro formas de emplear el DOR para reducir el número de patrones secuenciales y seleccionar sólo los patrones más discriminatorios, ya que la explosión del número de patrones es el principal problema en los clasificadores basados en patrones. Hasta donde sabemos, esta es la primera vez que algunos de estos enfoques han sido propuestos y comparados en la literatura científica.

Para conseguir el cuarto y el quinto objetivo, aunque este último objetivo se podría considerar general en todos los capítulos, en el **Capítulo 8**, basado en el artículo enviado a *Data Mining and Knowledge Discovery* [23], proponemos un original método para obtener un subconjunto reducido de patrones temporales novedosos que representan la evolución temporal sorpresiva del estado clínico del paciente, a los que llamaremos como *Jumping Diagnostic Odds Ratio Sequential Patterns (JDORSP)*. Usamos la medida DOR, presentada anteriormente, para seleccionar aquellos patrones secuenciales que representan un cambio drástico en la evolución del paciente, es decir, patrones que se convierten en un factor de protección cuando extendemos un patrón que era un factor de riesgo, o viceversa. Según nuestro conocimiento, es la primera vez que el DOR y los patrones secuenciales se utilizan de esta manera. Debido a la drástica reducción conseguida en el número de patrones, se puede realizar por expertos médicos una revisión manual de la sorpresividad y relevancia de los patrones descubiertos. De manera que se ha encontrado una alta sorpresividad (4.9 sobre 5) en los patrones secuenciales que inicialmente tienen un factor de riesgo, y sus extensiones se convierten en un factor de protección, es decir, pacientes que se recuperan a los pocos días de estar en alto riesgo de morir.

Por último, en el **Capítulo 9** exponemos las conclusiones obtenidas al realizar esta tesis y el trabajo futuro.

En la Tabla 2.1 se resume para cada capítulo los objetivos alcanzados y donde se ha realizado su publicación.

Capítulo	Objetivos	Publicación
4 Metodología para uso de patrones secuenciales con clasificación	1, 2	AIME 2015 [20]
5 Impacto de la discretización de series temporales en la clasificación	3	CAEPIA 2016 [21], Progress in Artificial Intelligence [22]
6 Evaluación de la consistencia de patrones secuenciales multivariantes	2	CAEPIA 2015 [19]
7 Clasificación mediante patrones discriminatorios usando DOR	4	JMIR Medical Informatics [24]
8 Descubriendo novedosos patrones secuenciales (JDORSP) usando DOR	4, 5	Data Mining and Knowledge Discovery [23]

Tabla 2.1: Objetivos y publicaciones realizadas por capítulo

2.3. Contribuciones científicas derivadas de la tesis doctoral

2.3.1. Publicaciones en congresos

- [19] Isidoro J. Casanova, Manuel Campos, Jose M. Juarez, Antonio Fernandez-Fernandez-Arroyo, y Jose A. Lorente. *Evaluación de consistencia de patrones secuenciales multivariable para predecir la supervivencia de pacientes en la unidad de quemados críticos.* En XVI Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2015), págs. 31–40. 2015. ISBN 978-84-608-4099-2.

- [20] Isidoro J. Casanova, Manuel Campos, Jose M. Juarez, Antonio Fernandez-Fernandez-Arroyo, y Jose A. Lorente. *Using multivariate sequential patterns to improve survival prediction in intensive care burn unit*. En Artificial Intelligence in Medicine (AIME 2015). Lecture Notes in Computer Science, tomo 9105, págs. 277–286. Springer International Publishing, 2015. ISBN 978-3-319-19551-3. doi:10.1007/978-3-319-19551-3_36. GII-GRIN Conference Rating 2015 (Class 3).

- [21] Isidoro J. Casanova, Manuel Campos, Jose M. Juarez, Antonio Fernandez-Fernandez-Arroyo, y Jose A. Lorente. *Impact of discretization with multivariate sequential patterns to do the classification of the survival prediction in intensive care burn unit*. En XVII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2016), págs. 847–856. 2016. ISBN 978-84-9012-632-5.

2.3.2. Publicaciones en revistas[1]

- [22] Isidoro J. Casanova, Manuel Campos, Jose M. Juarez, Antonio Fernandez-Fernandez-Arroyo, y Jose A. Lorente. *Impact of time series discretization on intensive care burn unit survival classification*. Progress in Artificial Intelligence, 7(1):41–53, 2018. ISSN 2192-6360. doi:10.1007/s13748-017-0130-8. 2018 SJR 0.513 (Q2).

- [23] Isidoro J. Casanova, Manuel Campos, Jose M. Juarez, Antonio Gomariz, Bernardo Canovas-Segura, Marta Lorente-Ros, y Jose A. Lorente. *Surprising and novel multivariate sequential patterns for temporal evolution in healthcare*. Data Mining and Knowledge Discovery, 2023, under review. 2021 SJR 1.615 (Q1). 2021 JIF 5.406 (Q1).

- [24] Isidoro J. Casanova, Manuel Campos, Jose M. Juarez, Antonio Gomariz, Marta Lorente-Ros, y Jose A. Lorente. *Using the Diagnostic Odds Ratio to Select Patterns to Build an Interpretable Pattern-Based Classifier in a Clinical Domain: Multivariate Sequential Pattern Mining Study*. JMIR Medical Informatics, 10(8):e32319, 2022. ISSN 2291-9694. PMID: 35947437. doi:10.2196/32319. 2021 SJR 0.805 (Q2). 2021 JIF 3.228 (Q3).

[1] Para cada revista se muestra el indicador SCImago Journal Rank (SJR) y adicionalmente, en caso de estar incluida en el Journal Citation Reports (JCR), el factor de impacto de la revista (Journal Impact Factor, JIF), mostrando el año base utilizado para realizar el cálculo y el mejor cuartil (Q) en el que se posiciona esa revista en relación con todas las de su área.

Capítulo 3

Estado del arte

3.1. Introducción

En este estado del arte, vamos a exponer las bases y referencias de los dos tópicos principales que componen esta tesis, la minería de patrones frecuentes y la clasificación con estos patrones.

La minería de datos es el proceso por el que se encuentran en los datos patrones novedosos, válidos y potencialmente útiles. En minería de datos, la minería de patrones frecuentes es uno de los problemas más investigados. Numerosos algoritmos se han propuesto para resolver la minería de patrones frecuentes o algunas de sus variantes. Los datos suelen tener un orden temporal en muchos escenarios, por lo que un problema relacionado con la minería de patrones frecuentes es la minería de patrones secuenciales, donde este orden en las transacciones es tenido en cuenta.

Se han definido diferentes marcos de trabajo para la minería de patrones frecuentes. El más común es el marco basado en el soporte, en el que se encuentran conjuntos de ítems cuya frecuencia se encuentra por encima de un umbral determinado. Sin embargo, de esta manera el número de patrones generados es muy grande, porque estos patrones varían muy ligeramente entre sí, y es probable que sólo unos pocos de estos patrones sean de interés para el experto en el dominio que analiza los datos. Así, muchos de estos patrones son irrelevantes u obvios, y no proporcionan nuevo conocimiento.

Para reducir el enorme conjunto de patrones frecuentes generados con la minería de datos, mientras se mantiene una alta calidad de estos patrones, estudios recientes se han centrado en la minería de un conjunto comprimido o aproximado de patrones frecuentes. En general, la compresión de patrones se puede dividir en dos categorías: compresión sin pérdida y compresión con pérdida, en términos de la información que contiene el conjunto de

resultados, en comparación con todo el conjunto de patrones frecuentes.

Para aumentar la utilidad, relevancia y el aprovechamiento de los patrones descubiertos, en lugar de utilizar únicamente el marco de trabajo basado en el soporte, se han propuesto una serie de medidas de interés para reducir el número de patrones. El desarrollo de estas medidas de interés es actualmente un área de investigación activa en el Descubrimiento de Conocimiento en Bases de Datos (o KDD, del inglés "Knowledge Discovery in Databases").

La clasificación es un proceso de análisis de datos basado en la búsqueda de un conjunto de modelos que describan y distingan entre dos o más clases de datos. Cada modelo se realiza analizando un conjunto de datos de entrenamiento que han sido etiquetados explícitamente con la clase a la que pertenecen. A continuación, el modelo se utiliza para predecir la clase de los objetos cuyas etiquetas de clase son desconocidas. La clasificación normalmente ayuda a proporcionar un buen entendimiento de los datos analizados. Es posible combinar el descubrimiento de patrones frecuentes con la clasificación cuando los patrones se usan como variables predictoras para construir el modelo de clasificación. En este caso surge la necesidad de reducir drásticamente el número de patrones usados porque se llegaría a un problema de sobredimensionamiento donde habría más variables productoras que instancias. Por ello se busca seleccionar patrones discriminatorios.

El descubrimiento de características distintivas y diferencias humanamente interpretables entre conjuntos de datos con etiquetas de clase es uno de los objetivos más importantes en la minería de datos. Esta característica inherente de los patrones discriminatorios los hace fácil de entender y por lo tanto pueden ser utilizados directamente por las personas.

3.2. Minería de datos

La minería de datos es el proceso de utilizar una o más técnicas de aprendizaje computacional para analizar conjuntos de datos observacionales y encontrar relaciones insospechadas, resumiendo estos datos en formas novedosas que son comprensibles y útiles para el propietario de los datos.

Las relaciones, estructuras y resúmenes extraídos a través de un proceso de minería de datos a menudo se denominan modelos o patrones. Entre ellos podemos encontrar reglas, ecuaciones lineales, clústeres, gráficos, estructuras de árboles o patrones recurrentes en series temporales. Estos modelos o patrones que se encuentran dentro de un conjunto de datos deben, por supuesto, ser novedosos para el propietario de los datos. No tiene mucho sentido descubrir relaciones que ya están bien establecidas.

La minería de datos es una parte integral del KDD, que es el proceso general de conversión de datos sin procesar en información útil. Este término se originó en el campo de

investigación de la inteligencia artificial.

El proceso de KDD implica varias fases, desde el preprocesamiento de datos hasta el posprocesamiento de los resultados de minería de datos: seleccionar los datos objetivo, preprocesar los datos, transformarlos si es necesario, realizar la minería de datos para extraer patrones y relaciones, y a continuación, interpretar y evaluar las estructuras descubiertas (consulte la Figura 3.1).

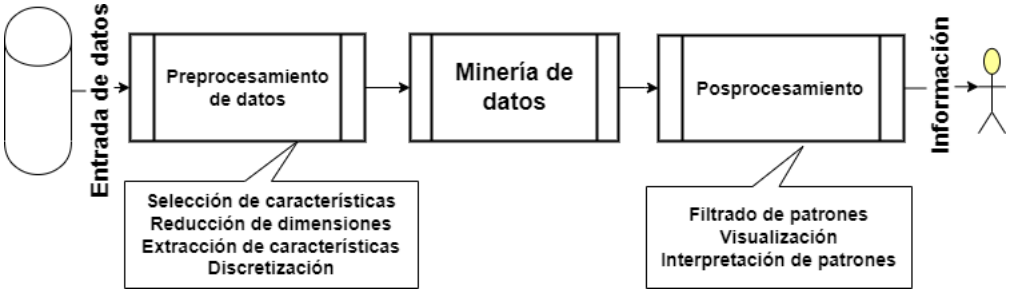

Figura 3.1: Etapas de transformación en el proceso de descubrimiento de conocimiento en bases de datos (KDD)

Los autores en [53] categorizan la minería de datos en cuatro tipos de tareas, correspondientes a diferentes objetivos para la persona que está analizando los datos. Esta categorización, que no es única y que sería posible dividirla en tareas más específicas, captura los principales tipos de actividades y algoritmos en minería de datos.

- **Análisis Exploratorio de Datos (AED)**: Como su nombre indica, el objetivo aquí es simplemente explorar los datos sin tener las ideas claras sobre lo que estamos buscando. Generalmente, las técnicas de AED son interactivas y visuales.
- **Modelado Descriptivo**: El objetivo de un modelo descriptivo es definir todos los datos (o el proceso de generación de los datos). Ejemplos de estas descripciones incluyen modelos para la distribución de probabilidad general de los datos (estimación de densidad), partición del espacio p-dimensional en grupos (análisis de clúster y segmentación) o modelos que describen la relación entre variables (modelado de dependencia). Así, por ejemplo, dentro de la partición del espacio en grupos, un análisis de la segmentación tiene como objetivo agrupar registros similares, como sucede en la segmentación del mercado de bases de datos comerciales. Aquí el objetivo es dividir los registros en grupos homogéneos para que personas similares (si los registros se refieren a personas) se pongan en el mismo grupo.
- **Modelado Predictivo (Clasificación y Regresión)**: El objetivo aquí es crear un modelo que permita predecir el valor de una variable a partir de los valores conocidos de

otras variables. En la clasificación, la variable que se está prediciendo es categórica, mientras que en regresión la variable es cuantitativa. El término "predicción" se utiliza aquí en un sentido general, y no implica ninguna noción de un continuo temporal. Así, por ejemplo, si bien podríamos predecir el valor del mercado de valores en una fecha futura, también podríamos querer determinar el diagnóstico de un paciente. Se han desarrollado un gran número de métodos en estadística o con aprendizaje automático para hacer frente a los problemas de modelado predictivo, y el trabajo en esta área ha dado lugar a avances teóricos significativos y a una mejor comprensión de las cuestiones profundas sobre la inferencia. La distinción clave entre predicción y descripción es que la predicción tiene como objetivo una variable única (la clase de enfermedad, el valor del mercado, ...), mientras que en los problemas descriptivos ninguna variable es fundamental para el modelo.

- **Descubriendo patrones y reglas**: Los tres tipos de tareas enumeradas anteriormente están relacionadas con la construcción de modelos. Otras aplicaciones de la minería de datos se ocupan de la detección de patrones. Un ejemplo es el descubrimiento de comportamientos fraudulentos mediante la detección de regiones en el espacio que definen diferentes tipos de transacciones donde los puntos de datos son significativamente diferentes del resto. Otro uso es la tarea de encontrar combinaciones de ítems que ocurren con frecuencia en las bases de datos transaccionales (por ejemplo, productos de alimentación que se compran a menudo juntos). Este problema ha sido el foco de mucha atención en minería de datos y se ha abordado utilizando técnicas algorítmicas basadas en reglas de asociación.

3.2.1. Preprocesamiento de datos

El propósito del preprocesamiento es transformar los datos de entrada brutos (sin tratar) en un formato adecuado para un siguiente análisis. Los pasos implicados en el preprocesamiento de datos incluyen la unión de datos de varias fuentes (agregación), la selección de un subconjunto de los datos objetivo que se van a analizar (muestreo), la limpieza de datos para eliminar el ruido y las observaciones duplicadas (reducción de dimensionalidad y selección de subconjuntos de características), la selección de registros y características que son relevantes para la tarea de minería de datos (extracción de características, construcción y asignación) y la transformación de un atributo continuo o discreto en un atributo categórico (discretización y binarización). Debido a las muchas maneras en que se pueden recopilar y almacenar datos, el preprocesamiento de datos es quizás el paso más laborioso y lento en el proceso general de KDD.

3.2.1.1. Discretización

La discretización es uno de los pasos del preprocesamiento dentro del proceso de KDD, en el que se transforma un atributo numérico de un conjunto de datos en un atributo cualitativo. Desde el punto de vista teórico, la discretización pretende obtener la mejor partición (mediante intervalos no superpuestos) del dominio continuo de dicho atributo [45].

Los métodos de discretización se utilizan esencialmente para dos propósitos. Por un lado, muchos estudios destacan que las tareas de inducción pueden beneficiarse de la discretización porque las características discretas están más cerca de una representación a nivel de conocimiento que las continuas. Por otro lado, los datos también se pueden simplificar y reducir a través de la discretización.

Tanto para usuarios no expertos como para expertos, las características discretas son fáciles de entender, explicar y usar [79]. En general, la discretización hace que el aprendizaje sea más preciso y rápido, y los modelos obtenidos (árboles de decisión, reglas de inducción) utilizando atributos discretos suelen ser más compactos y precisos que los continuos. Desde el punto de vista empírico, la discretización ayuda a examinar los resultados con detalle, al igual que a compararlos, usarlos y reutilizarlos. Además, existe una serie de algoritmos de aprendizaje de clasificación que solo se pueden utilizar cuando hay datos discretos disponibles. Así, por ejemplo, en el caso específico del clasificador Naive Bayes, cuando hay un atributo continuo no es posible calcular las probabilidades reales por cómputo de frecuencias. Una solución común es discretizar el atributo o utilizar una función de densidad de probabilidad.

Los primeros enfoques para discretizar usaban los métodos de misma frecuencia o mismo ancho (o, una forma de agrupación -binning-). Sin embargo, debido a la necesidad de mejorar la precisión y la eficiencia de los sistemas de clasificación, el interés por la discretización aumentó rápidamente. A lo largo de los años, muchos algoritmos de discretización han sido propuestos y probados para mostrar su potencial para reducir la cantidad de datos mientras se mejora la precisión predictiva [79].

Según [45], se puede hacer una categorización de los discretizadores teniendo en cuenta algunas de las siguientes características: técnicas supervisadas frente a técnicas no supervisadas, enfoques dinámicos frente a estáticos, objetivos globales frente a objetivos locales, división (de arriba hacia abajo) frente a métodos de mezcla (merging), y estrategias directas frente a incrementales.

En los métodos no supervisados, los intervalos continuos se dividen en subrangos atendiendo al ancho (rango de valores) o la frecuencia (número de instancias en cada intervalo) especificados por el usuario. El uso de un rango de valores puede no dar buenos resultados en

los casos en que la distribución de los valores continuos no es uniforme. Además, es vulnerable a los valores atípicos, ya que estos afectan significativamente a los rangos. Para superar esta deficiencia, se introdujeron métodos de discretización supervisados que utilizan información de clase para encontrar los intervalos adecuados causados por los puntos de corte [79].

En el problema de la discretización, se debe encontrar un compromiso entre la calidad de la información (intervalos homogéneos con respecto al atributo a predecir) y la calidad estadística (tamaño suficiente de la muestra en cada intervalo para garantizar la generalización). Los criterios basados en chi-cuadrado se centran en el punto de vista estadístico, mientras que los criterios basados en la entropía se centran en el punto de vista de la información. Otros criterios (como el criterio Gini o Fusinter) tratan de encontrar una compensación entre la información y las propiedades estadísticas. Alternativamente, otros autores sugieren métodos como el enfoque contenedor (wrapper) involucrado en el campo de selección de características o enfoques evolutivos [105].

Si consideramos exclusivamente el ámbito clínico, se ha prestado poca atención a la discretización de las características continuas. Exponemos a continuación algunos trabajos que han tratado este tema.

En [32], los autores consideran el papel de la discretización como parte de una evaluación más amplia de los clasificadores para un conjunto de datos de cirugía de trauma. Este estudio analiza árboles de decisión y clasificadores Naive Bayes, y evalúa su rendimiento con discretización de misma frecuencia, y un método supervisado basado en la entropía que incluye la supervisión de un experto del dominio. Los resultados muestran una mejora marginal pero estadísticamente significativa sobre el uso de cuartiles.

Clarke y Barton [27] propusieron un algoritmo de discretización utilizando datos clínicos del Instituto Nacional del Corazón, los Pulmones y la Sangre. Los autores utilizaron un método basado en la entropía para derivar particiones de ciertos atributos clínicos, incluyendo la presión arterial y el índice de masa corporal. En cada uno de estos casos, las etiquetas de clase eran conocidas por cada observación, y se utilizaban para minimizar la pérdida de información causada por la discretización.

Stacey y McGregor [110] desarrollaron un sistema de monitoreo que permite la detección de patrones temporales en múltiples flujos de datos de alta frecuencia para pacientes dentro del dominio de cuidados intensivos neonatales. El sistema utiliza dos métodos automatizados de discretización temporal no supervisados, la discretización de mismo ancho y Symbolic Aggregate approXimation (SAX).

3.2.1.2. Abstracción temporal de series temporales clínicas

El proceso de Abstracción Temporal (AT) es la agregación y/o segmentación de una serie de datos crudos, marcados en el tiempo y multivariantes en una representación simbólica de series de intervalos de tiempo, a menudo en un nivel más alto de abstracción [90] (por ejemplo, en lugar de una serie de valores de pH en crudo, queremos obtener caracterizaciones abstractas como '3 días de acidosis metabólica') y adecuado para una revisión manual humana o para la minería de datos.

La abstracción temporal se ha convertido en un asunto de gran interés en el análisis de datos médicos. Por ejemplo, la investigación en sistemas de Análisis de Datos Interactivos clínicos (Interactive Data Analysis, IDA) donde la alta dimensionalidad y el gran volumen de datos con marca de tiempo son la norma. Su objetivo es transformar los datos de los pacientes de una forma cuantitativa de bajo nivel a descripciones cualitativas de alto nivel, que están más cerca del lenguaje de los médicos [110].

Sin embargo, el contexto del análisis cambia debido al ajuste en la medicación y a la progresión del estado de la enfermedad, lo que significa que los valores de datos que se consideraron normales en un momento, o en un contexto particular, pueden ser peligrosamente anormales en otro. Cuando los datos oscilan a altas frecuencias y/o las frecuencias son irregulares, el proceso de abstracción temporal se vuelve cada vez más complejo.

Algunos dominios abarcan observaciones de baja frecuencia, como los niveles de glucosa en sangre en el dominio de la diabetes mellitus, mientras que otros implican características de alta frecuencia, como por ejemplo la fibrilación de frecuencia cardíaca en el dominio de los cuidados intensivos neonatales. La literatura revisada abstrae los datos, bien de una base de datos o de flujos de datos en línea. En cualquier caso, la frecuencia de muestreo de los datos es un factor importante para el diseño de los mecanismos de abstracción temporal y tiene un impacto en la granularidad de la medida de tiempo que determina si la abstracción temporal se empleará para compensar saltos en los datos o para resumir datos más densos [110].

Desde la perspectiva computacional, los enfoques adoptados son similares a la investigación sobre discretización. Algunos autores proponen explotar los conocimientos sensibles al contexto adquiridos de expertos humanos, un método conocido como Abstracción Temporal Basada en el Conocimiento (Knowledge-Based Temporal Abstraction, KBTA) [106]; otros son puramente automáticos, y dependen principalmente de una discretización de los valores sin procesar y del encadenamiento de valores [57, 102].

La abstracción temporal para la minería de series temporales en la forma de intervalos de tiempo fue ya propuesta por [57]. Varios métodos de discretización comunes, como la

discretización con el mismo ancho, que divide uniformemente los rangos de cada valor, o la discretización de la misma frecuencia, no tienen en cuenta el orden temporal de los valores; otros métodos, en particular, SAX [76] se centra en una discretización estadística de los valores, mientras que Persist [91] maximiza la duración de los intervalos de tiempo resultantes y considera explícitamente la dimensión temporal.

3.2.2. Posprocesamiento

Dentro del proceso de KDD, el conocimiento extraído en el paso anterior de minería de datos podría procesarse aún más. Una posibilidad es simplificar este nuevo conocimiento. Además, podemos evaluar el conocimiento extraído, visualizarlo o simplemente documentarlo para el usuario final. Podemos interpretar el conocimiento e incorporarlo a un sistema existente, y verificarlo en busca de posibles conflictos con el conocimiento previamente producido.

Según [18] el posprocesamiento incluye las siguientes técnicas:

- **Filtrado de conocimiento:** Si los datos de entrenamiento son ruidosos, entonces el algoritmo de aprendizaje genera hojas en un árbol de decisión o bien reglas de decisión que cubren un número muy pequeño de ejemplos de entrenamiento. Esto sucede porque el algoritmo de aprendizaje intenta dividir subconjuntos de los ejemplos de entrenamiento en subconjuntos aún más pequeños para lograr una mayor consistencia. Para superar este problema, el árbol o el conjunto de reglas de decisión debe reducirse, ya sea mediante poda (árboles de decisión) o truncamiento (reglas de decisión).
- **Interpretación y explicación:** Podemos usar el conocimiento adquirido directamente para la predicción o como modelo de base de conocimiento en un sistema experto. Si el proceso de descubrimiento de conocimiento se realiza para un usuario final, generalmente se documentan los resultados obtenidos. Otra posibilidad es visualizar el conocimiento producido, o transformarlo a una forma comprensible para el usuario final. Además, podemos verificar el nuevo conocimiento en busca de posibles conflictos con el conocimiento producido previamente. En este paso, también podemos resumir las reglas y combinarlas con un conocimiento específico del dominio proporcionado para la tarea dada. Cabe señalar que, especialmente, las aplicaciones de sistemas expertos deben ir acompañadas de la función de explicación para que un usuario final acepte la decisión del sistema experto.
- **Evaluación:** Cuando un sistema de aprendizaje induce hipótesis conceptuales (modelos) a partir de un conjunto de entrenamiento, su evaluación (o prueba) debe llevarse a cabo. Hay varios criterios ampliamente utilizados para este propósito, como la preci-

sión de la clasificación, la comprensibilidad o la complejidad computacional. La matriz de confusión se utiliza para evaluar el rendimiento de los enfoques de clasificación y agrupación.

- **Integración del conocimiento:** Los sistemas tradicionales de toma de decisiones han dependido generalmente de una sola técnica o modelo. En cambio, los nuevos sistemas sofisticados de ayuda a la decisión utilizan los resultados obtenidos de varios modelos (bases de conocimiento), cada uno de los cuales generalmente (pero no es obligatorio) se basa en un paradigma diferente, o los combinan o refinan de cierta manera. Por lo tanto, dicho sistema multiestrategia (híbrido) consta de dos o más "componentes" individuales que intercambian información y cooperan entre sí. Este proceso aumenta la precisión y el éxito.

3.3. Minería de patrones frecuentes y de patrones secuenciales

3.3.1. Minería de patrones frecuentes

Los patrones frecuentes son patrones (así como conjuntos de ítems o de subsecuencias) que aparecen con frecuencia en un conjunto de datos. Por ejemplo, un conjunto de ítems, como el pan y la leche, que aparecen juntos frecuentemente en un conjunto de datos compuesto por transacciones de un supermercado, es un conjunto de ítems frecuente. Una subsecuencia, como comprar primero un teléfono móvil, luego una funda de teléfono y, a continuación, una tarjeta de memoria, si se produce con frecuencia en una base de datos con historiales de compras, es un patrón secuencial frecuente.

En minería de datos, la minería de patrones frecuentes es uno de los problemas más intensamente investigado en relación con su desarrollo computacional y algorítmico. Durante las últimas décadas, se han propuesto muchos algoritmos para resolver la minería frecuente de patrones o algunas de sus variantes, y el interés en este problema todavía persiste [2].

La minería de patrones frecuentes detecta conjuntos de ítems que aparecen en un conjunto de datos con una frecuencia no inferior a un umbral especificado por el usuario y desempeña un papel esencial en la minería de asociaciones, correlaciones y muchas otras relaciones interesantes entre los datos. Fue propuesto por primera vez por Agrawal et al. (1993) [5] para el análisis de la cesta de la compra en la forma de minería de reglas de asociación, analizando los hábitos de compra de los clientes mediante la búsqueda de asociaciones entre los diferentes artículos que los clientes colocan en su "cesta de la compra". El descubrimien-

to de estas asociaciones puede ayudar a las empresas a desarrollar estrategias de marketing al obtener información sobre qué artículos son comprados al mismo tiempo con frecuencia por los clientes. Por ejemplo, si los clientes compran gel de ducha, evaluar la probabilidad que tendrían de comprar también acondicionador de cabello (e incluso qué marca) en la misma visita al supermercado. Esta información puede conducir a aumentar las ventas, ayudando a los propietarios de los supermercados a hacer un marketing selectivo y planificar dónde colocar sus productos en los estantes.

Diferentes marcos de trabajo se han definido para la minería de patrones frecuentes. El marco más común se encuentra basado en el soporte, en el que se encuentran conjuntos de ítems con una frecuencia por encima de un umbral determinado. Sin embargo, estos conjuntos de ítems a veces no representan conocimientos novedosos interesantes, y se podrían utilizar otras propiedades para seleccionar estos conjuntos de elementos, como cuantificadores estadísticos, que generarán conjuntos de ítems más interesantes desde una perspectiva estadística. En consecuencia, se han definido una serie de medidas alternativas de interés respecto al soporte en la literatura (puede consultar la Sección 3.5.2 para ampliar información).

3.3.1.1. Definición del problema

El problema de la minería de patrones frecuentes es el de encontrar relaciones entre los ítems en una base de datos. El problema, según [2], se puede definir de la siguiente manera:

Dada una base de datos $\mathcal{D}$ con transacciones $T_1, \ldots, T_N$, determine todos los patrones P que están presentes en al menos una fracción s de las transacciones.

La fracción s se conoce como el soporte mínimo. El parámetro s se puede expresar como un número absoluto o como una fracción del número total de transacciones en la base de datos. Cada transacción T_i se puede considerar como un vector binario disperso, o como un conjunto de valores discretos que representan los identificadores de los atributos binarios que son instanciados al valor de 1.

En el contexto de los datos relacionados con la cesta de la compra, para encontrar grupos de artículos que frecuentemente se compran juntos, cada atributo se corresponde con un artículo de un hipermercado y un valor binario representa si está presente o no en la transacción.

3.3.1.2. Reglas de asociación

En el modelo original de minería de patrones frecuentes [5], también se propuso el problema de encontrar reglas de asociación, el cual está estrechamente relacionado con el problema de patrones frecuentes. En general, las reglas de asociación se pueden considerar como

un resultado de "segunda etapa" que se deriva de los patrones frecuentes. Con el fin de seleccionar reglas de asociación interesantes del conjunto de todas las reglas posibles, se utilizan una serie de restricciones sobre diversas medidas de importancia e interés. Las restricciones más conocidas son los umbrales mínimos de soporte y confianza.

Formalmente, si consideramos los conjuntos de ítems (o elementos) U y V. La regla $U \Rightarrow V$ se considera una regla de asociación con el soporte mínimo s y la confianza mínima c, cuando las dos condiciones siguientes se mantienen verdaderas:

1. El conjunto $U \cup V$ es un patrón frecuente.

2. La ratio del soporte de $U \cup V$ respecto a U es al menos c.

La confianza mínima c es siempre una fracción menor que 1 porque el soporte del conjunto $U \cup V$ es siempre menor que el de U.

Debido a que el primer paso de encontrar patrones frecuentes suele ser el más exigente computacionalmente, la mayor parte de la investigación en esta área se ha centrado en este primer paso. Sin embargo, algunos problemas computacionales y de modelado también surgen durante la fase de posprocesamiento, especialmente cuando el problema de minería de patrones frecuentes se utiliza en el contexto de otros problemas de minería de datos, como puede ser la clasificación (puede consultar la Sección 3.4 para más información).

3.3.1.3. Técnicas de minería de patrones frecuentes

Las técnicas para la minería de patrones frecuentes comenzaron con los métodos basados en la unión de **tipo Apriori** [3]. En estos algoritmos se realiza una exploración en anchura, donde los conjuntos de ítems candidatos son generados en un orden creciente del tamaño del conjunto de ítems. A continuación, estos conjuntos de elementos se prueban contra la base de datos de transacciones subyacente y aquellos ítems frecuentes que satisfacen la restricción de soporte mínimo se conservan para una exploración posterior. Aunque también encontramos deficiencias en estos algoritmos del tipo Apriori. Así, para usar la técnica Apriori se necesitan generar conjuntos de ítems candidatos. Estos conjuntos de elementos pueden ser grandes en número si el conjunto de ítems en la base de datos es enorme, y Apriori necesita múltiples exploraciones de la base de datos para comprobar el soporte de cada conjunto de elementos generado, conduciendo a altos costos.

Estas deficiencias se pueden superar utilizando un **Algoritmo de crecimiento de patrones frecuentes** (en inglés, *Frequent Pattern Growth*, **FP growth** [52]). Este algoritmo es una mejora del método Apriori, produciendo patrones frecuentes sin la necesidad de generar candidatos. El algoritmo FP growth representa la base de datos en forma de árbol, denominada

árbol de patrones frecuentes o árbol FP. Esta estructura de árbol mantendrá la asociación entre los conjuntos de ítems. La base de datos se fragmenta utilizando un elemento frecuente y se analizan los conjuntos de elementos de estos patrones fragmentados. Por lo tanto, con este método, la búsqueda de conjuntos de ítems frecuentes se reduce comparativamente.

Otro enfoque se puede encontrar con el **Algoritmo Eclat** (*Equivalence Class Transformation*) [131], el cual es un algoritmo basado en la búsqueda en profundidad que utiliza un diseño de *base de datos vertical*. Requiere menos espacio que Apriori si los conjuntos de ítems son pequeños en número. Es adecuado para conjuntos de datos pequeños y requiere menos tiempo para la generación de patrones frecuentes que con Apriori.

3.3.2. Minería de patrones secuenciales

Un problema relacionado con la minería de patrones frecuentes es la minería de patrones secuenciales, en la que hay un orden en las transacciones. Un orden temporal es bastante natural en muchos escenarios, como el comportamiento en la compra de un cliente, donde los artículos se compran en marcas de tiempo específicas, y a menudo siguen un orden temporal natural. En estos casos, el problema se redefine al de la minería de patrones secuenciales, en la que es deseable determinar secuencias relevantes y frecuentes de ítems. Ejemplos de secuencias incluyen desde transacciones de compras de clientes minoristas hasta secuencias biológicas, tratamientos médicos, compra venta de acciones, así como muchos otros.

Una base de datos secuencial se basa en ítems o eventos ordenados, registrados con o sin una noción concreta de tiempo. El propósito de la minería de patrones secuenciales es detectar subsecuencias interesantes en una base de datos secuencial, es decir, detectar relaciones secuenciales entre elementos que son de interés para el usuario. Se pueden utilizar varias medidas para estimar lo interesante que es una subsecuencia. En el problema original de minería de patrones secuenciales, se utiliza como medida el soporte.

El **soporte** de una secuencia s en una base de datos secuencial se define como el número de secuencias que contienen s, y se denota por $soporte(s)$.

La minería de patrones secuenciales es la tarea de encontrar todas las subsecuencias frecuentes en una base de datos secuencial. Se dice que una secuencia s es una secuencia frecuente o un patrón secuencial si y sólo si $soporte(s) \geq \delta$, para un umbral δ establecido por el usuario. La suposición es que las subsecuencias frecuentes son de interés para el usuario.

La minería de patrones secuenciales, la minería de eventos frecuentes ordenados o subsecuencias como patrones, fue introducida por primera vez por Agrawal y Srikant [4] (1995) y se ha convertido en un importante problema de la minería de datos.

Desde entonces, se ha dedicado abundante literatura a esta investigación y se han hecho

progresos colosales. Las mejoras en los algoritmos de minería de patrones secuenciales han seguido una tendencia similar al área relacionada con la minería de patrones frecuentes y han estado motivadas por la necesidad de procesar más datos a una velocidad más rápida con un menor costo.

3.3.2.1. Clasificación de los algoritmos de minería de patrones secuenciales

Por lo general, según [2], los algoritmos de minería de patrones secuenciales, al igual que los algoritmos de minería de patrones frecuentes, se pueden clasificar en tres clases principales: (1) enfoques basados en Apriori como GSP, (2) inspirados en ECLAT, con un diseño de base de datos vertical, como SPADE y (3) algoritmos de crecimiento de patrones frecuentes (FP-Growth) como FreeSpan o PrefixSpan.

La primera clase de algoritmos (es decir, los *enfoques basados en Apriori*) forman la gran mayoría de los algoritmos propuestos en la literatura para la minería de patrones secuenciales. Dependen principalmente de la propiedad Apriori, que establece el hecho de que cualquier superpatrón de un patrón infrecuente no puede ser frecuente, y se basan en una generación de candidatos y en un paradigma de pruebas propuesto en la minería de reglas de asociación [5]. La minería de patrones secuenciales generalizados (en inglés, Generalized Sequential Pattern Mining , GSP) [109] fue el primer algoritmo para la minería de patrones secuenciales propuesto mediante un enfoque basado en Apriori. Por otra parte, SPADE [132] es un algoritmo alternativo que utiliza una búsqueda en profundidad primero, y evita algunos de los inconvenientes del algoritmo GSP. Además, utiliza una representación de *base de datos vertical* en lugar de una representación horizontal de la base de datos utilizada por los algoritmos basados en Apriori.

Los algoritmos anteriores tienen la desventaja de generar repetidamente un gran número de secuencias candidatas y escanear la base de datos para mantener la información del cómputo del soporte de estas secuencias durante cada iteración del algoritmo, lo que las hace costosas desde un punto de vista computacional.

Para superar estos problemas, el *enfoque de crecimiento de patrones* para la minería de patrones secuenciales adopta un paradigma de crecimiento de patrones del estilo divide y vencerás [52], donde las bases de datos secuenciales se proyectan recursivamente en un conjunto de bases de datos más pequeñas basadas en los patrones secuenciales actuales, y los patrones secuenciales son creados en cada base de datos proyectada explorando solo fragmentos frecuentes locales. El paradigma de crecimiento de patrones frecuentes elimina la necesidad de la generación de candidatos y los pasos de poda que se producen en los algoritmos basados en Apriori y restringe repetidamente el espacio de búsqueda dividiendo

una base de datos secuencial en un conjunto de bases de datos proyectadas más pequeñas, que se extraen por separado. El algoritmo de crecimiento de patrones más popular para la minería de patrones secuenciales es PrefixSpan [98]. PrefixSpan muestra normalmente un mejor rendimiento que GSP y SPADE, pero, cuando se trata con bases de datos densas, el rendimiento de PrefixSpan puede ser peor que el de SPADE.

Referimos al lector a consultar [43] y [36] para más información general sobre la minería de patones secuenciales.

3.3.2.2. Definición formal del problema

Ahora, definamos formalmente el problema de la minería de patrones secuenciales. Sea $I = \{i_1, i_2, ..., i_k\}$ un conjunto con k ítems. Un conjunto de ítems t es un subconjunto no vacío de I, ($t \subseteq I$). Una secuencia $\alpha = \langle t_1, t_2, ..., t_m \rangle$ es una lista ordenada de conjuntos de ítems (también llamados eventos). Los ítems dentro de una secuencia no están ordenados y los enumeraremos alfabéticamente. Un elemento puede producirse como máximo una vez en un conjunto de ítems, pero puede producirse varias veces en diferentes conjuntos de ítems de una secuencia. Una secuencia se considera multivariante cuando contiene múltiples atributos en cada ítem de la secuencia.

El número de instancias de ítems de una secuencia se denomina longitud de la secuencia. Una secuencia con longitud k se denomina una k-secuencia. Por ejemplo, $s = \langle a, ce, bd, bcde, f, dg \rangle$ es una secuencia que consta de 7 ítems distintos $\{a, b, c, d, e, f, g\}$ y 6 conjuntos de ítems. La longitud de la secuencia es 12 ítems.

Cada conjunto de ítems de una secuencia representa el conjunto de eventos que se producen al mismo tiempo (con la misma marca de tiempo). Diferentes conjuntos de ítems aparecerán en un momento diferente.

Una secuencia $\alpha = \langle a_1, a_2, \ldots, a_n \rangle$ es una subsecuencia de la secuencia $\beta = \langle b_1, b_2, ..., b_m \rangle$ (o β es una supersecuencia de la secuencia α), denotada como $\alpha \preceq \beta$, si existen enteros $i_1 < i_2 < \ldots < i_n$ tales que $a_1 \subseteq b_{i1}, a_2 \subseteq b_{i2}, \ldots, a_n \subseteq b_{in}$. Por ejemplo, $\langle a, bd, f \rangle$ es una subsecuencia de s.

Un grupo de secuencias almacenadas con sus identificadores se denomina una base de datos secuencial.

El propósito de la minería de patrones secuenciales es descubrir subsecuencias interesantes en una base de datos secuencial, es decir, relaciones secuenciales entre los ítems que son interesantes para el usuario. Se pueden utilizar varias medidas para estimar lo interesante que es una subsecuencia. En el problema original de la minería de patrones secuenciales se utiliza como medida el soporte.

3.3.2.3. Representación de los patrones temporales

La representación temporal de los patrones se lleva a cabo principalmente utilizando la representación con punto de tiempo o la representación con intervalo de tiempo.

En la representación con intervalo de tiempo, hay diferentes maneras de relacionar los intervalos entre sí, de las cuales las más conocidas son el álgebra de intervalos de Allen [8] o la representación de conocimiento de la serie temporal. En el álgebra de intervalos de Allen existen trece relaciones que configuran un lenguaje muy expresivo, lo que hace que sean mucho más complicadas las tareas relacionadas con el razonamiento temporal y la representación del patrón.

Los datos basados en puntos de tiempo son un caso especial de los datos basados en intervalos de tiempo, en los que los puntos inicial y final se producen al mismo tiempo (para cada intervalo) y las relaciones entre estos puntos se vuelven más simples (anterior, igual o coexiste y posterior), normalmente denotadas como $(<, =, >)$. Además, dado que el operador "posterior" $(>)$ es el inverso de la relación "anterior" $(<)$, si siempre consideramos una relación desde el punto que ocurre primero, no es necesario utilizar la relación "posterior". Por ejemplo, si tenemos $a > b$, en su lugar diremos $b < a$.

Por lo tanto, es posible definir patrones o secuencias con solo estas dos relaciones $(<, =)$. Dos patrones α y β son exactamente iguales si sus puntos son exactamente los mismos y tienen exactamente las mismas relaciones en las mismas posiciones, es decir, $\alpha \preceq \beta$ y $\beta \preceq \alpha$.

En esta tesis hemos utilizado el algoritmo FaSPIP [50] para descubrir patrones secuenciales multivariantes. FaSPIP se basa en la estrategia de clases de equivalencia y es capaz de extraer tanto puntos como intervalos, aunque en esta tesis utilizaremos únicamente la representación mediante puntos de tiempo. Además, FaSPIP utiliza un nuevo algoritmo de generación de candidatos basado en puntos de límite y métodos eficientes para evitar la generación de candidatos inútiles y comprobar su frecuencia.

En la generación de candidatos, FaSPIP distingue entre dos operaciones para extender una secuencia con un elemento, creando así una nueva secuencia: Extensiones de secuencia (S-extensiones), cuando los puntos frecuentes tienen lugar después, y extensiones de ítems (I-extensiones), cuando los puntos tienen lugar al mismo tiempo que el último elemento en el patrón. Por ejemplo, dada la secuencia $\alpha = \langle a < b \rangle$ y un punto $c \in I$, la secuencia $\beta = \langle a < b < c \rangle$ es una S-extensión y $\gamma = \langle a < b = c \rangle$ es una I-extensión.

3.4. Clasificación

En el análisis de datos, la tarea de la clasificación consiste en construir un modelo o clasificador para predecir etiquetas categóricas (el atributo de etiqueta de la clase). La clasificación tiene numerosas aplicaciones, incluyendo la detección de fraude, el marketing objetivo o el diagnóstico médico. Así, por ejemplo, podemos construir un modelo de clasificación para clasificar las solicitudes de préstamos bancarios con las etiquetas "seguro" o "arriesgado", o para un investigador médico que quiere analizar los datos del cáncer de mama y predecir cuál de los tres tratamientos específicos debe recibir un paciente, asignando las etiquetas "tratamiento A", "tratamiento B" o "tratamiento C".

Muchos métodos de clasificación han sido propuestos por investigadores en aprendizaje automático, reconocimiento de patrones y estadística.

3.4.1. Proceso de dos fases: aprendizaje y clasificación

En general, la clasificación de datos incluye el siguiente proceso de dos fases: una fase de aprendizaje (donde se selecciona, optimiza y construye un modelo de clasificación) y una fase de clasificación (donde el modelo se usa para predecir las etiquetas de clase para determinados datos).

En la **fase de aprendizaje** se crea un clasificador que describe un conjunto predeterminado de clases de datos. En esta etapa de entrenamiento, un algoritmo de clasificación crea un clasificador mediante el análisis o el aprendizaje de un conjunto de entrenamiento compuesto por tuplas de base de datos y sus etiquetas de clase asociadas. Se supone que cada tupla pertenece a una clase predefinida que se denomina atributo de etiqueta de clase. Normalmente, esta asignación se representa en forma de reglas de clasificación, árboles de decisión o fórmulas matemáticas.

Dado que se proporciona la etiqueta de clase en cada tupla de entrenamiento, este paso también se conoce como aprendizaje supervisado. Contrasta con el aprendizaje no supervisado (o clustering), en el que no se conoce la etiqueta de clase de cada tupla de entrenamiento, y el número o conjunto de clases que se deben aprender puede no ser conocido de antemano.

En la **fase de clasificación**, el modelo creado anteriormente se utiliza para la clasificación.

La precisión de un clasificador en un conjunto de pruebas determinado es el porcentaje de tuplas del conjunto de pruebas que el clasificador clasifica correctamente. Si tuviéramos que usar solo el conjunto de entrenamiento para medir la precisión del clasificador, esta estimación probablemente sería optimista, porque el clasificador tiende a sobreajustar los

datos. Por lo tanto, se debe usar un conjunto de pruebas compuesto por tuplas de prueba y sus correspondientes etiquetas de clase asociadas. La etiqueta de clase asociada a cada tupla de prueba se compara con la clase aprendida utilizando la predicción del clasificador para esa tupla. Si la precisión del clasificador se considera aceptable, el clasificador se puede utilizar para clasificar futuras tuplas de datos para las que no se conoce la etiqueta de clase.

3.4.2. Integración de la minería de patrones y la clasificación: Clasificación basada en patrones

En los últimos años cada vez se estudia en mayor profundidad la combinación de la minería de patrones frecuentes y la clasificación. La idea principal con estos métodos es que los patrones se pueden utilizar para definir características o como reglas y la clasificación hará uso de estas características o reglas, siendo más precisa y fácil de entender.

Con la clasificación logramos una mejor comprensión de los datos y un análisis de datos más potente, realizándose normalmente después de la minería de patrones frecuentes en un paso intermedio. Cuando la minería de patrones frecuentes se utiliza al mismo tiempo como un paso para la extracción de reglas para usarlas en la clasificación, a menudo se conoce como *clasificación basada en patrones*.

Un ejemplo sencillo se proporciona en la Figura 3.2.

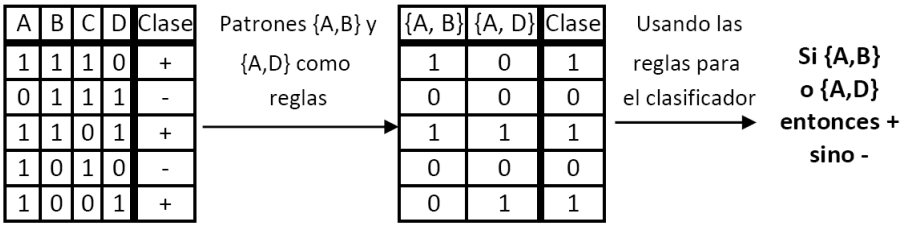

Figura 3.2: Ejemplo del proceso de clasificación basada en patrones

Esencialmente, un patrón frecuente es un conjunto de ítems que se observa en una serie de ejemplos. En este ejemplo, {A, B} es un patrón que se produce en el primer y tercer ejemplo y {A, D} en el tercer y quinto ejemplo. Una clasificación basada en este ejemplo podría predecir que el ejemplo es positivo, incluyendo solamente los ítems A, B y D.

En minería de datos, cuando se construye el conjunto de patrones, se puede hacer de dos maneras [17]:

- **Posprocesamiento**: Podemos ejecutar un algoritmo de minería de patrones una vez para encontrar un gran número de patrones, y posprocesar su resultado para obtener un conjunto más pequeño.

- **Minado Iterativo**: Podemos ejecutar iterativamente un algoritmo de minería de patrones, encontrando en cada ronda un número muy pequeño de patrones (a menudo sólo uno), teniendo en cuenta los patrones anteriores en cada ronda.

Los enfoques iniciales que combinaban los modelos de clasificación y la minería de patrones adoptaron un **enfoque estricto paso a paso**, en el que un conjunto de patrones se calcula una vez y estos patrones (*posprocesamiento del conjunto de patrones*) se utilizan posteriormente en los modelos de clasificación. Sin embargo, últimamente se ha propuesto un gran número de métodos que tienen como objetivo **integrar la minería de patrones, la selección de características y la construcción de modelos** (mediante *minería iterativa*) [17].

Esta integración se realiza centrándose en un subconjunto especial de reglas de asociación que siempre tienen una etiqueta de clase como su consecuente. Nos referimos a este subconjunto de reglas como las Reglas de la Asociación de Clases (Class Association Rules, CAR). El siguiente paso es usar estas reglas (*patrón* $\Rightarrow$ *etiqueta_de_clase*) para crear un clasificador seleccionando aquellas reglas más adecuadas para clasificar nuevos registros de datos. Muchos clasificadores basados en patrones se han propuesto mediante la adopción de algoritmos de minería de reglas de asociación eficientes basados entre otros por los métodos de Apriori [3] o FP-growth [52].

Un método bien conocido de clasificación de reglas de asociación, la Clasificación Basada en Asociaciones (Classification Based on Associations, CBA) fue propuesto por Liu et al. [77] e implementa el algoritmo Apriori [3] con el fin de generar CAR. Una versión mejorada del método, CBA2 [78], resolvió el problema de los conjuntos de datos no balanceados mediante el uso de varios valores mínimos de soporte de clase (puede consultar la Sección 3.4.5 para ampliar más información sobre el aprendizaje de clases poco frecuentes).

En otro trabajo, Li et al. [74] propusieron la "Clasificación basada en Reglas de Asociación Múltiple" (Classification based on Multiple Association Rules, CMAR) extendiendo el algoritmo FP-growth para minar grandes bases de datos de manera más eficiente que CBA. En CMAR, se emplean varias reglas en lugar de una sola regla para evitar el sobreajuste inherente a CBA. Además, la clasificación del conjunto de reglas establecida en CMAR se basa en el Chi-cuadrado ponderado de cada regla, reemplazando la confianza y el soporte de cada regla utilizada en CBA.

Yin y Han [129] propusieron CPAR (Classification based on Predictive Association Rules), que funciona mucho más rápido tanto en la generación de reglas como en la clasificación, aunque su precisión es parecida a la de CBA y CMAR.

Muchos estudios experimentales han demostrado que la clasificación basada en la minería de reglas de asociación es un enfoque con un alto potencial que construye sistemas de

clasificación más predictivos y precisos que los métodos de clasificación tradicionales [92]. Además, muchas de las reglas encontradas por los métodos asociativos de clasificación no se pueden encontrar utilizando técnicas de clasificación tradicionales.

Numerosos algoritmos se han derivado de los enfoques explicados anteriormente, con nuevas versiones y mejoras. El lector puede encontrar más información en [17, 92].

3.4.3. Clasificación de patrones secuenciales

La clasificación de secuencias tiene una amplia gama de aplicaciones en el mundo real. Por ejemplo, en la informática médica, la clasificación de las series temporales de ECG (la serie temporal de las frecuencias cardíacas) indica si los datos provienen de una persona sana o provienen de un paciente con enfermedad cardíaca [124], o en los sistemas financieros, la clasificación de los datos de las secuencias de transacciones en un banco se puede utilizar para luchar contra el blanqueo de dinero [81].

Normalmente, una secuencia es una lista ordenada de eventos. Un evento se puede representar como un valor simbólico, un valor real numérico, un vector de valores reales o un tipo de datos complejo.

Si suponemos que un conjunto S de n secuencias, denotado por $s_1 ... s_n$, está disponible para la construcción del modelo de entrenamiento. Dado L como el conjunto de k clases de etiquetas, cada una de estas secuencias se anota con una etiqueta de clase extraída de $l_1 ... l_k$. Estos datos de entrenamiento se utilizan para construir un modelo C que puede predecir la etiqueta de secuencias de prueba desconocidas. Por lo tanto, la tarea de clasificación de secuencias consiste en aprender un clasificador de secuencias C, que es una función que asigna una secuencia s en una etiqueta de clase $l \in L$, escrito como, $C : s \rightarrow l, l \in L$.

Muchas técnicas de modelado, como los clasificadores de los vecinos más cercanos, los métodos basados en reglas y los métodos basados en grafos, son comunes a las series temporales y a la clasificación de secuencias discretas debido a la naturaleza temporal de los dos tipos de datos.

En la clasificación de secuencias convencional, cada secuencia está asociada a una sola etiqueta de clase y toda la secuencia está disponible para un clasificador antes de la clasificación. También hay otros escenarios de aplicación para la clasificación de secuencias, así, por ejemplo, para una secuencia de síntomas de un paciente durante un largo período de tiempo, la condición de salud del paciente puede cambiar.

3.4.3.1. Desafíos principales en la clasificación de secuencias

Existen tres desafíos principales en la clasificación de secuencias [128]. En primer lugar, la mayoría de los clasificadores, como los árboles de decisión y las redes neuronales, solo pueden tomar los datos de entrada como un vector de características. Sin embargo, no hay características explícitas en los datos de una secuencia. En segundo lugar, incluso con varios métodos de selección de características, podemos transformar una secuencia en un conjunto de características y la selección de estas características está lejos de ser trivial. La dimensionalidad del espacio de características para los datos de secuencia puede ser muy alta y su cálculo puede ser costoso. Por último, además de unos resultados de clasificación precisos, en algunas aplicaciones, es posible que también deseemos obtener un clasificador interpretable. Crear un clasificador de secuencias interpretable es difícil, ya que no hay características explícitas.

3.4.3.2. Métodos de clasificación de secuencias

Los métodos de clasificación convencionales, como las redes neuronales o los árboles de decisión, están diseñados para clasificar vectores de características. Una forma de resolver el problema de la clasificación de secuencias es transformar una secuencia en un vector de características a través de selecciones de estas características. De este modo, las secuencias se pueden clasificar mediante un método de clasificación convencional, como máquinas vectoriales de soporte (Support Vector Machines, SVM) o árboles de decisión. Sin embargo, hay otras maneras de clasificar secuencias. En general, los métodos de clasificación de secuencias se pueden dividir en tres grandes categorías [128]:

- **Clasificación basada en características**: Transforma una secuencia en un vector de características y, a continuación, aplica métodos de clasificación convencionales. La selección de las características juega un papel importante en este tipo de métodos.
- **Clasificación basada en la distancia de las secuencias**: La función de distancia que mide la similitud entre secuencias determina significativamente la calidad de la clasificación.
- **Clasificación basada en modelos**: Como el Modelo oculto de Markov (Hidden Markov Model, HMM) y otros modelos estadísticos para clasificar secuencias.

Existen varias investigaciones en las que construyen clasificadores de secuencias basados en patrones secuenciales frecuentes.

Lesh et al. [68] propusieron un algoritmo para la clasificación de secuencias utilizando patrones frecuentes como características en el clasificador. En su algoritmo, las subsecuencias se extraen y transforman en conjuntos de características. Después de la extracción de características, se pueden utilizar algoritmos de clasificación general como SVM, Naive Bayes o redes neuronales para realizar la clasificación. Su algoritmo es el *primer intento en la combinación de clasificación y minería de patrones secuenciales*.

Tseng y Lee [118] propusieron el algoritmo *Classify-By-Sequence (CBS)* para clasificar grandes conjuntos de datos de secuencias. La metodología principal del método CBS es la minería de patrones secuenciales clasificables (Classifiable Sequential Patterns, CSP) y, a continuación, asignar una puntuación al nuevo objeto de datos para cada clase mediante una función de puntuación. Propusieron una serie de funciones de puntuación alternativas y probaron su rendimiento. Los resultados mostraron que la longitud de un CSP es el mejor atributo para la puntuación de clasificación y, en general, CBS puede crear un clasificador más preciso usando un soporte mínimo más bajo.

Exarchos et al. [39] propusieron una metodología de dos etapas para la clasificación de secuencias basada en la minería y optimización de patrones secuenciales. En la primera etapa, se utiliza la minería de patrones secuenciales y se crea un modelo de clasificación de secuencias basado en los patrones secuenciales extraídos. A continuación, las ponderaciones se aplican tanto a los patrones secuenciales como a las clases. En la segunda etapa, las ponderaciones se ajustan con una técnica de optimización para lograr una precisión de clasificación óptima. La precisión de su algoritmo es mayor que la de CBS. Sin embargo, la optimización es un procedimiento que consume mucho tiempo.

En [133] las secuencias se utilizan para clasificar planes buenos y malos en los sistemas de producción. Otros autores utilizan técnicas de abstracción temporal [56], donde los datos originales se abstraen en secuencias de intervalos y, a continuación, se descubren reglas de asociación.

En cuanto a la Unidad de Cuidados Intensivos, pocos trabajos han abordado el problema de la predicción de supervivencia utilizando el aprendizaje automático o el análisis inteligente de datos [62].

El lector puede encontrar un breve estudio sobre la clasificación de secuencias en [128].

3.4.4. Medidas para la evaluación de la clasificación

La base para analizar el rendimiento de un clasificador es una matriz de confusión [63]. Una matriz de confusión describe qué tan bien un clasificador puede reconocer diferentes clases. Para n clases, la matriz de confusión es una tabla nxn, donde cada entrada (i, j) indica

	Test de referencia	
Test	Positivo real	Negativo real
Positivo previsto	TP	FP
Negativo previsto	FN	TN

Tabla 3.1: Matriz de confusión 2x2. Las abreviaturas TP, FP, FN y TN denotan el número de, respectivamente, verdaderos positivos (True Positives), falsos positivos (False Positives), falsos negativos (False Negatives) y verdaderos negativos (True Negatives).

el recuento de instancias de la clase *i* clasificadas como *j*. Esto significa que las instancias clasificadas correctamente están en la diagonal principal de la matriz de confusión.

La forma más común y sencilla de una matriz de confusión es una matriz de dos clases, tal y como se muestra en la Tabla 3.1.

Dadas dos clases (positivas o negativas), normalmente usamos una terminología especial que describe a los miembros de la matriz de confusión, por lo que los Verdaderos Positivos (True Positives, TP) son instancias positivas que se clasificaron correctamente, los Verdaderos Negativos (True Negatives, TN) también son instancias correctamente clasificadas, pero de la clase negativa. Por el contrario, los Falsos Positivos (False Positives, FP) son instancias incorrectamente clasificadas como positivas y los Falsos Negativos (False Negatives, FN) son instancias incorrectamente clasificadas como negativas.

Sensibilidad, especificidad y precisión son los términos que se utilizan comúnmente para medir el rendimiento de una prueba de clasificación. La **sensibilidad** (también denominada tasa de verdaderos positivos o exhaustividad) indica el porcentaje de instancias verdaderamente positivas que se clasificaron como positivas (consulte la Ecuación 3.1) y la **especificidad** mide el porcentaje de instancias verdaderamente negativas que se clasificaron como negativas (consulte la Ecuación 3.2). Mientras que la **precisión** mide qué tan bien predice la prueba ambas categorías. La precisión de un clasificador en un determinado conjunto es el porcentaje de instancias correctamente clasificadas. En general, podemos definir la precisión como la división del número de instancias correctamente clasificadas entre el número total de instancias.

$$sensibilidad = \frac{TP}{TP + FN} \tag{3.1}$$

$$especificidad = \frac{TN}{TN + FP} \tag{3.2}$$

Para el caso de 2 clases, la precisión estaría definida en la Ecuación 3.3.

$$precisión = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.3}$$

La precisión es la medida más utilizada, pero produce resultados no esperados en ciertos casos, como cuando las clases están significativamente desbalanceadas. Un buen ejemplo lo tenemos con datos médicos cuando tratamos con una enfermedad rara, donde por lo general hay una mayoría de casos negativos (por ejemplo, 99 %) y sólo unos pocos (1 %) casos positivos. En este caso se podría conseguir una precisión impresionante simplemente clasificando todas las instancias como negativas, lo que es absolutamente inaceptable para fines médicos. Por lo tanto, es interesante utilizar alternativas para la medida de la precisión, como podría ser la sensibilidad o la especificidad, entre muchos otros.

3.4.5. Aprendizaje en bases de datos desbalanceadas

Como acabamos de ver, la distribución de clases en muchas aplicaciones no se encuentra normalmente balanceada. Otro ejemplo de esta característica se encontraría en un escenario de detección de fraude en el que existen registros de datos que representan la actividad de tarjetas de crédito, y estos registros se deben etiquetar como "normales" o "fraudulentos". Por lo general, esta distribución de clases estaría muy desbalanceada. Así, en este escenario será habitual encontrar más del 98 % de los puntos de datos con normalidad, mientras que sólo menos del 2 % de los puntos de datos pueden ser fraudulentos. La aplicación directa de algoritmos de clasificación puede dar lugar a resultados engañosos debido a la preponderancia de la clase normal. En este escenario de detección de fraude, y en muchos otros, queremos lograr una alta precisión de clasificación en los casos raros. El problema es que en este ejemplo un clasificador bayesiano tendrá antecedentes sesgados que favorecen a la clase normal y con un clasificador que utilice un árbol de decisión le resultará difícil separar las instancias que pertenecen a la clase rara. Como resultado, la mayoría de los clasificadores, si no se modifican adecuadamente, clasificarán muchas instancias raras en la clase mayoritaria.

La clasificación de casos que ocurren en raras ocasiones es un problema desafiante en muchas aplicaciones de la vida real. Implica el procesamiento de un conjunto de datos desbalanceados para distinguir los casos que ocurren en raras ocasiones respecto a los otros casos abrumadores. Afortunadamente, estos objetivos se pueden lograr realizando cambios en los algoritmos de clasificación existentes. En [63] los autores dividen estos cambios en tres grupos principales de acuerdo con la estrategia de su uso: *métodos a nivel de algoritmo, métodos a nivel de datos y métodos combinados*. Los *métodos a nivel de algoritmo* modifican el clasificador o el proceso de aprendizaje para resolver el desbalanceo. Los *métodos a nivel de datos* se utilizan en el preprocesamiento y suelen utilizar varias formas de remuestreo. La última estrategia se basa en la **combinación de varios métodos** para aumentar el rendimiento.

El objetivo de los **métodos a nivel de algoritmo** es modificar un clasificador o un proceso de aprendizaje en lugar de cambiar la distribución del conjunto de datos, descartando o replicando instancias. Estos métodos se basan principalmente en el sobreponderamiento de la clase minoritaria, discriminando la clase mayoritaria y penalizando la clasificación errónea o sesgando el algoritmo de aprendizaje.

Otro enfoque para tratar con conjuntos de datos desbalanceados son los **métodos a nivel de datos** donde se cambian las distribuciones de clases aumentando el número de instancias de la clase minoritaria (sobremuestreo), disminuyendo el número de instancias de la clase mayoritaria (submuestreo), mediante combinaciones de estos métodos o utilizando otras formas avanzadas de muestreo. El modelo de clasificación se aprende en los datos remuestrados. Generalmente se ha observado que el submuestreo de la clase mayoritaria tiene una serie de ventajas sobre el sobremuestreo de la clase rara. Cuando se utiliza el submuestreo, los datos de entrenamiento muestreados son mucho más pequeños que el conjunto de datos original, lo que conduce a una mejor eficiencia de entrenamiento [1]. En algunas variaciones, todas las instancias de la clase rara se utilizan en combinación con una pequeña muestra de la clase mayoritaria. Esto también se conoce como selección unilateral. La lógica de este enfoque es que las instancias de clase raras son demasiado valiosas como datos de entrenamiento para modificar cualquier tipo de muestreo.

3.5. Calidad e interés de los patrones

Los algoritmos de minería frecuente de conjuntos de ítems a menudo descubren un gran número de patrones. Si el umbral de calidad mínimo se establece demasiado bajo, el número de patrones descubiertos suele ser enorme y los patrones son muy redundantes. En este caso, suele ser muy difícil y lento para los seres humanos analizar los patrones encontrados. En consecuencia, se ha invertido mucho esfuerzo de investigación para mejorar la calidad de los patrones descubiertos. A continuación, daremos una visión general de cómo encontrar patrones significativos y cómo aplicar medidas de interés y otras técnicas de reducción de redundancia que se han propuesto con este fin.

3.5.1. Representaciones comprimidas de patrones frecuentes

Una primera limitación importante del problema tradicional de minería de patrones frecuentes es el gran número de conjuntos de ítems que pueden ser encontrados por los algoritmos. Este número dependerá de las características de la base de datos y del valor del umbral de soporte mínimo establecido por los usuarios. Asimismo, a medida que se encuentran más

patrones, el rendimiento de los algoritmos suele disminuir en términos de memoria y tiempo de ejecución. Encontrar demasiados patrones es un problema porque los usuarios normalmente no tienen mucho tiempo para analizar una gran cantidad de patrones. Para presentar un pequeño conjunto de patrones significativos al usuario, los investigadores han diseñado algoritmos para extraer representaciones comprimidas de conjuntos de ítems de alta utilidad. Una representación comprimida es un subconjunto de todos los conjuntos de ítems que es significativo y resume todo el conjunto de patrones frecuentes. Una ventaja de las representaciones comprimidas es que pueden ser varias órdenes de magnitud más pequeñas que el conjunto de ítems originales. Además, si se realiza directamente una minería de representaciones comprimidas puede ser mucho más rápida que descubrir todos los conjuntos de ítems mediante algoritmos de minería de patrones tradicionales. Asimismo, algunas de estas representaciones comprimidas también proporcionan una mayor precisión para las tareas de clasificación en comparación con el uso de todos los patrones frecuentes. Se han propuesto varias representaciones comprimidas de patrones frecuentes y se han diseñado varios algoritmos para descubrir directamente estas representaciones sin que sea necesario extraer todos los patrones frecuentes.

Existen diferentes tipos de representaciones compactas en la minería de patrones frecuentes que conservan diferentes niveles de conocimiento acerca del conjunto original de patrones frecuentes y sus valores de soporte [1]. Las representaciones más conocidas son los conjuntos de ítems frecuentes maximales (maximal frequent itemset) y los conjuntos de ítems frecuentes cerrados (closed frequent itemsets), aunque existen otras representaciones aproximadas. Todas estas representaciones varían en el grado de pérdida de información que presentan.

La representación comprimida de conjuntos de ítems se puede hallar de dos maneras. Así, los algoritmos se pueden diseñar para encontrar directamente la representación comprimida durante el proceso de detección de patrones frecuentes o bien se puede establecer el conjunto de ítems frecuentes con un nivel de soporte mínimo determinado, y la representación comprimida se puede derivar de este conjunto.

A continuación veremos los diferentes tipos de representación comprimida de manera más detallada.

Definición 1 *Dada una colección S de conjuntos de ítems con un nivel de soporte mínimo dado, se dice que un conjunto de ítems $M \in S$ es un* ***Conjunto Maximal de ítems Frecuentes (Maximal Frequent Itemset)*** *en S si no hay otro superconjunto frecuente de M en S.*

Los conjuntos maximales de ítems frecuentes no se pueden ampliar sin que su soporte caiga por debajo del umbral. Por lo tanto, esta representación comprimida solo informa del

patrón frecuente más largo de cada rama. La minería de patrones maximales fue estudiado por primera vez por Bayardo [14], donde se propuso MaxMiner, un método de búsqueda basado en Apriori, que utilizando una búsqueda primero en amplitud lograba encontrar conjuntos de ítems maximales mediante la realización de la poda de frecuencia del superconjunto y la poda de infrecuencia de los subconjuntos para la reducción del espacio de búsqueda.

Las representaciones maximales tienen pérdidas respecto al soporte, en cambio, no tienen pérdidas respecto a la pertenencia de los conjuntos de ítems. Aunque todos los conjuntos de ítems se pueden generar a partir de los conjuntos de ítems maximales utilizando el enfoque de creación de subconjuntos, no se pueden llegar a obtener sus valores de soporte. Por lo tanto, los conjuntos de ítems maximales tienen pérdida ya que no conservan información sobre los valores de soporte. Para proporcionar una representación sin pérdidas en términos de los valores de soporte, se utiliza la noción de minería de conjuntos de ítems cerrados.

Definición 2 *Dada una colección S de conjuntos de ítems con un nivel de soporte mínimo dado, se dice que un conjunto de ítems $C \in S$ es un* ***Conjunto Cerrado de ítems Frecuentes (Closed Frequent Itemset)*** *en S si no hay otro superconjunto frecuente de C en S con el mismo soporte.*

La minería de los conjuntos cerrados de ítems frecuentes fue propuesta por Pasquier et al. en [97], donde se propuso un algoritmo basado en Apriori llamado A-Close para dicha minería. Las representaciones cerradas son totalmente sin pérdidas con respecto al soporte y a la pertenencia de los conjuntos de ítems. La búsqueda de patrones cerrados proporciona dos beneficios al mismo tiempo: una reducción en el número de candidatos, y una salida más compacta manteniendo la máxima cantidad de información.

Los patrones cerrados son una forma particularmente interesante de representación comprimida. Un conjunto de ítems C está configurado para cerrarse si ninguno de sus superconjuntos tiene el mismo soporte que C. Por lo tanto, al determinar todos los patrones frecuentes cerrados, se puede obtener no sólo el conjunto exhaustivo de conjuntos de ítems frecuentes, sino también sus soportes. Según hemos visto anteriormente, los valores de soporte se pierden mediante la minería de patrones maximales. En otras palabras, el conjunto de patrones maximales no se puede utilizar para derivar los valores de soporte de subconjuntos que faltan. Sin embargo, los valores de soporte de conjuntos de ítems frecuentes cerrados se pueden usar para derivar los valores de soporte de los subconjuntos que faltan.

Definición 3 *Dada una colección S de conjuntos de ítems con un nivel de soporte mínimo dado, se dice que un conjunto de ítems $N \in S$ es un* ***Conjunto Minimal de ítems Frecuentes (Minimal Frequent Itemset)*** *en S si no hay otro subconjunto frecuente de N en S.*

Es fácil ver que el conjunto minimal de ítems es el más general de la colección. Esta representación fue propuesta por Fan [40] y es lo opuesto al conjunto maximal de ítems frecuentes, es decir, consiste en los conjuntos más pequeños de ítems con la suposición de que a menudo son los más interesantes. Por ejemplo, con fines de marketing, un minorista puede estar más interesado en encontrar los conjuntos más pequeños de artículos que generan un alto beneficio, ya que es más fácil promocionar un pequeño conjunto de artículos dirigidos a muchos clientes en lugar de un gran conjunto de artículos dirigidos a pocos clientes. Esta representación tiene pérdida con respecto al soporte y con respecto a la pertenencia a los conjuntos de elementos, pero a menudo proporciona la mejor alternativa práctica en escenarios basados en aplicaciones.

3.5.2. Medidas de interés para la minería de patrones

La minería de patrones frecuentes se basa en la suposición de que los patrones frecuentes son interesantes. Pero esta suposición no se mantiene para numerosas aplicaciones. Por ejemplo, en una base de datos con transacciones de un supermercado, el patrón $\{pan, leche\}$ puede ser muy frecuente, pero al mismo tiempo poco interesante, ya que representa un comportamiento de compra que es común, y puede producir un bajo beneficio. Por el contrario, existen patrones tales como $\{champan, caviar\}$ que pueden no ser frecuentes, pero que podrían llegar a producir un mayor beneficio.

El uso del soporte para la minería de patrones frecuentes, en muchos casos, no es la medida ideal. Por un lado, elegir un soporte mínimo con un valor alto puede llevar a obtener sólo reglas que contengan conocimientos obvios y a que se pierdan casos excepcionales que son interesantes. Por otro lado, establecer un soporte mínimo con un valor bajo producirá un gran número de reglas que normalmente son redundantes o con ruido. Por lo tanto, un soporte mínimo es difícil de ajustar [59]. Para encontrar patrones interesantes en los datos se pueden considerar otros aspectos como el grado de interés, el beneficio o la utilidad.

3.5.2.1. Integración de las restricciones en la minería de patrones frecuentes

Para encontrar patrones más interesantes y reducir el número de patrones encontrados, los investigadores han propuesto integrar restricciones en la minería de patrones frecuentes. Una restricción es un conjunto adicional de criterios que el usuario proporciona para indicar con mayor precisión los tipos de patrones interesantes que se quieren encontrar. Numerosos tipos de restricciones se han utilizado ampliamente en el aprendizaje automático, existiendo generalmente dos maneras de aplicar estas restricciones para encontrar patrones interesantes [43]. La primera forma es aplicar estas restricciones como un paso de posprocesamiento en

el conjunto de todos los patrones frecuentes, para así filtrar patrones poco interesantes. Sin embargo, un problema con este enfoque es que enumerar todos los patrones frecuentes puede consumir mucho tiempo y requiere una gran cantidad de memoria. La segunda manera de abordar este problema es promocionando las restricciones profundamente en el proceso de minería. En otras palabras, las restricciones se aplican durante la búsqueda de patrones para lograr así reducir el espacio de búsqueda. Los algoritmos que adoptan este enfoque pueden ser órdenes de magnitud más rápidos y generar mucho menos patrones que los algoritmos de minería de patrones frecuentes tradicionales, dependiendo de las restricciones utilizadas.

3.5.2.2. Restricciones en la minería secuencial de patrones

En relación con la minería secuencial de patrones, uno de los primeros algoritmos en integrar restricciones fue GSP [109]. Este algoritmo introdujo las restricciones de tiempo mínimo y máximo entre dos conjuntos consecutivos de ítems en patrones secuenciales (restricciones de intervalo), así como una duración de tiempo máxima para cada patrón secuencial (restricción de duración). Otro tipo de restricciones consideradas en la minería de patrones secuenciales son las expresiones regulares (restricciones de expresión regular). El algoritmo SPIRIT [46] permite a los usuarios especificar expresiones regulares en los patrones que se quieren buscar. Convierte restricciones en un autómata para podar patrones al realizar una búsqueda de amplitud.

Varios investigadores han estudiado las características de las restricciones que se pueden situar profundamente en el proceso de minería de patrones secuenciales [43]. Se han identificado tres tipos principales de restricciones: *Anti-monótona, Convertible y Sucinta*. *Las restricciones anti-monótonas* que utilizan un umbral de soporte mínimo con restricciones de longitud, intervalo o duración son algunas de las más fáciles y ventajosas de integrar en un algoritmo de minería de patrones secuencial, ya que se pueden utilizar para podar el espacio de búsqueda aplicando la propiedad de cierre descendente. *Las restricciones convertibles* son restricciones que no son ni monótonas ni anti-monótonas, pero que se pueden convertir en restricciones anti-monótonas si se aplican algunas estrategias adicionales. Por último, *las restricciones sucintas* son restricciones que se pueden comprobar para un patrón mirando solo los elementos individuales que contiene. Por ejemplo, la restricción sobre la suma de los pesos de un patrón secuencial que no debe ser mayor o menor que un valor determinado se puede comprobar simplemente agregando los pesos de sus elementos.

En [55] los autores presentan una nueva técnica de minería de datos basada en el principio de la longitud mínima de la descripción (Minimum Description Length principle, MDLP), que descubre características interesantes en una secuencia ordenada en el tiempo.

En [99] los autores introducen un método con el que identificar de forma exacta y eficiente los k patrones más interesantes en una base de datos secuencial para la que la diferencia entre su frecuencia observada y la esperada es máxima: una medida denominada apalancamiento (leverage). Otros autores se centran en medidas para la selección de patrones, como el riesgo relativo o una medida de cobertura [69].

3.5.2.3. Medidas de interés de un patrón

Hay muchas medidas de interés ampliamente utilizadas como restricciones en el aprendizaje automático, la minería de datos y la estadística.

Sin embargo, todavía no hay una definición formal del interés (o grado de interés, en inglés "interestingness"). En un estudio, Geng y Hamilton [47] han reunido 9 criterios diferentes sobre el interés de un patrón. Estos 9 criterios son concisión, generalidad, fiabilidad, peculiaridad, diversidad, novedad, sorpresa, utilidad y accionabilidad.

Estos criterios pueden tener solapamientos o conflictos con otros. Por ejemplo, un patrón conciso, debido a su simplicidad, normalmente es general y la generalidad también puede conducir a la fiabilidad. Por otro lado, la generalidad está en conflicto con la peculiaridad y la novedad.

Además de los criterios mencionados que pueden definir el interés de un patrón, existen 3 categorías principales en las que se pueden clasificar las medidas del grado de interés: *objetivas, subjetivas y basadas en la semántica* [47].

Medidas objetivas son aquellas que dependen sólo de los datos sin procesar, no dependiendo de una aplicación o específicamente del usuario. Las *medidas subjetivas* son aquellas que tienen en cuenta el conocimiento de los usuarios, así como los datos. Como un tipo especial de medidas subjetivas, las *medidas basadas en la semántica* tienen en cuenta la explicación y la semántica de un patrón, y son, al igual que las medidas subjetivas, específicas del dominio [59].

Debido al gran número de patrones, muchos de los métodos de descubrimiento están basados en los k principales (top k) con varias medidas de interés para ordenar un subconjunto de patrones [73]. Un método "top k" solo selecciona un subconjunto de patrones con la más alta calidad hasta el número k. Los top k subpatrones todavía pueden contener patrones redundantes. Esto ha sido observado por [120].

Existen varios artículos donde los autores comparan diferentes medidas de interés.

Por ejemplo, Bayadro y Agrawel [15] han propuesto un algoritmo para extraer reglas optimizadas bajo un orden parcial de las reglas (en lugar del orden total típico en las reglas) de acuerdo con diferentes medidas de interés como soporte, confianza, convicción, ganancia,

ganancia de entropía, gini o chi-cuadrado entre otros.

Tan et al. [112] han introducido 21 medidas objetivas diferentes de reglas que se pueden utilizar para evaluar las reglas de asociación. Mostraron que dependiendo de sus propiedades, cada medida es útil para alguna aplicación, pero no para otras. También propusieron un enfoque para encontrar la mejor medida de interés para los patrones de un dominio específico. Para ello, en primer lugar, un especialista en dominios debe realizan manualmente un ranking de un conjunto de patrones en ese dominio. A continuación, el ranking más similar realizado usando diferentes medidas mostrará la mejor medida que se puede usar para ese dominio de aplicación específico. En los casos en los que hay un gran número de patrones, sólo aquellos que tienen una alta desviación estándar en la evaluación de las diferentes medidas serán elegidos como muestra que se presentará a los expertos del dominio.

En otro trabajo, Ohsaki et al. [95] han aplicado diferentes medidas de interés en la minería de reglas de asociación para examinar la utilidad de estas medidas para encontrar reglas interesantes extraídas de los datos clínicos. Otra solución es utilizar un enfoque de prueba de hipótesis para identificar o eliminar patrones redundantes mediante la evaluación por retención (holdout) [122].

Korn et al. propusieron las reglas de ratio [66], mediante el análisis del eigensystem para calcular correlaciones entre los valores de los atributos, lo que revela los ejes de mayor variación y, por lo tanto, las correlaciones más importantes.

Las reglas de ratio de Korn fueron desarrolladas aún más por Malone y McGarry añadiéndoles un elemento temporal en la forma de reglas de ratio diferencial (dFr) capaces de detectar patrones interesantes en datos espaciotemporales [83]. La técnica dFr incorpora todos los aspectos de los patrones espaciotemporales y se utilizó para detectar cambios en las imágenes digitalizadas de geles proteicos 2D dentro de una serie temporal. Las proteínas específicas fueron marcadas como interesantes y clasificadas de acuerdo con la cantidad en la que se habían alterado. Alteraciones particulares incluyen la ausencia/presencia de proteínas y variaciones morfológicas a lo largo del tiempo.

Después de mostrar que incluso las reglas de confianza pueden tener correlaciones negativas, Arunasalam y Chawla [10] propusieron una nueva medida denominada Complement Class Support (CCS) que garantiza que las reglas estén correlacionadas positivamente. Esta medida se basa en la propiedad anti-monótona del CCS, junto con el hecho de que las reglas “buenas” tienen valores CCS bajos, y descubre reglas sólidas mediante un algoritmo de enumeración de filas [29].

Además de los anteriores trabajos, referimos al lector a [86] y [47] para obtener información más general sobre las medidas de interés y [44] para patrones de alta utilidad o beneficio.

Medida	Fórmula
Soporte	$P(Ac)$
Confianza	$P(c\|A)$
Cobertura	$P(A)$
Prevalencia	$P(B)$
Especificidad	$P(\neg c\|\neg A)$
Precisión	$P(Ac) + P(\neg A \neg c)$
Razón de probabilidades (OR)	$\frac{P(Ac)P(\neg A \neg c)}{P(A \neg c)P(\neg Ac)}$
Riesgo relativo	$\frac{P(c\|A)}{P(c\|\neg A)}$

Tabla 3.2: Medidas clínicas habituales de interés para las reglas en la forma $A \rightarrow c$

3.5.2.4. Medidas objetivas de interés en ambientes clínicos

El soporte y la confianza son las medidas predeterminadas de interés utilizadas universalmente para descubrir las reglas de asociación relevantes. El marco de trabajo basado en soporte-confianza es el más utilizado en la mayoría de los métodos de minería de reglas de asociación y para la minería y selección de reglas para patrones discriminatorios [59].

Aunque el soporte y la confianza son medidas apropiadas para construir un modelo fuerte en muchos casos, no se pueden considerar como las medidas ideales.

Otros estudios utilizan pruebas estadísticas en el descubrimiento de patrones, como en [122] que utilizan la corrección de Bonferroni o la evaluación por retención (holdout), o algunas definiciones sintácticas para eliminar redundancias, por ejemplo, el cierre [120], las restricciones [16] o la relevancia [51].

En esta tesis nos centraremos en las medidas objetivas basadas en la probabilidad que normalmente se utilizan en el dominio clínico. Algunos ejemplos de medidas objetivas de interés de las reglas, que a menudo se utilizan en la epidemiología como métricas estadísticas, se muestran en la Tabla 3.2.

El riesgo relativo y la razón de probabilidades (odds ratio, OR) son métricas estadísticas que a menudo se utilizan en estudios epidemiológicos. Son consistentes: una razón de probabilidades más grande conduce a un riesgo relativo mayor, y viceversa. Bajo la suposición de enfermedad rara, la razón de probabilidades se aproxima al riesgo relativo [72]. La razón de probabilidades se utiliza generalmente en estudios de control de casos.

Li et al. [73] presentó un procedimiento para podar reglas redundantes basadas en la superposición del intervalo de confianza de la razón de probabilidades. La razón de probabilidades generalmente se informa con su intervalo de confianza para mostrar la precisión de la estimación. Li et al. utilizaron intervalos de confianza para determinar si una regla y su padre

son estadísticamente diferentes. Si los intervalos de confianza no se superponen, las reglas deben llevar información diferente, en otro caso, se consideran equivalentes y la subregla es podada.

En [72] y [71], los autores utilizan una métrica epidemiológica, el riesgo relativo, para medir el interés del patrón, y concluyen que es una medida óptima con la que encontrar patrones de alto riesgo. El método propuesto era más eficaz en lo que respecta a la cobertura del espacio de búsqueda y producía un número menor de reglas. Sin embargo, el número de reglas en la salida todavía podría ser demasiado grande para poder realizar una interpretación fácil. Los autores aplicaron el método a un conjunto de datos médicos y farmacéuticos vinculados del mundo real y reveló algunos patrones que son potencialmente útiles en la práctica clínica.

3.5.2.5. Medidas de interés para mejorar la clasificación

Todas las medidas objetivas de interés propuestas para reglas de asociación que se basan en probabilidades se pueden aplicar directamente a la evaluación de las reglas de clasificación, ya que sólo incluyen las probabilidades del antecedente de una regla, del consecuente, o de ambos, y representan la correlación, generalidad y fiabilidad entre el antecedente y el consecuente. Sin embargo, cuando estas medidas se utilizan de esta manera, evalúan el grado de interés de la regla con respecto a los datos dados (el conjunto de datos de entrenamiento), mientras que el enfoque clave en la minería de reglas de clasificación se centra en la precisión de la predicción [47].

Los autores de [12] utilizaron una prueba estadística derivada de la distribución binomial para evaluar el valor predictivo de las reglas de asociación obtenidas para crear un modelo de clasificación. Debido al alto número de reglas de asociación obtenidas a pesar de este procedimiento, los mismos autores [84] definieron más adelante un algoritmo voraz para evaluar los posibles subconjuntos de patrones predictivos mínimos.

El clasificador asociativo SPARCCC [121], fue introducido por Verhein y Chawla y utiliza el valor p de la prueba exacta de Fisher para extraer sólo reglas estadísticamente significativas. Estos autores también utilizan una nueva medida llamada Relación de Correlación de Clases (CCR) para seleccionar solo aquellas reglas que estén más positivamente correlacionadas con la clase que predicen en lugar de las otras clases. La clasificación se lleva a cabo mediante el uso de una puntuación de fuerza para clasificar las reglas. Esta puntuación es una combinación de confianza, el valor p y el CCR. Los autores muestran que SPARCCC puede superar a otros algoritmos cuando se utiliza un conjunto de datos desbalanceado para el entrenamiento.

Respecto a los patrones secuenciales, la mayoría de los algoritmos de clasificación convencionales basados en patrones secuenciales frecuentes siguen un enfoque estricto paso a paso (véase en la Sección 3.4 la clasificación basada en patrones). El primer paso consiste en extraer un conjunto completo de patrones secuenciales dado un soporte mínimo, mientras que el segundo consiste en seleccionar una serie de patrones discriminatorios con los que crear un clasificador [127]. En la mayoría de los casos, la minería de un conjunto completo de patrones secuenciales de una base de datos grande consume mucho tiempo, y el enorme número de patrones descubiertos indica que la selección de patrones y la creación del clasificador también consumirán mucho tiempo.

De hecho, la consideración más importante en la clasificación de secuencias no es la de encontrar el conjunto completo de reglas, sino la de descubrir los patrones más discriminatorios. A este respecto, recientemente se ha prestado más atención al descubrimiento discriminatorio de patrones frecuentes para una clasificación efectiva.

En [39], los autores proponen una técnica de optimización con la que ponderar los patrones secuenciales utilizados en la clasificación.

En el ámbito clínico, los episodios frecuentes univariados de subpuntuaciones SOFA (Sequential Organ Failure Assessment) durante los primeros días después de la admisión fueron identificados en [114]. A continuación, los autores seleccionaron un número reducido de patrones utilizando el Criterio de Información de Akaike para crear un modelo de regresión logística con el fin de predecir la supervivencia de los pacientes que sufren de fallos multiorgánicos. Más tarde, en [115] los mismos autores mostraron que el uso de patrones univariados como predictores es al menos tan eficaz como las puntuaciones clínicas.

En [48] después de extraer patrones emergentes de salto (JEP), el autor utiliza modelos ocultos de aprendizaje de Markov acoplados (Coupled Hidden Markov learning models, CHMM) para crear clasificadores secuenciales robustos basados en patrones. Esto hizo posible predecir el riesgo de hipotensión, un episodio hipotenso agudo (AHE) o incluso un shock séptico, utilizando las mediciones de la presión arterial media, la frecuencia cardíaca y la frecuencia respiratoria.

3.5.3. Minería de patrones discriminatorios

La minería de patrones discriminatorios es un grupo popular de técnicas de minería de patrones diseñadas para descubrir un conjunto de patrones significativos que se producen con frecuencias poco razonables en diferentes conjuntos de datos con clases etiquetadas [54].

Los patrones discriminatorios pueden proporcionar una comprensión importante en los conjuntos de datos con etiquetas de clase al contrastar las características de las clases dadas

e identificar diferencias humanamente interpretables. Además, los patrones discriminatorios generalmente revelan la distribución subyacente de los datos etiquetados y ayudan a desarrollar una solución a problemas complejos como la construcción de modelos de clasificación precisos. Sobre todo, el valor práctico de los patrones discriminatorios ha sido demostrado por una amplia gama de estudios en varios ámbitos, incluyendo la bioinformática, la ciencia médica, la gestión del marketing, etc.

3.5.3.1. Conjuntos de contraste, patrones emergentes y subgrupos

La investigación sobre patrones discriminatorios evoluciona rápidamente bajo varias definiciones no uniformes como los *conjuntos de contraste (contrast sets)* [13], *patrones emergentes (Emerging Patterns, EP)* [34] y *los subgrupos* [65, 125].

De acuerdo con [13, 123], la minería de *conjuntos de contraste* tiene como objetivo descubrir patrones que capturan diferencias de frecuencia excepcionales en diferentes grupos de sujetos definidos por el usuario.

La minería de *patrones emergentes* detecta patrones donde el crecimiento de frecuencia cambia de una clase a otra [34, 40]. El problema de la minería de patrones emergentes se puede expresar de la siguiente manera: Dadas dos clases de datos y un umbral de tasa de crecimiento α, encontrar todos los patrones (conjuntos de elementos) cuyas tasas de crecimiento - la relación de su frecuencia entre las dos clases - son mayores que el umbral dado [40]. Por lo tanto, un patrón X es emergente si $\frac{soporte2(X)}{soporte1(X)} \geq \alpha$.

Mientras que el descubrimiento de *subgrupos* intenta encontrar subgrupos de población que son estadísticamente más interesantes dada una población de individuos y una propiedad de aquellos individuos que estamos interesados [65, 125].

En general, aunque tales patrones se describen bajo diferentes nombres, casi son equivalentes, utilizando en esencia el poder discriminatorio entre las clases.

Al igual que otras reglas o patrones compuestos de combinaciones conjuntivas de elementos, los EP pueden ser fácilmente entendidos y utilizados directamente por los médicos.

Además, se ha propuesto el concepto de patrones emergentes de salto (Jumping Emerging Patterns, JEP) [35] para describir aquellas características discriminatorias que se producen sólo en los casos positivos de entrenamiento, y por lo tanto no se producen en absoluto en la clase negativa. Si $\frac{soporte2(X)}{soporte1(X)} = \infty$ entonces X es un JEP. Los JEP que aparecen con más frecuencia se han utilizado para crear clasificadores precisos ([70] [37]).

3.5.3.2. Exploración de patrones discriminatorios: procedimiento de dos pasos

La exploración de patrones discriminatorios generalmente incluye dos aspectos: la frecuencia y una restricción discriminatoria. Por un lado, la **frecuencia** de un patrón normalmente se estima por su soporte, que se define como el porcentaje de transacciones en una clase que contienen este patrón. Un patrón es frecuente si su valor de soporte no es menor que un umbral determinado. Por otro lado, la **restricción discriminatoria** generalmente utiliza un test estadístico como chi-cuadrado [13], razón de probabilidades (odds ratio) [47, 73] o la ganancia de información [25] para obtener patrones discriminatorios con una significación estadística. Por lo general, cualquier medida estadística que sea capaz de cuantificar las diferencias entre clases es utilizable.

En general, los algoritmos para la minería de patrones discriminatorios se pueden diferenciar en dos categorías de acuerdo con sus estrategias de ranking [54]: algoritmos con umbrales especificados por el usuario y algoritmos basados en las normas de la significancia estadística. Los algoritmos que entran en la primera categoría suelen establecer algunos umbrales alcanzables de medidas estadísticas para reducir el espacio de búsqueda y filtrar patrones insignificantes. Los algoritmos incluidos en la segunda categoría tienen como objetivo principal encontrar un cierto número de patrones que tengan el mayor poder discriminatorio.

La mayoría de estos algoritmos adoptan un **procedimiento de dos pasos**: primero generar un conjunto de patrones candidatos (es decir, patrones frecuentes en una clase); y entonces, aplicar la restricción discriminatoria (generalmente una prueba de significancia estadística) para medir su poder discriminatorio y podar patrones insignificantes. Sin embargo, debido a la gran cantidad de ítems que normalmente se generan en el primer paso, algunos métodos tratan de extraer directamente patrones discriminatorios en un solo paso, sin la necesidad de generar previamente los patrones candidatos.

Cuando se utilizan patrones discriminatorios, una pregunta importante es cómo seleccionar una restricción discriminatoria adecuada en algunas situaciones prácticas específicas. Además, [41] presenta una interesante formulación para dividir los patrones discriminatorios en varias categorías con respecto a sus diferentes tipos de poder discriminatorio. En particular, la eficacia de una medida de discriminación puede ser diferente con la distinción de los objetivos fijados, los tipos de datos y las categorías de patrones discriminatorios. Por lo tanto, la elección de medidas apropiadas para la evaluación del poder discriminatorio a veces necesita conocimientos del dominio y un claro reconocimiento de la naturaleza de los problemas y de los datos.

En un trabajo reciente [84], los autores utilizan un algoritmo voraz para evaluar los posibles subconjuntos de patrones predictivos mínimos (utilizados en [12]) con el objetivo de

definir un nuevo algoritmo para extraer patrones predictivos en lugar de utilizar un proceso de dos fases. Utilizaron una prueba estadística derivada de la distribución binomial para evaluar las reglas de asociación obtenidas como candidatas para ser covariables en el modelo de clasificación.

También se han propuesto métodos de análisis discriminatorio sobre datos secuenciales. Ji et al. [61] estudió en la minería de subsecuencias las características minimales que se producen frecuentemente en secuencias de una clase y con poca frecuencia en secuencias de la otra clase. Extraer información comprimida y de fuerte contraste entre dos conjuntos de datos secuenciales puede ser útil en la evolución clínica de los pacientes, o en la construcción de modelos de clasificación secuenciales. Por lo que estos autores diseñaron un algoritmo eficiente, llamado ConSGapMiner, para encontrar todas las subsecuencias distintivas generando primero candidatos, luego calculando su soporte de frecuencia, probando la satisfacción del hueco y finalmente usando un procesamiento posterior para eliminar todas las respuestas no minimales.

Los autores en [6] proponen un novedoso algoritmo evolutivo para la minería de los top-k patrones emergentes secuenciales, y de esta manera recomendar la secuencia de cursos más prometedora a los estudiantes en base a la secuencia de cursos ya realizada por estudiantes considerados excelentes.

Capítulo 4

Metodología para uso de patrones secuenciales con clasificación

Tal y como vimos en la Sección 1.2, la reanimación y la estabilización son problemas clave en las Unidades de Quemados Críticos y las predicciones de supervivencia temprana ayudan a decidir la mejor acción clínica durante estas fases. Las puntuaciones actuales de supervivencia por quemaduras se centran en variables clínicas como la edad o la superficie corporal quemada. Sin embargo, la evolución de otros parámetros (por ejemplo, la diuresis o el balance de fluidos) durante los primeros días también es un conocimiento valioso. En este capítulo, basado en el trabajo aceptado en AIME 2015 *(*15^{th} *Conference on Artificial Intelligence in Medicine)* [20], sugerimos una metodología con la que se realiza un proceso de minería temporal de datos para estimar la condición de supervivencia de la evolución del paciente.

4.1. Introducción

En las Unidades de Quemados Críticos se registra la evolución del paciente, pero en cambio no se considera esta evolución en las puntuaciones para la predicción de la mortalidad. Creemos que esto podría ser una mejora relevante para el conocimiento actual. Este conocimiento puede ser descubierto usando la minería de datos temporal (MDT), en la que se infieren asociaciones de proximidad contextual y temporal, algunas de las cuales también pueden indicar una asociación causa-efecto.

En este trabajo nos enfrentamos a tres desafíos principales: a) proporcionar a los médicos un entendimiento del conocimiento de la evolución del paciente, b) extraer una cantidad manejable de conocimiento, y c) generar modelos comprensibles que podrán ser interpreta-

Atributo	Mín.	Máx.	Media	Desviación estándar
Edad (años)	9	95	46.57	20.75
Peso (kg)	25	120	71.1	10.76
Duración de la estancia (días)	5	163	26.11	25.31
Superficie total quemada (%)	1	90	31.78	20.61
Superficie quemada profunda (%)	0	90	17.67	18.16
SAPS	6	58	20.76	9.57

Tabla 4.1: Resumen de atributos (379 pacientes).

dos por el experto. Con este fin, hemos llevado a cabo diferentes experimentos dentro de un proceso de descubrimiento de conocimiento de 4 pasos.

Los antecedentes de este tema concreto pueden verse en el Capítulo 3, concretamente en la Sección 3.3.2 sobre minería de patrones secuenciales, o en la Sección 3.4.3 sobre su clasificación. Además sería interesante leer la Sección 3.5 sobre la calidad e interés de los patrones, concretamente las secciones 3.5.1 sobre las representaciones comprimidas de patrones frecuentes y 3.5.3 sobre la minería de patrones discriminatorios.

4.2. Preprocesamiento de datos del caso de estudio

La base de datos tiene 480 pacientes registrados entre 1992 y 2002. Para realizar este primer experimento, de la base de datos hemos eliminado todos los pacientes con datos faltantes sobre las variables seleccionadas para este estudio (101 pacientes). Aunque se podrían haber preservado los datos de más pacientes para realizar la minería de patrones secuenciales, eliminamos esta gran cantidad de registros para poder comparar los resultados con las puntuaciones de gravedad de la quemadura (según se explica en la Sección 1.2.4). Tras esta limpieza, quedan 379 pacientes, de los que el 79.95 % (303/76) sobreviven, el 69.39 % (263/116) son hombres y el 47.23 % (179/200) tienen lesiones por inhalación. La Tabla 4.1 muestra un resumen de los atributos estáticos de la base de datos.

4.3. Proceso de descubrimiento del conocimiento en 4 pasos

Con el fin de crear modelos para predecir la mortalidad en las Unidades de Quemados Críticos, definimos un proceso de descubrimiento de conocimiento en 4 pasos. Los dos primeros pasos se centran en el procesamiento previo de la base de datos y se utiliza una técnica

de descubrimiento de patrones para mostrar la evolución de los pacientes. A continuación, proponemos un posprocesamiento de los patrones con el fin de reducir el número de patrones descubiertos. Por último, con el fin de obtener modelos interpretables, los patrones que permanecen se utilizan para construir modelos de clasificación en forma de reglas y árboles de decisión.

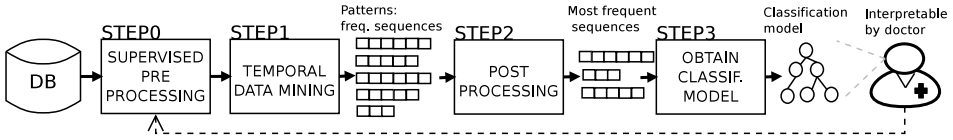

Figura 4.1: Proceso de descubrimiento de conocimientos en 4 pasos.

4.3.1. Paso 0: discretización de atributos temporales

El primer paso se centra en la discretización de todos los atributos temporales por tres razones: (1) evitamos la variabilidad en los puntos de división de atributos continuos calculados en cada rama del árbol de decisión, (2) los patrones son fáciles de interpretar por los médicos ya que contienen su lenguaje habitual, y (3) aumentamos la legibilidad de los patrones debido al menor número de variables. Los antecedentes de la discretización se pueden consultar en el Capítulo 3 del estado del arte, concretamente en la Sección 3.2.1.1.

Con el fin de discretizar, por un lado, se podrían usar técnicas basadas en la teoría de la información o técnicas estadísticas. Por otro lado, un experto puede proporcionar una abstracción cualitativa. Este último puede introducir un sesgo, e incluso la capacidad predictiva de la variable puede disminuir respecto a discretizaciones automáticas.

En nuestro caso, inicialmente adoptamos un enfoque de discretización por rangos de referencia realizada por un experto con el objetivo de que se puedan interpretar fácilmente los patrones resultantes en su lenguaje habitual.

En el Capítulo 5 se estudian como afectan diferentes discretizaciones a los resultados de la clasificación. Se pueden consultar los intervalos de discretización establecidos por el experto en la Sección 1.2.4.1.

4.3.2. Paso 1: minería de patrones secuenciales multivariantes

Según [4], la minería de patrones secuenciales es “la extracción de todas las subsecuencias frecuentes presentes en un conjunto de secuencias, donde cada secuencia es un conjunto ordenado de elementos”. En nuestro caso, la extracción de secuencias frecuentes se procesa a partir de un conjunto de series clínicas temporales de los pacientes.

Puede encontrar más información sobre la definición de este problema en la Sección 3.3.2.

En esta tesis usamos el algoritmo FaSPIP [50] para realizar la minería de patrones secuenciales multivariantes. FaSPIP se basa en la estrategia de clases de equivalencia y es capaz de extraer tanto puntos como intervalos. Además, FaSPIP utiliza un nuevo algoritmo de generación de candidatos basado en puntos límite y métodos eficientes para evitar la generación de candidatos inútiles y comprobar su frecuencia.

4.3.3. Paso 2: posprocesamiento

El fenómeno de la explosión de patrones es un importante inconveniente del uso de patrones como predictores para clasificadores. Si el soporte usado como umbral por los algoritmos es bajo, el número de patrones frecuentes aumenta bruscamente. Este problema se vuelve extremadamente limitante cuando trabajamos con bases de datos grandes. Puede ampliar más información sobre la calidad e interés de los patrones en la Sección 3.5.

Un enfoque interesante utilizado para resolver este problema consiste en utilizar una representación comprimida de los patrones, buscando patrones con propiedades específicas como los patrones cerrados [97], patrones maximales [14] o patrones minimales [40]. Podemos adaptar la definición del conjunto de ítems cerrado, maximal o minimal explicados en la Sección 3.5.1 a su utilización con secuencias del siguiente modo:

Definición 4 *Una secuencia frecuente S_i es cerrada si no existe otra supersecuencia frecuente con el mismo soporte.*

Definición 5 *Una secuencia frecuente S_i es maximal si no existe otra supersecuencia frecuente de ella.*

Definición 6 *Una secuencia frecuente S_i es minimal si no existe otra subsecuencia frecuente de ella.*

El número final de secuencias frecuentes extraídas al realizar este posprocesamiento va disminuyendo progresivamente en el conjunto de secuencias cerradas, maximal y minimal. Nótese que los patrones minimales se encuentran sólo después de haber procedido a reducir el número inicial de patrones, eliminado aquellos patrones que no son interesantes. En otro caso, no podrían encontrarse debido a la monotonicidad en la búsqueda de patrones.

La búsqueda de patrones cerrados proporciona dos beneficios al mismo tiempo: una reducción en el número de candidatos, y una salida más compacta mientras que se mantiene

la máxima cantidad de información [97]. Sin embargo, con las secuencias maximales o minimales se pierde información del soporte y no seremos capaces de reconstruir exactamente el conjunto de patrones frecuentes original. Los patrones minimales son los más generales y cortos, y posiblemente los más robustos y adecuados para clasificación. Según [40] los patrones JEP minimales son los más expresivos, ya que un JEP más corto significa que tiene menos ítems (atributos). Si podemos usar menos atributos para distinguir dos clases de datos, agregar más atributos no contribuirá a la clasificación y, lo que es peor, generará ruido en el clasificador.

En este capítulo exploramos el uso de patrones secuenciales cerrados y maximales en lugar de usar únicamente los patrones frecuentes como predictores con el fin de aumentar la capacidad discriminatoria.

Por otra parte, se puede obtener un conjunto aún más reducido de patrones predictores generando los conjuntos de patrones emergentes [34] o los patrones JEP (Jumping Emerging Patterns). Los JEP son aquellos patrones que se encuentran en el conjunto de los supervivientes y que no se encuentra en el conjunto de los fallecidos (y viceversa). De esta forma, se eliminan los patrones con comportamiento que sea común, o que no son buenos discriminantes para clasificación. Además, para obtener reglas más robustas al ruido en los datos, hemos incluido patrones emergentes con un pequeño grado de solapamiento en ambas clases.

4.3.4. Paso 3: algoritmos de clasificación con modelos interpretables

En el proceso de descubrimiento de conocimiento preferimos elegir un modelo fácil de interpretar por el médico. En este capítulo usamos la poda incremental repetida para producir la reducción de errores (Repeated Incremental Pruning to Produce Error Reduction, RIPPER [28]) para el aprendizaje de reglas. Con este algoritmo de cobertura secuencial, las reglas se aprenden de una en una y cada vez que se aprende una regla, se eliminan las tuplas cubiertas por la regla. Este proceso se repite hasta que no se encuentren más ejemplos de entrenamiento o si la calidad de una regla obtenida está por debajo de un umbral especificado por el usuario. Según [94], JRIP (la implementación de RIPPER en WEKA), junto con PART, es uno de los mejores algoritmos de clasificación para combinar legibilidad y precisión.

En los otros capítulos optamos además por el uso de otro algoritmo diferente para cubrir posibles diferencias en la estructura de la base de datos, de manera que adicionalmente se ha elegido el árbol de decisiones J48 implementado por WEKA para el algoritmo C4.5 de Quinlan [101]. Entre algunas de las ventajas de los árboles de decisión tenemos que: son fáciles de entender, se convierten fácilmente en un conjunto de reglas de producción, y pueden clasificar tanto datos categóricos como numéricos. El algoritmo J48 emplea una técnica

voraz que es una variante de ID3, que determina el atributo más predictivo en cada paso y divide un nodo en función de este atributo. Según [89], con J48 se produce en general una clasificación de alta precisión y con una estructura de árbol simple. Además, los autores en [62] muestran que el árbol de decisión J48 proporciona el modelo más simple utilizando un conjunto de datos de la Unidad de Quemados Críticos, de esta manera, es más fácil de interpretar por los médicos.

4.4. Experimentos

Hemos llevado a cabo varios experimentos que realizan la minería de patrones y los usamos como predictores en los clasificadores. Como primer experimento consideramos simplemente todos los patrones extraídos de la base de datos con FaSPIP utilizando diferentes soportes. A continuación, hemos explorado como posprocesamiento de todos los patrones extraidos la selección de patrones cerrados y de patrones maximales, y hemos estudiado sus características principales.

La Tabla 4.2 muestra el número total de patrones descubiertos utilizando diferentes valores de soporte, la longitud máxima de los patrones en días y el número máximo de eventos en los patrones. Se considera como un evento cada uno de los ítems de un patrón. Un patrón que ocurra en pocos días y tenga pocos eventos se considerará más general que otro patrón específico que incluya todos los días (el máximo de días posibles recogidos son 5) con un alto número de eventos. Si usamos un soporte alto, sólo se encuentran patrones generales muy comunes y cortos; estos patrones tienen menor valor discriminatorio, ya que aparecen tanto en sobrevivientes como en no sobrevivientes con un soporte similar. No utilizamos soportes por debajo del 5 % ya que el patrón más largo encontrado con un soporte del 10 % ya se extiende a los 5 días que abarca la recogida de datos.

El conjunto de patrones cerrados es menor que el original, y el conjunto de patrones maximales es el más pequeño. En general, los patrones largos al ser más específicos deberían ser más propensos a sobreajustar, pero si consideramos la propiedad de monotonicidad, un patrón cerrado resume todos sus subpatrones sin perder información (el soporte). Cuando elegimos sólo patrones maximales, obtenemos la evolución esencial, pero perdemos información sobre el soporte. Por lo general, un conjunto de patrones largos (que contienen más de 5 elementos en nuestro caso) tendría un prefijo común y sólo diferiría en unos pocos elementos, con una capacidad discriminatoria similar a la de sus superpatrones.

Llevamos también a cabo un segundo experimento que añade al posprocesamiento además de la representación comprimida de los patrones (con patrones cerrados o maximales), la exploración de patrones discriminatorios para intentar reducir aún más su número. De for-

Soporte	Tipo patrones	Número	Cerrados	Maximales	días/eventos
40 %	Todos	311	289	227	3 / 4
20 %	Todos	4295	4269	3137	4 / 6
15 %	Todos	11525	11498	8424	4 / 7
10 %	Todos	43193	43151	31236	5 / 8
5 %	Todos	373051	372013	261560	5 / 9
15 %	Emergentes	82	81	77	4 / 6
10 %	Emergentes	16154	14867	13546	5 / 9
15 %	JEP	43	42	41	4 / 6
10 %	JEP	16115	14828	13526	5 / 9
10 %	Con clase	24878	24873	18123	4 / 7

Tabla 4.2: Número de patrones y longitud máxima en cada test.

ma que extraeremos patrones emergentes con un pequeño grado de superposición en ambas clases (patrones que se encuentran en los sobrevivientes y como máximo en el 2 % de los no sobrevivientes), y adicionalmente restringiremos más aún el número de patrones, al extraer patrones JEP consistentes en patrones exclusivos del subconjunto de sobrevivientes o del subconjunto de no sobrevivientes, con el fin de eliminar totalmente el comportamiento común o la evolución del paciente que no es discriminatoria.

En la Tabla 4.2 se muestra la reducción producida en el número de patrones al utilizar estas características. Así por ejemplo, con un soporte del 15 % y una representación cerrada de los patrones encontramos únicamente 81 patrones emergentes y 42 patrones JEP. Como era de esperar, el número de patrones (predictores) se reduce claramente utilizando el mismo soporte.

Llevamos a cabo un tercer experimento para tratar de mantener patrones más discriminatorios. Con ese fin, incluimos la clase como una nueva transacción en el día 6 en cada secuencia del paciente. Después de minar los patrones, descartamos los patrones frecuentes que no contenían la clase como último elemento. A continuación, eliminamos la clase y comprobamos de nuevo a los pacientes cuya secuencia contiene el patrón (independientemente de que sobrevivan o no). Los nombramos como patrones *con clase* en las Tablas 4.2 y 4.3. Contrariamente a lo que pensábamos, encontramos aún más patrones de este tipo usando el mismo soporte porque la clase es muy frecuente en la base de datos y se encuentra en muchos patrones.

En la Tabla 4.3 comparamos los clasificadores creados exclusivamente con conocimiento estático y los creados utilizando patrones como predictores con los tres experimentos anteriores. Para los predictores, utilizamos los patrones obtenidos con un soporte del 10 % para evitar el sobreajuste con patrones demasiado específicos (experimentos con valores de

Tipos patrones	Sensibilidad	Especificidad	Precisión	AUC
Atributos estáticos	91.9 %	38.16 %	80.47 %	0.6681
Todos	88.78 %	19.74 %	74.93 %	0.54
Todos (Cerrados)	91.09 %	31.58 %	79.16 %	0.606
Todos (Maximales)	94.06 %	19.74 %	79.16 %	0.574
Emergentes (Maximales)	100 %	78.95 %	95.78 %	0.87
JEP (Cerrados)	100 %	80.26 %	96.04 %	0.899
Con clase (Todos)	94.72 %	25 %	80.74 %	0.6031

Tabla 4.3: Resultados de los experimentos de clasificación con el algoritmo RIPPER (JRIP en WEKA).

soporte más altos proporcionaron peores resultados). En ambos casos, configuramos los clasificadores con el mismo número mínimo de elementos en cada hoja o regla al 2 % de las instancias. En la base de datos con atributos estáticos, no usamos el atributo LOS, ya que contiene información que se extiende más allá de los cinco primeros días después de la admisión. La precisión, sensibilidad, especificidad y AUC se calculan con una validación cruzada de 10 iteraciones (10-fold cross validation).

4.5. Discusión

Como era de esperar, la sensibilidad es muy alta en todos los experimentos, ya que la proporción de sobrevivientes es de aproximadamente el 80 % en la base de datos. La baja especificidad en los experimentos que no utilizan patrones discriminatorios puede deberse al alto número de patrones secuenciales y al bajo número de pacientes en la base de datos. En general, los patrones cerrados mantienen más información que los maximales, produciendo mejor especificidad.

El uso de patrones emergentes o JEP mejora la especificidad dada por el clasificador respecto a cuando solo se usan atributos estáticos. Esto se debe al hecho de que los modelos son capaces de incluir reglas para una mejor predicción de los fallecidos. Entre los diferentes tipos de patrones discriminatorios, los modelos no muestran diferencias significativas ni en precisión ni en AUC.

El uso de patrones añadiendo la clase no mejoró los clasificadores. Esto puede deberse a que la variable de clase es muy común y los patrones obtenidos no fueron discriminatorios. RIPPER generó sólo 2 reglas para patrones normales, cerrados y maximales.

Con los datos médicos proporcionados se genera un alto número de patrones redundantes, ya que algunas de las variables están estrechamente relacionadas, así por ejemplo, el balance

de fluidos (BAL) está relacionado con los entrantes (INC) y la diuresis (DIU). Sin embargo, el significado de estas variables es diferente ya que los entrantes son un signo relacionado con los tratamientos, mientras que la diuresis es un signo de cómo se comporta el riñón. Aunque los algoritmos no incluyen patrones correlacionados en el mismo modelo, sería interesante hacer el mismo experimento usando menos variables.

```
RULE 1:
IF (BIC_LOW => BAL_VERY_HIGH => EB_NORMAL) AND NOT (EB_NORMAL => PH_NORMAL)
      AND NOT (BAL_POSITIVE => (BIC_NORMAL+PH_NORMAL)) THEN NOT SURVIVE
RULE 2:
IF (EB_VERY_LOW => (DIU_AUGMENTED + EB_VERY_LOW))
        AND NOT (INC_HIGH + BIC_NORMAL) => DIU_AUGMENTED)
        AND NOT (INC_HIGH => PH_NORMAL => DIU_AUGMENTED)
        AND NOT (BAL_VERY_POSITIVE => (INC_LOW + BAL_POSITIVE)) THEN NOT SURVIVE
RULE 3: ELSE SURVIVE
```

Figura 4.2: Un modelo RIPPER utilizando patrones secuenciales multivariantes.

Por último, también ha llamado nuestra atención el tamaño de los modelos. Encontramos que los modelos RIPPER son más pequeños que los árboles de decisión. Las reglas suelen contener solo unos pocos patrones (de 3 a 11), y pueden ser fácilmente comprendidas e interpretadas por el experto. Por ejemplo, en la Figura 4.2 muestra un modelo que muestra una regla predeterminada para los supervivientes y varias reglas para los fallecidos. Las reglas generalmente contienen un patrón con una evolución, y varios patrones negados (ver AND NOT “y no”) que contienen eventos que no se observan en el paciente.

Con el fin de comparar nuestro clasificador con las puntuaciones clínicas, hemos calculado la puntuación de Brier para las puntuaciones de Baux (0.12), R-Baux (0.12) y ABSI (0.14) para los pacientes con quemaduras, además de la puntuación SAPSII (0.14) y la de nuestro clasificador (0.04). Nuestro clasificador supera claramente las puntuaciones clínicas para la predicción de mortalidad.

4.6. Conclusiones

En este capítulo proponemos un proceso de descubrimiento de conocimiento de 4 pasos para construir el modelo de clasificación en el caso de estudio para predecir la mortalidad.

Las principales aportaciones de este trabajo son las siguientes:

- Hasta donde sabemos, este es el primer trabajo donde patrones secuenciales multivariantes se utilizan como predictores de mortalidad en una Unidad de Quemados Críticos o en una UCI.

- La utilización de patrones emergentes o JEP ha permitido una gran reducción en su número, haciendo frente al problema de la explosión de patrones, así por ejemplo, con

un 10 % de soporte, de encontrar inicialmente 43193 patrones, pasamos a seleccionar solamente 16154 patrones emergentes (-62.6 %) o 16115 patrones JEP (-62.7 %), siendo esta reducción drástica cuando se aumenta el soporte.

- La utilización de una representación comprimida de los patrones logra disminuir aún más su número, de forma que se consigue una mayor reducción con los patrones maximales que con los patrones cerrados. Así por ejemplo, con un 10 % de soporte, de los 16154 patrones emergentes iniciales, encontramos 14867 patrones cerrados (-8 %) y 13546 patrones maximales (-9 % adicional respecto al número de patrones cerrados).

- Los resultados de las pruebas de clasificación muestran que solamente se consiguen buenos resultados cuando se mejora la calidad de los patrones, proporcionando los mejores resultados de clasificación la utilización de patrones JEP. Nuestro enfoque además supera en puntuación de Brier a las puntuaciones de gravedad de quemaduras utilizadas actualmente por los médicos. A diferencia de estas puntuaciones, basadas en los datos recogidos en el momento de la admisión, nuestra propuesta se basa en los datos del paciente en los primeros cinco días. Destacamos el interés de los médicos en este tipo de patrones ya que pueden proporcionar información sobre la posible evolución de las variables del paciente (y por lo tanto, la respuesta al tratamiento).

Capítulo 5

Impacto de la discretización de series temporales en la clasificación

En el proceso de descubrimiento de conocimiento, dentro del paso previo de preprocesamiento de datos, el método de discretización elegido puede tener un impacto notable en el rendimiento y la precisión de los algoritmos de clasificación. En este capítulo, basado en un artículo publicado en *Progress in Artificial Intelligence* [22], que al mismo tiempo amplía el trabajo presentado a la *XVII Conferencia de la Asociación Española para la Inteligencia Artificial* (CAEPIA 2016) [21], analizamos y comparamos algoritmos de discretización generados de manera automática o bien por expertos. En particular, estudiamos su impacto para predecir la supervivencia de los pacientes en el contexto de las Unidades de Quemados Críticos. Evaluaremos la calidad de los diferentes algoritmos de discretización analizando el número de intervalos generados, la cantidad de patrones producidos y el rendimiento en la clasificación en un problema clínico específico.

5.1. Introducción

La clasificación de datos temporales, especialmente de series temporales multivariantes y univariantes, es una tarea muy desafiante e importante en muchos dominios. Es esencial en una multitud de dominios médicos diferentes, en los que la correcta clasificación de los datos de series temporales tiene implicaciones inmediatas para el diagnóstico, para la evaluación de la calidad y para la predicción de resultados significativos [90].

Hay dos enfoques principales en la clasificación de series temporales: la extracción de características y la comparación directa. Sin embargo, cuando las series temporales son multivariantes, ordenadas y no periódicas, esos enfoques tienen limitaciones. Alternativamente,

podemos explotar estos datos temporales extrayendo los patrones para llevar a cabo un problema de clasificación. En tales situaciones, la discretización desempeña un papel clave.

Sin embargo, el proceso para transformar características continuas en intervalos discretos (métodos de discretización) puede sesgar los modelos obtenidos por algoritmos de clasificación. A pesar de que este es un problema abierto clave, se le ha prestado poca atención.

Según [85], la discretización debe dar lugar a particiones que (a) reflejen la distribución original del atributo continuo, (b) mantengan los atributos de cualquier patrón sin agregar ficticios, y (c) sean significativas e interpretables para los expertos del dominio.

La discretización se realiza normalmente utilizando la distribución de la probabilidad o utilizando parámetros estadísticos como la frecuencia de cada clase. En algunos ámbitos, como la medicina, la discretización usualmente es realizada de forma manual por expertos. Una amplia gama de algoritmos de discretización han sido ideados y evaluados, sin embargo, pocos estudios han examinado la discretización de datos clínicos [75], y ninguno de ellos, que conozcamos, han tratado con patrones secuenciales específicamente.

En este capítulo comparamos el efecto de la discretización en la clasificación de la supervivencia de pacientes en la Unidad de Quemados Críticos a partir de la evolución diaria de los pacientes utilizando patrones secuenciales multivariantes. Extraemos seis series temporales de datos fisiológicos y de laboratorio de la historia clínica, y evaluamos diferentes enfoques de discretización (discretización experta y otros 25 métodos de discretización utilizando el paquete Keel [7]). Analizamos el número de intervalos generados con cada algoritmo de discretización, estudiando cómo la discretización afecta al número de patrones extraídos. Por último, discutimos su efecto en el rendimiento de la clasificación, la importancia de los patrones obtenidos y la interpretabilidad de los resultados.

Los antecedentes de la discretización se pueden consultar en el Capítulo 3 del estado del arte, concretamente en la Sección 3.2.1.1.

5.2. Métodos de discretización

En el Capítulo 4 utilizamos patrones discriminatorios realizando una abstracción temporal basada en el conocimiento de un experto y luego construimos clasificadores de la supervivencia de los pacientes con una alta sensibilidad y especificidad. Para ello presentamos un proceso de descubrimiento de conocimiento de 4 pasos (véase Sección 4.3), con el fin de crear modelos para predecir la mortalidad en la Unidad de Quemados Críticos utilizando datos temporales. Este proceso consta de un paso inicial (Paso 0) para preprocesar el conjunto de datos clínicos. A continuación, el conjunto de datos es minado para extraer patrones secuenciales presentes en la serie temporal multivariante (Paso 1). Estos patrones secuenciales

representan la evolución más frecuente de los pacientes. Debido al gran número de patrones, el siguiente paso (Paso 2) es un posprocesamiento de los patrones para reducir su número. El último paso es obtener modelos interpretables (Paso 3), donde los patrones seleccionados se procesan para construir modelos de clasificación en forma de reglas y árboles de decisión.

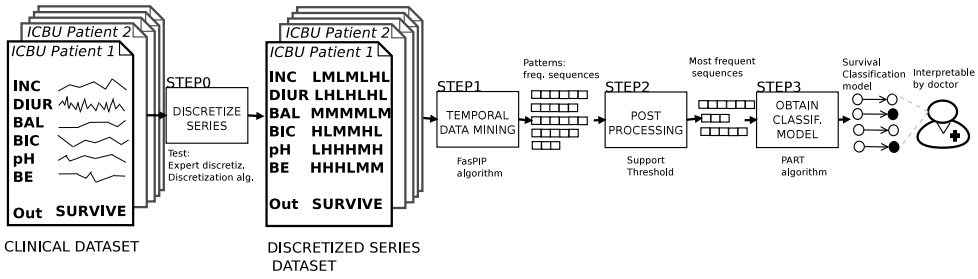

Figura 5.1: Proceso de descubrimiento de conocimiento en 4 pasos.

En este capítulo, adaptamos este proceso en 4 pasos para estudiar el efecto de los métodos de discretización. Para ello, el paso de preprocesamiento (Paso 0) incluirá la discretización de datos temporales en diferentes formas. En el Paso 1, los patrones secuenciales se obtienen utilizando el algoritmo FaSPIP [50] teniendo en cuenta un umbral de soporte. El paso posterior al procesamiento (Paso 2) reduce el número de patrones frecuentes. Por último, se obtiene un modelo de clasificación basado en reglas mediante PART [94]. Las reglas extraídas predicen la supervivencia o no supervivencia de los pacientes. La Figura 5.1 ilustra un ejemplo de este proceso.

Con el fin de analizar las ventajas que puede proporcionar la discretización automática, hemos considerado 25 algoritmos diferentes de discretización automática incluidos en KEEL Software Tool [7]. Hemos utilizado los valores predeterminados propuestos por KEEL para cada algoritmo, pero hemos fijado el número de intervalos a 4 cuando ha sido posible. Esto es debido a que los facultativos están acostumbrados a tener al menos tres intervalos en sus dominios médicos, con al menos lo que ellos consideran valores bajos, normales o altos. Por otro lado, la discretización con la misma frecuencia es un algoritmo muy utilizado de discretización, estándar y muy simple, donde los datos se dividen generalmente en cuatro cuartiles. Por lo tanto, no vamos a elegir algoritmos de discretización con un número muy bajo de intervalos (menos de 3), o con un número muy alto de ellos (superior a 10). De esta manera, vamos a comparar entre algoritmos de discretización con un número similar de intervalos, incluyendo la discretización experta, donde cada intervalo puede tener un significado para el médico y conseguiremos que no se produzca un sobreajuste de los datos.

Después de realizar la discretización obtenemos el siguiente número de puntos de corte para cada atributo (consulte la Tabla 5.1).

Algoritmo de discretización	Intervalos	Número de puntos de corte					
		INC	DIUR	BAL	BIC	pH	BE
Experta (Rango de referencia)	Fijos	3	3	4	4	6	4
1R (1R)	Automáticos	128	140	156	4	1	6
Ameva (Ameva)	Automáticos	1	1	1	1	1	1
Bayesian (Bayesian)	Automáticos	448	513	612	15	4	28
Class-Atribute Interdependence Maximization (CAIM)	Automáticos	1	1	1	1	1	1
Chi2 (Chi2)	Automáticos	26	24	38	2	3	4
Chi-Merge (ChiMerge)	Automáticos	1	0	1	1	1	1
Cluster Analysis (ClusterAnalysis)	Automáticos	231	150	172	49	51	90
Distribution-Index-Based Discretizer (DIBD)	Fijos (4)	2	1	1	1	1	1
Extended Chi2 (ExtendedChi2)	Automáticos	0	3	4	2	3	3
Fayyad e Irani (Fayyad)	Automáticos	0	0	1	2	1	2
Fusinter (FUSINTER)	Automáticos	3	4	5	2	4	3
Hypercube Division-Based (HDD)	Automáticos	1	1	1	1	1	1
Hellinger-based Discretizer (HellingerBD)	Fijos (4)	3	3	3	3	3	3
Iterative Dicotomizer 3 (ID3)	Automáticos	669	649	658	83	47	148
Interval Distance-Based Method for Discretization (IDD)	Fijos (4)	3	3	3	3	3	3
Khiops Discretizer (Khiops)	Automáticos	17	14	19	12	18	14
Mantaras Distance-Based (MantarasDist)	Automáticos	0	0	1	1	1	1
Modified Chi2 (ModifiedChi2)	Automáticos	355	397	461	20	10	31
Multivariate Discretization (MVD)	Fijos (4)	0	0	0	0	0	3
Proportional (Proportional)	Automáticos	47	47	47	31	24	41
Unsupervised Correlation Preserving Discretization (UCPD)	Fijos (3,5)	4	4	4	4	4	4
Uniform Frequency (UniformFrequency)	Fijos (4)	3	3	3	3	3	3
Uniform Width (UniformWidth)	Fijos (4)	3	3	3	3	3	3
Unparametrized Supervised Discretizer (USD)	Automáticos	454	530	614	44	16	41
Zeta (Zeta)	Automáticos	1	1	1	1	1	1

Tabla 5.1: Número de puntos de corte de los atributos del conjunto de datos clínicos (INC, DIUR, BAL, BIC, pH, BE) utilizando diferentes métodos de discretización.

Al seleccionar el número de intervalos, generalmente hay una compensación entre una mayor precisión y una mayor consistencia [85]. Podemos observar que el número de intervalos generados automáticamente por cada algoritmo de discretización es muy diferente, existiendo discretizaciones de 1 a cientos de puntos de corte.

Un algoritmo de discretización debe generar el menor número posible de intervalos discretos. Sin embargo, muy pocos intervalos pueden ocultar información sobre la relación entre la variable de clase y la variable de intervalo. Pero, si el número de intervalos es grande, entonces hay una mayor probabilidad de que un modelo sobreajuste los datos [26].

En las siguientes secciones detallamos solamente los métodos de discretización que tienen un número de intervalos entre 3 y 10, tal y como se ha justificado previamente, y se muestran los puntos de corte generados para cada atributo.

5.2.1. Algoritmos de discretización utilizados

5.2.1.1. Discretización experta

La discretización con rangos de referencia realizada por un experto se puede consultar en la Sección 1.2.4.1. En nuestro caso, los atributos relacionados con los fluidos se normalizan en valores por hora y por kg de peso del paciente, y después se han particionado según cuartiles, mientras que los atributos relacionados con el balance ácido-base se discretizan según los valores de referencia dados por el laboratorio.

5.2.1.2. Distribution-Index-Based Discretizer (DIBD)

El discretizador basado en índices de distribución (Distribution-Index-Based Discretizer, DIBD) [126] tiene en cuenta la distribución natural de los valores de datos y se basa en definiciones de entropía dicotómica y en un índice distributivo compuesto. La entropía dicotómica indica el grado de homogeneidad de la distribución del valor de la decisión, y se aplica para determinar el mejor punto de división. Este índice combina tanto los grados de homogeneidad de las distribuciones de valor del atributo como la distribución del valor de la decisión y se aplica para determinar si se acepta el nuevo punto de corte. El índice de distribución compuesto para un intervalo siempre disminuye cuando un intervalo grande se divide en dos intervalos pequeños. Sobre la base de este enfoque, un área de valor con alta ocurrencia y alto grado de homogeneidad es dividida en pequeños intervalos; de lo contrario, será dividida en grandes intervalos.

En la Tabla 5.2 puede ver los diferentes puntos de corte generados para cada atributo.

Punto de corte	INC	DIUR	BAL	BIC	pH	BE
Primero	0.0548	3.9256	4.0816	19.5	7.29	0.3
Segundo	0.0549					

Tabla 5.2: Puntos de corte para cada atributo usando la discretización DIBD.

Punto de corte	INC	DIUR	BAL	BIC	pH	BE
Primero		1.9574	-0.4494	19.95	7.285	-4.85
Segundo		1.9587	-0.4477	25.95	7.375	-4.75
Tercero		6.1487	0.1819		7.485	0.85
Cuarto			10.1706			

Tabla 5.3: Puntos de corte para cada atributo usando la discretización Extended Chi2.

5.2.1.3. Extended Chi2

El algoritmo de discretización ChiMerge introducido inicialmente por Kerber [64] es un método de discretización global supervisado de abajo hacia arriba ampliamente utilizado en la literatura. El principal inconveniente de ChiMerge es que el usuario tiene que proporcionar varios parámetros como los intervalos máximos y mínimos. Chi2 mejora el algoritmo original ChiMerge calculando automáticamente el valor del nivel de significancia, pero todavía requiere que los usuarios proporcionen una tasa de inconsistencia para detener el procedimiento de fusión.

El discretizador Extended Chi2 [111] determina la tasa de clasificación errónea predefinida a partir de los datos en sí. También considera el efecto de la varianza en los dos intervalos adyacentes. Con estas modificaciones, el algoritmo Extended Chi2 no solo maneja datos mal clasificados o inciertos, sino que también se convierte en un método de discretización completamente automatizado y su precisión predictiva es mejor que el algoritmo Chi2.

La Tabla 5.3 muestra los puntos de corte generados automáticamente por el discretizador Extended Chi2. El atributo de los entrantes (Incoming, INC) no se considerará, ya que no tiene ningún punto de corte.

5.2.1.4. Fayyad e Irani

Fayyad e Irani [119] propusieron una discretización supervisada basada en la entropía. En general, un método basado en la entropía utilizará la entropía con respecto a la clase de las particiones candidatas para seleccionar límites para la discretización. Esta entropía es una medida de pureza y mide la cantidad de información que se necesitaría para especificar a qué clase pertenece una instancia. En este proceso se tiene en cuenta un gran intervalo que

Punto de corte	INC	DIUR	BAL	BIC	pH	BE
Primero			4.087	19.55	7.295	-6.05
Segundo				25.25		0.35

Tabla 5.4: Puntos de corte para cada atributo mediante la discretización de Fayyad e Irani.

Punto de corte	INC	DIUR	BAL	BIC	pH	BE
Primero	0.0710	1.3178	-1.8016	19.05	7.29	-6.0
Segundo	0.0873	1.5401	0.9901	25.05	7.35	-1.9
Tercero	0.0925	1.6940	2.0073		7.42	0.75
Cuarto		3.9303	3.7648		7.485	
Quinto			8.0468			

Tabla 5.5: Puntos de corte para cada atributo mediante la discretización FUSINTER.

contiene todos los valores conocidos de una característica y, a continuación, recursivamente este intervalo se divide en subintervalos más pequeños hasta que se logra algún criterio de detención o un número óptimo de intervalos. Fayyad e Irani propusieron un criterio de detención para esta generalización utilizando el principio de longitud mínima de la descripción (Minimum Description Length Principle, MDLP), que afirma que la mejor hipótesis es la que tiene una longitud de descripción mínima.

Los intervalos generados por este algoritmo de discretización se representan en la Tabla 5.4. En los experimentos no vamos a utilizar los atributos de entrantes (INC) y de diuresis (DIU) porque se han reducido a un solo intervalo.

5.2.1.5. Fusinter

El método Fusinter [134] es un algoritmo de discretización supervisado que construye los intervalos de forma recursiva, siguiendo una estrategia ascendente. El punto de corte se elige a medida que se minimiza la entropía cuadrática. Este proceso se repite de forma recursiva y se detiene cuando no hay más combinaciones posibles. En la Tabla 5.5 se pueden ver los diferentes puntos de corte generados para cada atributo.

5.2.1.6. Hellinger-based Discretizer (HellingerBD)

El algoritmo de discretización basado en Hellinger (HellingerBD) [67] discretiza los valores numéricos para que el contenido de información de cada intervalo sea el máximo posible y tan igual entre ellos como sea posible. Con este fin, los autores definen la cantidad de información para cada intervalo (con respecto al atributo objetivo) en función de la diferencia entre las frecuencias de clase. Esta diferencia se calcula utilizando la distancia Hellinger.

Punto de corte	INC	DIUR	BAL	BIC	pH	BE
Primero	0.0874	0.9438	-1.3912	10.9	7.015	-9.9
Segundo	0.1205	0.9450	0.1290	15.6	7.07	-9.05
Tercero	0.1283	1.2504	4.9987	16.05	7.16	-4.75

Tabla 5.6: Puntos de corte para cada atributo mediante la discretización HellingerBD.

Punto de corte	INC	DIUR	BAL	BIC	pH	BE
Primero	0.0549	1.3383	4.1792	15.1	7.30	-0.7
Segundo	0.0814	1.4182	4.2262	16.1	7.35	-0.3
Tercero	0.0968	4.072	16.6375	17.5	7.40	0.1

Tabla 5.7: Puntos de corte para cada atributo mediante la discretización IDD.

Cuanto más diferentes sean estas dos frecuencias de clase, más información se define del intervalo para dar al atributo objetivo.

Los puntos de corte generados se pueden ver en la Tabla 5.6.

5.2.1.7. Interval Distance-Based Method for Discretization (IDD)

El Método de Discretización basado en Intervalos de Distancia (Interval Distance-Based Method for Discretization, IDD) [103] se fundamenta en distancias de intervalo mediante el concepto de Entorno-Delta en el espacio de destino, de forma que se tiene en cuenta el orden del atributo de clase, si existe, para que se pueda utilizar con clases discretas ordinales, así como con clases continuas, en caso de problemas de regresión. Además, el IDD también es aplicable cuando no hay ningún orden en la variable de clase, utilizando una distancia adecuada en la distribución de la variable de salida. El IDD, a diferencia de las técnicas de discretización supervisadas habituales, calcula los puntos de corte en un solo paso.

En primer lugar, el IDD ordena los distintos valores del atributo para discretizar. Estos son los puntos de corte potenciales del esquema de discretización. Luego, cuando es posible, crea para cada punto de corte sus intervalos a la izquierda y la derecha basados en su Entorno-delta para cada lado. Por último, el IDD calcula la distancia entre estos dos intervalos y le asigna el punto de corte específico. Aquellos con valores más altos se seleccionan como puntos de corte.

Los puntos de corte generados por el algoritmo IDD se describen en la Tabla 5.7.

5.2.1.8. Unsupervised Correlation Preserving Discretization (UCPD)

El método de Discretización de Preservación de la Correlación No Supervisada (Unsupervised Correlation Preserving Discretization, UCPD) [87] utiliza la distribución de atribu-

Punto de corte	INC	DIUR	BAL	BIC	pH	BE
Primero	0.2303	1.4941	0.5701	16.5	7.2767	-7.8833
Segundo	0.2913	1.7749	4.7804	17.6667	7.29	-6.63333
Tercero	0.3203	2.0857	7.4763	19.6167	7.3667	-4.3167
Cuarto	0.5588	2.1697	13.1833	23.0333	7.3983	-0.2333

Tabla 5.8: Puntos de corte para cada atributo mediante la discretización UCPD.

Cuartil	INC	DIU	BAL	BIC	pH	BE
Primer cuartil (Q1)	0.1057	1.0786	-0.0694	22	7.36	-2.15
Segundo cuartil (Q2)	0.1766	1.4713	1.0868	24	7.41	0
Tercer cuartil (Q3)	0.2971	2.0238	3.0750	26	7.44	2.1

Tabla 5.9: Cuartiles de cada atributo utilizando la discretización de frecuencia uniforme (UniformFrequency).

tos categóricos y continuos y la estructura de correlación subyacente en el conjunto de datos para obtener los intervalos discretos. Para hacer esto, UCPD utiliza un análisis de componentes principales (Principal Component Analysis, PCA) para identificar la estructura de correlación entre los atributos continuos y también patrones de asociación para capturar correlaciones cuando también hay atributos categóricos. PCA también ayuda a tratar con un conjunto de datos con una dimensionalidad muy alta porque permite reducir el número de dimensiones.

Los puntos de corte calculados por el método UCPD se muestran en la Tabla 5.8.

5.2.1.9. Uniform Frequency Discretizer

El algoritmo de discretización uniforme de frecuencia (UniformFrequency) (o con la misma frecuencia, en este caso utilizaremos cuartiles) crea una serie de intervalos de densidad iguales (cada intervalo tiene el mismo número de instancias). Hemos dividido todos los datos registrados durante cinco días de cada atributo en cuatro intervalos que contienen aproximadamente el mismo número de valores. Los tres cuartiles de cada atributo se encuentran en la Tabla 5.9.

5.2.1.10. Uniform Width Discretizer

El algoritmo de discretización de ancho uniforme (UniformWidth) indica los valores mínimos y máximos del atributo discretizado y a continuación divide el intervalo en el número definido por el usuario de intervalos discretos de igual ancho. Hemos definido el número de intervalos discretos en 4. Los puntos de corte generados por el algoritmo UniformWidth se

Punto de corte	INC	DIUR	BAL	BIC	pH	BE
Primero	4.6686	14.1550	36.5825	18.0	7.1525	-13.675
Segundo	9.3346	28.2722	79.7483	26.0	7.3250	-5.3500
Tercero	14.0006	42.3895	122.9142	34.0	7.4975	2.9750

Tabla 5.10: Puntos de corte para cada atributo mediante la discretización del mismo ancho (UniformWidth).

muestran en la Tabla 5.10.

El principal inconveniente de este método del mismo ancho es que en los casos en que las observaciones de resultados no se distribuyen uniformemente, se puede perder una gran cantidad de información importante después del proceso de discretización.

5.3. Experimentos

En esta sección analizamos empíricamente el impacto de los métodos de discretización de atributos temporales en un clasificador en el dominio clínico. Los experimentos se llevan a cabo después del proceso de detección de conocimiento de 4 pasos propuesto en la Sección 4.3 y utilizando el conjunto de datos clínicos de la Unidad de Quemados Críticos descrito en Sección 1.2.4.

En el Paso 0 de "discretización de atributos temporales", se considera que en el dominio médico una buena discretización debe generar un pequeño número de patrones. De esta forma se tiene la ventaja de que estos patrones son más significativos, ya que pocos patrones son generados, por lo que es más fácil encontrar un significado médico para un patrón específico, siendo este patrón más robusto. Además, simplifica el tratamiento a realizar en todos los patrones.

En el proceso de minería de datos hemos observado que el número de patrones generados depende en gran medida de la discretización elegida. Solamente se han aplicado los diferentes métodos de discretización seleccionados en la Sección 5.2.1.

En el Paso 1 de "minería de patrones" hemos utilizado el algoritmo FaSPIP. FaSPIP se basa en la Estrategia de Clases de Equivalencia y utiliza un nuevo algoritmo de generación de candidatos basado en puntos de límite y métodos eficientes para evitar la generación de candidatos inútiles y para comprobar su frecuencia [50].

En el Paso 2 como "posprocesamiento" extraemos patrones secuenciales exclusivos del subconjunto de sobrevivientes o del subconjunto de no sobrevivientes con el fin de eliminar el comportamiento común o la evolución del paciente que no es discriminatoria. De esta manera obtenemos los patrones JEP (véase Sección 3.5.3.1) que usaremos en el paso de

Algoritmo de Discretización	Número de patrones secuenciales (supervivientes + no supervivientes)		
	soporte 60 %	soporte 40 %	soporte 20 %
Experto	31 + 32	306+358	4,437+5,424
DIBD	3,580,206 + 306,064	46,252,122+3,675,313	NC+61,737,772
ExtendedChi2	307+278	3,043+2,484	56,587+36,618
Fayyad e Irani	1,111+716	10,899+4,939	145,155+53,936
Fusinter	114+67	1250+685	28,299+11,111
HellingerBD	64,940+42,799	713,772 + 436,284	13,479,327 + 8,410,687
IDD	8,656+2,167	127,969+22,969	3,028,049+434,845
UCPD	242+66	3,918+808	102,325+17,260
UniformFrequency	0+19	91+288	1,305+4,956
UniformWidth	1,449,774+1,169,449	17,284,432+12,135,602	NC+159,377,848

Tabla 5.11: Número de patrones secuenciales generados a partir del subconjunto de supervivientes y de no supervivientes utilizando un soporte del 60 %, 40 % y 20 % para cada método de discretización seleccionado.

clasificación para generar modelos interpretables.

Para estudiar inicialmente cómo influye la discretización en el número de patrones secuenciales generados, hemos calculado cuántos patrones se generan a partir del subconjunto de supervivientes y desde el subconjunto de no supervivientes, aplicando un soporte del 60 %, 40 % y 20 % (consulte Tabla 5.11).

Es razonable que si el soporte es alto, se generen pocos patrones, y este número debería aumentar razonablemente de manera gradual a medida que se reduce el soporte aplicado.

Con un alto soporte del 60 % ya podemos ver que hay una gran diferencia entre el número de patrones generados. Así que pasamos de no tener ningún patrón en el subconjunto de supervivientes (usando cuartiles con frecuencia uniforme) a obtener varios millones de patrones (con DIBD o con la discretización del mismo ancho -UniformWidth-).

Teniendo como objetivo la clasificación, no tiene sentido trabajar con discretizaciones que generan un gran número de patrones, y más si actualmente estamos utilizando un soporte alto.

Cuando reducimos el soporte, las discretizaciones anteriores siguen creciendo o incluso no se pueden calcular (marcadas como NC en la Tabla 5.11) debido al alto número de patrones generados. Además, hay discretizaciones con un número razonable de patrones cuando se utiliza un soporte alto, pero cuando el soporte disminuye, el número de patrones aumenta desproporcionadamente. Este es el caso de las discretizaciones HellingerBD e IDD.

Por lo tanto, para obtener patrones relevantes, vamos a trabajar sólo con discretizaciones que generan menos de medio millón de patrones utilizando un soporte bajo. Este sería el

caso de las siguientes discretizaciones: Experto, ExtendedChi2, Fayyad, Fusinter, UCPD y UniformFrequency.

De esta manera obtendremos un número relativamente pequeño de patrones, que tendrán una mayor significancia. En el siguiente paso, reduciremos significativamente el número de patrones seleccionando sólo patrones JEP. Idealmente, sólo se deberían encontrar cientos de patrones JEP, y por lo tanto se podría hacer una revisión manual del significado de cada patrón.

Por último, en el Paso 3 de "algoritmos de clasificación con modelos interpretables", con el fin de construir un modelo de clasificación interpretable utilizamos PART como algoritmo de aprendizaje de reglas que sean entendibles. PART (junto con JRIP) es uno de los mejores algoritmos de clasificación que combinan legibilidad y precisión (véase [94]).

Se ha elegido PART en este capítulo porque es una combinación de clasificación de aprendizaje de reglas C4.5 y RIPPER, siendo estos los dos algoritmos propuestos para realizar la clasificación en el Capítulo 4. PART es un algoritmo de aprendizaje de reglas mediante la estrategia de divide y vencerás que produce conjuntos de reglas llamadas listas de decisiones, que se ordenan como conjuntos de reglas. PART construye un árbol parcial de decisión C4.5 en cada iteración y convierte la mejor hoja en una regla.

Consideramos diferentes soportes, desde el 14 % al 6 % para generar los patrones, tratando de encontrar el soporte más alto que genera menos patrones con los mejores resultados de clasificación. Por lo tanto, obtenemos desde un pequeño número de patrones JEP (cientos) a miles de ellos.

En todos los casos, configuramos el clasificador con el mismo número mínimo de elementos en cada hoja o regla al 2 % de las instancias. La precisión, sensibilidad, especificidad y AUC se calculan con una validación cruzada de 10 iteraciones.

La Tabla 5.12 muestra los resultados de los experimentos utilizando diferentes algoritmos de discretización y variando el soporte de las reglas. También se muestra el número de patrones generados en el subconjunto de los supervivientes y en el conjunto de los no supervivientes. Además, mostramos el número final de patrones JEP obtenidos.

5.4. Discusión

Según los experimentos realizados (ver Tabla 5.12), los algoritmos de discretización con menos patrones JEP y alto soporte son los realizados por un experto (utilizando rangos de referencia) y con frecuencia uniforme (cuartiles). En cambio, el mayor número de patrones JEP es generado por Fayyad e Irani. La situación ideal sería encontrar pocos patrones JEP y con ellos obtener una gran precisión en la clasificación. Nuestro propósito es obtener

Discretización	Soporte	Patrones iniciales Supervivientes+Muertes	Patrones JEP	Sensibilidad	Especificidad	Precisión	AUC
Experto	10 %	46,041+83,015	391	100.00 %	43.68 %	89.46 %	0.709
	8 %	88,084+241,866	4,931	100.00 %	**56.32 %**	**91.83 %**	**0.782**
	6 %	224,952+492,504	47,113	100.00 %	44.83 %	89.68 %	0.720
Fayyad e Irani	14 %	377,769+143,939	7,641	100.00 %	0 %	81.29 %	0.486
	12 %	524,378+234,772	18,752	100.00 %	51.72 %	90.97 %	0.756
	10 %	794,546+417,847	56,679	100.00 %	52.87 %	91.18 %	0.750
Cuartiles	10 %	14,856+81,670	430	100.00 %	48.28 %	90.32 %	0.728
	8 %	28,144+249,486	8,978	100.00 %	49.43 %	90.54 %	0.749
	6 %	73,272+522,697	49,379	100.00 %	43.68 %	89.46 %	0.722
ExtendedChi2	14 %	197,383+120,219	322	100.00 %	42.53 %	89.25 %	0.709
	12 %	311,613+217,946	5,907	100.00 %	45.65 %	89.36 %	0.761
	10 %	563,470+445,943	52,103	100.00 %	41.38 %	89.03 %	0.716
Fusinter	14 %	110,037+39,487	1,390	100.00 %	55.17 %	91.61 %	0.748
	12 %	183,030+75,861	6,167	100.00 %	**60.92 %**	**92.69 %**	**0.796**
	10 %	354,383+168,650	37,220	100.00 %	43.66 %	89.46 %	0.739
UCPD	16 %	238,337+49,947	2,179	100.00 %	52.87 %	91.18 %	0.789
	14 %	396,238+68,654	7,556	100.00 %	**66.67 %**	**93.76 %**	**0.851**
	12 %	647,943+137,546	22,940	100.00 %	59.77 %	92.47 %	0.796

Tabla 5.12: Resultados de los experimentos con el algoritmo PART.

patrones comprensibles con una alta significancia clínica.

Por otro lado, el mejor rendimiento en la clasificación, tanto en precisión como en especificidad, se consigue con la discretización UCPD y Fusinter. También obtenemos un resultado aceptable con Fayyadd o la discretización de un Experto. El peor resultado se obtiene utilizando cuartiles o la discretización ExtendedChi2. Con Fayadd, cuando elevamos el soporte al 14 %, obtenemos una especificidad del 0 % porque no obtenemos patrones del subconjunto de los no supervivientes.

Garcia y otros [45] hicieron un análisis empírico en el aprendizaje supervisado y concluyeron que Fusinter, Chi2, Fayyad y UCPD obtienen un balance satisfactorio entre el número de intervalos producidos y la precisión.

Los resultados que obtuvimos corroboran las conclusiones de [45], excepto que Chi2 obtiene demasiados intervalos y por lo tanto preferimos ExtendedChi2.

Si dejamos de lado las medidas estadísticas de clasificación (como precisión, especificidad, etc.), los médicos preferirán un proceso de discretización en el que los intervalos representen evidencia clínica. Es decir, la semántica de los atributos también juega un papel esencial cuando hay que tomar una decisión médica. Como se muestra en la Tabla 5.12, la discretización del Experto obtiene resultados bastante aceptables que superan a muchos algoritmos de discretización automática.

La discretización automática produce cortes arbitrarios que generalmente no se corresponden con el conocimiento clínico y complica la comprensión y la interpretación. Al me-

nos, los cortes en los cuartiles tienen una interpretación matemática simple y útil para los médicos, mientras que la discretización basada en la información no la tiene. Pero hemos obtenido un resultado pobre utilizando la discretización con la misma frecuencia, por lo que es necesario buscar otras discretizaciones.

Como se indica en [45], hay discretizadores clásicos (EqualWidth, EqualFrequency) que ya no son los más eficaces, teniendo en cuenta que hay otros discretizadores también clásicos (1R, ID3) que se han mejorado a lo largo de los años. Pero en nuestro caso, 1R o ID3 se descartaron porque generan demasiados intervalos, mientras que EqualFrequency crea una gran cantidad de patrones debido a los valores atípicos existentes en los datos.

El mejor resultado lo obtiene UCPD, teniendo la especificidad alcanzada menos variación cuando cambiamos el soporte. UCPD se basa en el análisis de componentes principales y explota la estructura de correlación subyacente en los datos para obtener los intervalos discretos y asegurarse de que se conserven las correlaciones inherentes. La correlación entre las características continuas es considerada, así como las interacciones entre las características continuas y las categóricas. En nuestros experimentos, la variable BAL se asocia a la perfusión de líquidos, siendo normal en el dominio clínico que una variable afecta a otras. Con UCPD los intervalos de discretización resultantes son significativos y podrían revelar patrones ocultos en los datos [42].

5.5. Conclusiones

En este capítulo analizamos el efecto de los diferentes algoritmos de discretización en la clasificación utilizando patrones secuenciales multivariantes a partir de los datos clínicos de la Unidad de Quemados Críticos.

La discretización y el establecimiento de los diferentes intervalos se ha realizado bien mediante el uso de la abstracción basada en el conocimiento (discretización experta), o bien mediante discretizaciones basadas en técnicas probabilísticas, estadísticas y de la información incluidas en KEEL Software Tool [7].

A continuación se exponen las principales conclusiones obtenidas:

- Un algoritmo de discretización debe generar el menor número posible de intervalos discretos [26]. Sin embargo, hemos visto que este número tiene una gran variación cuando se genera automáticamente. Hemos elegido sólo aquellos algoritmos de discretización con al menos 3 intervalos y no superiores a 10. De esa manera tendremos una serie de intervalos similares a la discretización experta o con la misma frecuencia (cuartiles), que serán fácilmente entendidos por los médicos y nos aseguraremos de

que no vamos a sobreajustar los datos. Por lo tanto, no hemos utilizado las discretizaciones 1R, Bayesian, Chi2, ClusterAnalysis, ID3 o USD, porque tienen demasiados intervalos, mientras que las discretizaciones Ameva, CAIM, Chimerge, HDD, MantarasDist, MVD y Zeta se han desechado porque tienen muy pocos.

- En el proceso de minería de datos, el número de patrones generados depende en gran medida de la discretización elegida. Una buena discretización en el dominio médico debe generar un pequeño número de patrones y dar lugar a una buena precisión de clasificación. Si encontramos pocos patrones, estos son más significativos y normalmente más generales, y por lo tanto es más fácil encontrar un significado médico simple para un patrón específico. Otra ventaja adicional es que se simplifica el tratamiento a realizar en todos los patrones. Por lo tanto, descartamos las discretizaciones DIBD, HellingerBD, IDD y UniformWidth, porque generan demasiados patrones en comparación con las demás. Los algoritmos de discretización con menos patrones JEP y alto soporte son los realizados por un experto (mediante rangos de referencia) y la de frecuencia uniforme (cuartiles). En cambio, el mayor número de patrones JEP es generado por el método de Fayadd.

- En los experimentos, no hemos encontrado ninguna asociación entre el número de patrones JEP generados y la precisión del clasificador. Sin embargo, cuando dos clasificadores proporcionan la misma precisión, el que necesita menos patrones es más fácil de interpretar.

- Nuestros resultados muestran que muchos algoritmos automáticos tienen un rendimiento inferior a la discretización experta y que es necesario tener en cuenta la correlación existente entre las características continuas para obtener la mejor precisión.

- El mejor rendimiento en la clasificación se consigue con la discretización UCPD y Fusinter. Además, con UCPD la especificidad alcanzada tiene menor variación cuando cambiamos el soporte. También se obtiene un resultado aceptable con Fayyadd y la discretización de un Experto. El peor resultado se obtiene utilizando cuartiles y la discretización ExtendedChi2.

- Llegamos a la conclusión, que en el caso de la clasificación de patrones secuenciales en el dominio clínico, con los resultados ofrecidos aquí, podrían ayudar a reducir el conjunto de candidatos para elegir un discretizador. Nuestro estudio se limita al dominio específico de las Unidades de Quemados Críticos, sin embargo, nuestros experimentos coinciden con [45], de forma que se esperan obtener unos resultados similares en otros campos clínicos.

Capítulo 6

Evaluación de la consistencia de patrones secuenciales multivariantes

Este capítulo se basa en el artículo presentado en la *XVI Conferencia de la Asociación Española para la Inteligencia Artificial* (CAEPIA 2015) [19]. El problema abordado es el de reducir el alto número de patrones secuenciales multivariantes obtenidos en el proceso de minería a la vez que mantener la significación clínica de los mismos. Para ello, se ha realizado una validación en varias particiones estratificadas. Los experimentos realizados con una base de datos de 480 pacientes muestran una clara reducción del número de patrones y una mejora en la predicción de la supervivencia del paciente.

6.1. Introducción

En el Capítulo 4 usamos patrones emergentes y JEP para después crear clasificadores de la supervivencia de un paciente con una alta sensibilidad y especificidad, obteniendo unos buenos resultados desde el punto de vista de la clasificación. Sin embargo, necesitamos reducir el número de patrones, para que los patrones predictores tengan la mayor relevancia médica posible y se encuentren uniformemente distribuidos por toda la base de datos de pacientes.

El objetivo de este trabajo es mejorar la selección de patrones secuenciales multivariantes mediante una evaluación de la consistencia con distintas particiones, y explotar las propiedades basadas en soporte para descartar aquellos patrones no significativos. Para realizar esto, basándonos en la evaluación por retención (holdout) propuesta por [122] vamos a particionar los datos en 2 subconjuntos para obtener de cada partición los patrones correspondientes, y quedarnos solamente con aquellos patrones consistentes, de forma que se encuentren unifor-

memente distribuidos en la base de datos, demostrando por lo tanto su solidez, persistencia y equilibrio. A partir de estos patrones generaremos clasificadores que deberán ser entendibles por el médico sobre la posible evolución del paciente. Para lograr este objetivo proponemos un proceso de 6 pasos para el descubrimiento de conocimiento.

6.2. Extensión del método: Proceso en 6 pasos de descubrimiento del conocimiento

Para poder construir modelos predictivos de la mortalidad de los pacientes quemados en una Unidad de Quemados Críticos, hemos definido 6 pasos para la realización del descubrimiento del conocimiento:

1. Transformación y discretización de los atributos temporales
2. **Partición estratificada de los datos**
3. Minería de patrones secuenciales en cada partición
4. **Identificación de los patrones consistentes**
5. Posprocesamiento de los patrones consistentes
6. Clasificación con modelos interpretables

En este capítulo proponemos los pasos 2 y 4 como nuevos pasos que vamos a realizar respecto a los 4 pasos indicados en el Capítulo 4, en la Sección 4.3, de forma que nos permitirán reducir el número de patrones y quedarnos con los que sean clínicamente más significativos.

El descubrimiento de patrones se realizará de forma independiente en cada partición de los datos, y luego realizaremos un posprocesamiento para reducir el número de patrones mediante dos técnicas: primero nos quedamos con aquellos patrones que coincidan de las diferentes particiones y posteriormente realizamos una selección de patrones cerrados, maximales o minimales.

En el último paso (6) utilizaremos los patrones que nos queden para obtener modelos interpretables de clasificación en la forma de reglas.

A continuación explicamos con mayor nivel de detalle cada uno de estos pasos.

6.2.1. Paso 1: transformación y discretización de los atributos temporales

El primer paso consiste en la transformación y discretización de todos los atributos temporales. En nuestro caso hemos optado por seguir los términos clínicos indicados por los expertos con el objetivo de que se puedan interpretar los patrones resultantes en su lenguaje habitual. De esta manera, evitamos la variabilidad de los puntos de división de los atributos continuos calculados globalmente y localmente por los algoritmos de aprendizaje. La discretización experta finalmente aplicada se puede consultar en la Sección 1.2.4.1. Se podrían usar técnicas basadas en teoría de la información o técnicas estadísticas, pero perderíamos la ventaja buscada a pesar del riesgo de introducir cierto sesgo. Además, según mostramos en el Capítulo 5, la discretización experta obtiene unos resultados competitivos en la clasificación (tanto en precisión como en especificidad), existiendo muchos algoritmos automáticos que producen un rendimiento inferior a la discretización experta.

6.2.2. Paso 2: partición estratificada de los datos

Como el resultado final que se quiere obtener es clasificar adecuadamente los pacientes que van a morir o a sobrevivir, es importante que en cada una de las particiones que se creen exista el mismo porcentaje de pacientes que fallecen para que así se genere el mayor número de patrones coincidentes posibles.

6.2.3. Paso 3: minería de patrones secuenciales multivariantes en cada partición

Los patrones que vamos a obtener como predictores son patrones secuenciales multivariantes. Se puede encontrar una definición de estos patrones y el problema de minería de patrones temporales en el Capítulo 3 del estado del arte, concretamente en la Sección 3.3.

Para la minería usamos el algoritmo FaSPIP [50], que está basado en la estrategia de equivalencia de clases y es capaz de minar tanto puntos como intervalos. Además, FaSPIP usa un novedoso algoritmo de generación de candidatos basado en los puntos frontera y en métodos eficientes para evitar la generación de candidatos no válidos y para calcular su frecuencia.

6.2.4. Paso 4: identificación de los patrones consistentes

El fenómeno de la explosión de patrones es un importante inconveniente en el uso de patrones como predictores para los clasificadores. Si el soporte dado es bajo, el número de patrones frecuentes se incrementa severamente. Este problema llega a ser extremadamente limitador cuando se trabajan con grandes bases de datos.

En [122] para encontrar los patrones más significativos realiza el particionamiento de los datos en un conjunto de entrenamiento y otro de test, obteniendo primero los patrones del conjunto de entrenamiento y luego estos son evaluados estadísticamente contra el conjunto de test. En nuestro caso, vamos a utilizar solamente las propiedades de soporte del patrón para elegir los patrones significativos, de forma que eliminaremos aquellos patrones que se puedan dar en una partición de los pacientes, pero que no se dan en la otra partición. Se puede incluso aumentar el porcentaje de patrones eliminados si para un determinado patrón exigimos que aparezca distribuido de manera uniforme en ambas particiones.

Con esta selección de patrones, nos quedaremos con solamente aquellos que se encuentren uniformemente distribuidos en la base de datos (en todas las particiones).

6.2.5. Paso 5: posprocesamiento de los patrones consistentes

El objetivo de este posprocesamiento es obtener un conjunto más reducido de patrones predictores. Por un lado, hemos generado los conjuntos de patrones JEP y los patrones emergentes [34].

Por otro lado, se puede realizar la eliminación adicional de patrones no interesantes, buscando patrones con unas propiedades específicas, como son los patrones cerrados [97], los patrones maximales [14] o los patrones minimales [40].

Puede encontrar más información sobre este posprocesamiento en la Sección 4.3.

6.2.6. Paso 6: algoritmos de clasificación con modelos interpretables

En el proceso de descubrimiento del conocimiento tenemos que elegir un modelo interpretable por el médico, tal y como se explica en la Sección 4.3.4. Para ello hemos seleccionado las reglas de decisión, cuyas ventajas incluyen que son fáciles de entender, son fácilmente convertibles a reglas de producción y pueden clasificar tanto datos cualitativos como cuantitativos. Como clasificador, hemos elegido el algoritmo de cobertura secuencial RIPPER (Repeated Incremental Pruning to Produce Error Reduction) [28]. JRIP (la implementación de RIPPER en WEKA) es uno de los mejores algoritmos de clasificación para combinar legibilidad y precisión [94].

Patrón	Nº patrones	Posprocesado	Sensib.	Especif.	Precisión	AUC
Emergente	12512	Ninguno	92.33 %	51.72 %	84.73 %	0.695
Emergente	11523	Cerrado	94.97 %	52.87 %	87.10 %	0.753
Emergente	9949	Maximal	94.44 %	50.57 %	86.24 %	0.723
Emergente	8303	Minimal	93.92 %	57.47 %	87.10 %	0.750
JEP	4931	Ninguno	100 %	58.62 %	92.26 %	0.777
JEP	4515	Cerrado	100 %	57.47 %	92.04 %	0.800
JEP	4192	Maximal	100 %	48.28 %	90.32 %	0.762
JEP	3892	Minimal	100 %	51.72 %	90.97 %	0.762

Tabla 6.1: Experimento 1 de referencia: Resultados de la clasificación JRIP en WEKA, sin usar pasos 2 y 4.

Para realizar los experimentos hemos configurado los clasificadores con un mínimo del 2 % de instancias en cada regla. La precisión, sensibilidad, especificidad y AUC han sido calculados con una validación cruzada de 10 iteraciones.

6.3. Experimentos

Hemos diseñado varios experimentos con los patrones evolutivos de los 465 pacientes descritos en la Sección 1.2.4, en los que hemos ido reduciendo el número de patrones predictores en los clasificadores.

En el primer experimento, que utilizaremos de referencia y cuyos resultados se ven en la Tabla 6.1, no se han realizado los pasos 2 y 4 propuestos en este trabajo, y sí el resto de los pasos. Se ha usado un 8 % de soporte en la minería de patrones secuenciales multivariantes ya que con ese soporte se obtienen patrones que representan toda la base de datos al abarcar hasta los 5 días de evolución registrados. Estos son los resultados que usamos como referencia para contrastar los resultados de la propuesta actual.

Lo primero que se puede observar en la Tabla 6.1 es que el número de patrones JEP es bastante menor que el número de patrones emergentes, y que disminuye el número de patrones generados conforme vamos seleccionando patrones cerrados, maximales y minimales. Por otra parte, se puede ver que los patrones JEP proporcionan en general una mayor precisión, al tener una sensibilidad del 100 %. Esta sensibilidad se alcanza por la propia característica del patrón, que solo tiene pacientes que mueren o que viven. Sin embargo, el número de patrones predictores es alto.

En el segundo experimento introducimos los pasos 2 y 4 propuestos en este capítulo para obtener patrones consistentes y reducir el número de patrones predictores. Aunque de manera general se podrían usar k-particiones, debido al tamaño de nuestra base de datos sólo

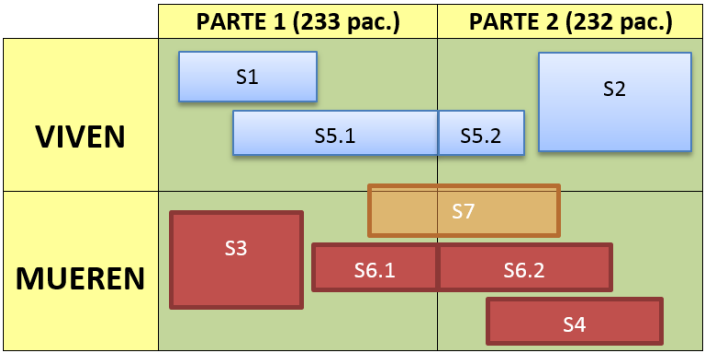

Figura 6.1: Conjunto de pertenencia de los posibles pacientes que puede tener un patrón extraído de cada una de las particiones.

hemos usado dos particiones estratificadas: la primera con 233 pacientes y la segunda mitad con 232, ambas con 189 supervivientes. Para clarificar más el experimento, mostramos la Figura 6.1.

En primer lugar, se extraen por separado los patrones de las dos particiones, los cuales pertenecerán a alguno de los subconjuntos indicados. Se pueden considerar como patrones JEP aquellos que contienen únicamente pacientes que sobreviven (S1, S2, S5) o sin ningún superviviente (S3, S4, S6). De estos patrones JEP seleccionaremos únicamente aquellos que son consistentes. Para ello, en nuestro ejemplo seleccionaremos aquellos patrones en los que se den casos con pacientes tanto en la partición 1 como en la 2, tal y como sería el patrón $S5 = S5.1 \cup S5.2$ con únicamente supervivientes, y el patrón $S6 = S6.1 \cup S6.2$ con solamente no supervivientes. En este caso, no hemos restringido el desbalanceo entre los subconjuntos $S5.1$ y $S5.2$, ni entre los subconjuntos $S6.1$ y $S6.2$.

El resultado de este experimento se muestra en la Tabla 6.2. Se puede observar que se ha logrado disminuir el número de patrones en más de un 50 % respecto al primer experimento. No se han encontrado apenas diferencias en la precisión y con los patrones JEP se ha mantenido la sensibilidad en el 100 % y se ha mejorado la especificidad con los patrones maximales y minimales. Las diferencias entre los patrones JEP y los emergentes, y los cerrados, maximales y minimales son similares a las del primer experimento.

El tercer experimento es similar al segundo, pero restringiendo el desbalanceo a un máximo del 20 % entre las dos particiones. Al imponer condiciones más estrictas, conseguimos en primer lugar obtener patrones más consistentes, y en segundo lugar, reducir aún más el número de patrones. En la Tabla 6.3 se muestran los resultados obtenidos en este tercer experimento.

Patrón	Nº patrones	Posprocesado	Sensib.	Especif.	Precisión	AUC
Emergente	5957	Ninguno	94.97 %	48.28 %	86.24 %	0.705
Emergente	5440	Cerrado	94.71 %	49.43 %	86.24 %	0.704
Emergente	4702	Maximal	93.65 %	50.57 %	85.59 %	0.709
Emergente	4135	Minimal	94.97 %	49.43 %	86.45 %	0.718
JEP	2223	Ninguno	100 %	54.02 %	91.40 %	0.751
JEP	1970	Cerrado	100 %	51.72 %	90.97 %	0.757
JEP	1850	Maximal	100 %	56.32 %	91.83 %	0.784
JEP	1757	Minimal	100 %	59.77 %	92.47 %	0.797

Tabla 6.2: Experimento 2: Resultados de la clasificación JRIP en WEKA con Pasos 2 y 4 y sin restringir el desbalanceo.

Patrón	Nº patrones	Posprocesado	Sensib.	Especif.	Precisión	AUC
Emergente	5151	Ninguno	96.03 %	55.17 %	88.39 %	0.775
Emergente	4663	Cerrado	97.62 %	51.72 %	89.03 %	0.750
Emergente	4113	Maximal	94.71 %	42.53 %	84.95 %	0.682
Emergente	3706	Minimal	95.50 %	52.87 %	87.53 %	0.735
JEP	2128	Ninguno	100 %	50.57 %	90.75 %	0.747
JEP	1876	Cerrado	100 %	47.13 %	90.11 %	0.733
JEP	1768	Maximal	100 %	45.98 %	89.89 %	0.721
JEP	1685	Minimal	100 %	55.17 %	91.61 %	0.766

Tabla 6.3: Experimento 3: Resultados de la clasificación JRIP en WEKA con Pasos 2 y 4 y con desbalanceo restringido al 20 %.

6.4. Discusión

A modo de discusión general, con respecto al Experimento 1 de referencia, donde no se incluía la detección de patrones consistentes, se puede observar que se obtiene en general una mejor precisión cuando se restringe el nivel de desbalanceo al 20 % y se utilizan patrones emergentes, y en cambio empeoran los resultados cuando se utilizan patrones JEP. Sin embargo, el número de patrones JEP es mucho más reducido, siendo aproximadamente un 60 % menor en este tercer experimento con respecto a la situación de partida y habiéndose reducido en general, un 10 % adicional con respecto al segundo experimento.

Asimismo, podemos indicar que los patrones cerrados conservan más información que los maximales y obtienen buenos resultados, sin embargo, los mejores resultados se obtienen con los patrones minimales, que son los más generales y cortos (se puede ampliar información de este posprocesamiento en la Sección 4.3). Además, se ha comprobado que todos los patrones minimales generados en los diferentes experimentos abarcan desde un mínimo de 2 días hasta los 5 días que disponemos, no perdiendo por tanto expresividad ni significación clínica. Si todos los patrones fueran cortos, y abarcaran solamente un par de días, su utilidad sería más limitada cuando se quiere analizar el contenido de los patrones.

En nuestro caso, no se ha realizado una selección de variables clásica, con medidas basadas en correlación o teoría de la información por diversos motivos. El más importante, es que aunque muchos de los patrones están correlacionados, incluso únicamente cambiando el orden entre dos ítems, la interpretación clínica de la evolución de los pacientes es distinta. Al usar sólo variaciones del soporte como propiedad fundamental conseguimos eliminar patrones que siguen representando el máximo posible de información de los patrones frecuentes. En cualquier caso, finalmente los algoritmos de clasificación realizarán una selección reducida de patrones que estarán en el modelo.

Respecto a la complejidad de los modelos, son necesarias aproximadamente 14 reglas para poder generar el clasificador en los diferentes experimentos, no habiendo diferencias significativas en el número de reglas que se generan en los diferentes experimentos. El hecho de contener pocas reglas hace que puedan ser interpretados por los clínicos.

El proceso propuesto en este trabajo es genérico pudiendo realizarse para k-particiones en vez de sólo para dos. No obstante, la principal limitación de los experimentos viene dada por el reducido tamaño de la base de datos, que no permite realizar un mayor número de particiones con las que comprobar la consistencia. Una alternativa sería realizar una validación con remuestreo, similar a bootstrap, pero el alto coste computacional de los algoritmos de extracción de patrones secuenciales hace que esta validación sea prohibitiva.

6.5. Conclusiones

En este capítulo se ha propuesto la mejora del proceso de descubrimiento del conocimiento para construir un modelo de clasificación mediante patrones secuenciales multivariantes para predecir la mortalidad de los pacientes quemados de una base de datos de la unidad de cuidados intensivos, de forma que introduciendo un proceso general de validación de $k - particiones$ estratificadas se ha podido comprobar que:

- Con la selección de los patrones consistentes se ha logrado reducir significativamente el número de patrones usados como predictores en los clasificadores. Así por ejemplo, pasamos de obtener 12512 patrones emergentes y 4931 patrones JEP en el experimento de referencia (sin utilizar representaciones comprimidas), a quedarnos con 5957 patrones emergentes (-52.4 %) y 2223 patrones JEP (-55 %) sin restringir el balanceo. Si se restringe el desbalanceo al 20 %, se consigue aumentar esta reducción, quedándonos 5151 patrones emergentes (-13.5 % adicional) y 2128 patrones JEP (-4.2 % adicional).

- Se han obtenido patrones distribuidos uniformemente en las diferentes particiones y con significación clínica, sin tener que recurrir a métodos de selección de variables que pasan por alto las propiedades de los patrones secuenciales.

- Con los patrones JEP y seleccionando los patrones consistentes sin restringir el desbalanceo entre las clases se obtienen los mejores resultados en la clasificación. De forma que seleccionando patrones minimales, se consigue la mejor precisión del 92.47 % y la mejor especificidad del 59.77 %.

- En general, el uso de patrones minimales mejora los resultados de clasificación con respecto a los patrones cerrados y maximales, además de generar también un conjunto más reducido de patrones. Así por ejemplo, utilizando patrones JEP y seleccionando patrones consistentes sin restringir el desbalanceo, pasamos de obtener 2223 patrones a 1970 patrones cerrados (-11.3 %), 1850 patrones maximales (-6 % adicional) y 1757 patrones minimales (-5 % adicional respecto a los patrones maximales).

Capítulo 7

Clasificación mediante patrones discriminatorios usando la razón de probabilidades diagnóstica (DOR)

En este capítulo, basado en un artículo publicado en *JMIR Medical Informatics* [24], proponemos utilizar la razón de probabilidades diagnóstica (Diagnostic Odds Ratio, DOR) para seleccionar los patrones utilizados en la clasificación en un dominio clínico, en lugar de emplear las propiedades de frecuencia. Esto permite utilizar una terminología más cercana al lenguaje de los médicos, en la que se considera que un patrón tiene un factor de riesgo o un factor de protección. Comparamos cuatro maneras de emplear el DOR para la selección de patrones: 1) Lo usamos como umbral para seleccionar patrones con un DOR mínimo; 2) Seleccionamos patrones cuyos DOR diferenciales son superiores a un umbral en lo que respecta a sus extensiones; 3) Seleccionamos patrones cuyos intervalos de confianza DOR no se superponen; y 4) Proponemos la combinación de umbrales e intervalos de confianza no superpuestos con el fin de seleccionar los patrones más discriminatorios. Como referencia, comparamos nuestras propuestas con los patrones JEP (Jumping Emerging Patterns), una de las técnicas más utilizadas para la selección de patrones que utilizan la propiedad de la frecuencia.

7.1. Introducción

Muchos indicadores cuantitativos específicos del rendimiento de las pruebas diagnósticas se han introducido en el dominio clínico, como la sensibilidad y especificidad, los valores predictivos positivos y negativos, las ratios de probabilidad o el área bajo la curva carac-

terística de funcionamiento del receptor (AUC), entre otros. Pero hay un único indicador de rendimiento diagnóstico, denominado Razón de Probabilidades Diagnóstica (Diagnostic Odds Ratio, DOR), que está estrechamente vinculado a los indicadores existentes, facilita el metaanálisis formal de los estudios sobre el rendimiento de las pruebas diagnósticas y se deriva de modelos logísticos [49].

En este capítulo proponemos y comparamos cuatro enfoques en los que el DOR se utiliza como medida estadística para seleccionar un número reducido de patrones, y planteamos el uso de estos patrones como predictores en un modelo de clasificación. El cálculo del DOR para un patrón nos permite utilizar una terminología más cercana al lenguaje de los médicos, en la que un patrón se considera que tiene un factor de riesgo o un factor de protección.

El primer enfoque consiste en utilizar el DOR como un umbral mínimo con el que seleccionar patrones. En el segundo enfoque, calculamos la diferencia en el DOR de un patrón secuencial con respecto a sus extensiones, y establecemos un umbral para esta diferencia con el fin de reducir el número de patrones seleccionados. Una ventaja de este enfoque es que se podría utilizar como una poda temprana dentro del algoritmo de descubrimiento de patrones. En tercer lugar, calculamos un intervalo de confianza para el DOR, y utilizamos este intervalo de confianza para podar patrones que no son estadísticamente diferentes de sus patrones de extensión. Por último, combinamos el segundo y tercer enfoque para seleccionar patrones con diferentes propiedades.

Hemos verificado que estas propuestas proporcionan resultados aceptables mediante la construcción de un modelo para la clasificación de la supervivencia de los pacientes utilizando su evolución diaria en una Unidad de Quemados Críticos, empleando patrones secuenciales multivariantes. Además, hemos comparado los cuatro enfoques con la selección de patrones utilizando medidas clásicas basadas en la frecuencia, como los patrones JEP (Jumping Emerging Patterns).

Recomendamos la lectura del estado del arte realizada en el Capítulo 3, concretamente sobre la minería de patrones secuenciales en la Sección 3.3.2, de su clasificación en la Sección 3.4, y de cómo mejorar la calidad e interés de estos patrones, mediante los patrones discriminatorios (concretamente los patrones JEP), en la Sección 3.5.3 o las medidas de interés para la clasificación de secuencias dentro de la Sección 3.5.2.

	Test de referencia	
Test	Trastorno objetivo	No trastorno objetivo
Positivo previsto	TP	FP
Negativo previsto	FN	TN

Tabla 7.1: Tabla de contingencia 2x2. Las abreviaturas TP, FP, FN y TN denotan el número de, respectivamente, verdaderos positivos (True Positives), falsos positivos (False Positives), falsos negativos (False Negatives) y verdaderos negativos (True Negatives).

7.2. Razón de probabilidades diagnóstica e intervalo de confianza

Los médicos deben basarse en la correcta interpretación de los datos diagnósticos en una variedad de entornos clínicos. Una tabla 2x2 es una herramienta esencial con la que presentar los datos relativos a los estudios epidemiológicos para la evaluación de pruebas diagnósticas (véase Tabla 7.1). Los términos comúnmente utilizados con las pruebas diagnósticas son Sensibilidad, Especificidad y Precisión, y miden estadísticamente el rendimiento de la prueba. La sensibilidad indica qué tan bien predice la prueba una categoría (véase Ecuación 7.1) y la especificidad mide qué tan bien predice la prueba la otra categoría (véase Ecuación 7.2), mientras que se espera que la precisión mida qué tan bien predice la prueba ambas categorías.

$$sensibilidad = \frac{TP}{TP + FN} \tag{7.1}$$

$$especificidad = \frac{TN}{TN + FP} \tag{7.2}$$

También se han sugerido otras múltiples pruebas con las que mejorar la toma de decisiones diagnósticas en diferentes situaciones clínicas. Por ejemplo, Glas et al. [49] propuso el uso del DOR como un único indicador del rendimiento diagnóstico.

El DOR se utiliza para medir el poder discriminatorio de una prueba diagnóstica, consistiendo en la razón de las probabilidades de un resultado positivo de la prueba entre los enfermos, y las probabilidades de un resultado positivo de la prueba entre los no enfermos. El DOR no depende de la prevalencia, y puede ser más fácil de entender, ya que es una medida epidemiológica familiar. Se puede expresar en términos de sensibilidad y especificidad (véase Ecuación 7.3).

$$DOR = \frac{\frac{TP}{FN}}{\frac{FP}{TN}} = \frac{\frac{sensibilidad}{1-sensibilidad}}{\frac{1-especificidad}{especificidad}} \tag{7.3}$$

El valor del DOR oscila entre 0 e infinito. Para calcular el DOR, los problemas potenciales que implican división por cero se resuelven agregando 0.5 a las celdas seleccionadas en la tabla de diagnóstico 2x2.

Cuanto mayor sea la razón de probabilidades de 1, más probable es que las personas con la enfermedad estén expuestas en comparación con las personas sin la enfermedad (factor de riesgo). Un valor de 1 significa que una prueba no discrimina entre los pacientes con la enfermedad y aquellos que no la tienen. Los valores inferiores a 1 sugieren un menor riesgo de enfermedad asociada con la exposición (factor de protección).

Los Intervalos de Confianza (IC) para las estimaciones de rango se pueden calcular convencionalmente como se muestra en la Ecuación 7.4, donde Xhm es el Mantel-Haenszel chi-cuadrado y $Z = 1.96$ si se emplea una confianza del 95 %. En la práctica, un IC del 95 % se usa a menudo como indicador de la presencia de significación estadística si no se superpone al valor nulo (DOR=1).

$$IC = DOR^{(1\pm\frac{Z}{Xhm})}, Xhm = \sqrt{\frac{(n-1)(TP\times TN - FP\times FN)^2}{(TP+FP)(FN+TN)(TP+FN)(FP+TN)}} \tag{7.4}$$

Li et al. [73] creó un algoritmo basado en la siguiente suposición: si la agregación de una exposición a una regla no produce un cambio significativo en el DOR, entonces la regla no debe ser notificada. El DOR entre dos reglas es significativamente diferente si sus intervalos de confianza del 95 % no se superponen.

Han aparecido varios trabajos basados en la no superposición del DOR. En [117] los autores discuten las diferencias en el rendimiento obtenidas al extraer reglas con las diferentes definiciones de una población no expuesta, cuando no se utilizan criterios de poda para filtrar reglas redundantes, o cuando se agrega un criterio de poda de reglas redundantes basadas en la superposición del IC 95 %. Confirmaron que la minería sin criterio de poda produce un alto número de reglas redundantes, lo que demuestra la necesidad de un proceso con el que eliminarlas. En [116] los mismos autores explican que las métricas de interés tradicionales de soporte y confianza deben sustituirse por métricas que se centran en las variaciones de riesgo causadas por diferentes exposiciones. Tienen dos criterios de poda posteriores al procesamiento: una regla se poda si su IC 95 % para el DOR cruza el valor de 1 o si no hay superposición del IC 95 % de la regla con todos sus padres.

7.3. Experimentos

Llevamos a cabo los experimentos siguiendo el proceso de descubrimiento de conocimiento de 4 pasos explicado en el Capítulo 4: 1) preprocesamiento, 2) minería, 3) selección

Discre-tización	Soporte patrones	Patrones iniciales Viven+Mueren	Base JEP	Experim. 1 DOR		Experim. 2 Difer. DOR		Experim. 3 No sup. DOR		Experim. 4 Dif+nosup.DOR	
				< .08, > 16	< .04, > 32	todos	mejor	todos	mejor	todos	mejor
Experto	10 %	46,041+83,015	391	2,065	750	2,795	2,359	858	746	236	198
	8 %	88,084+241,866	4,931	14,424	5,798	10,655	8,781	2,195	1,856	701	504
	6 %	224,952+492,504	47,113	51,352	41,059	32,406	26,157	4,545	3,803	1,556	1,293
UCPD	16 %	238,337+49,947	2,179	14,158	2,766	2,401	1,990	1,529	1,415	325	272
	14 %	396,238+68,654	7,556	33,979	7,483	4,153	3,465	2,296	2,052	487	411
	12 %	647,943+137,546	22,940	65,564	16,272	9,907	8,173	6,418	5,228	1,397	1,212

Tabla 7.2: Número de patrones interesantes seleccionados después de la minería en el subconjunto de supervivientes y en el conjunto de no supervivientes, con la discretización Experta y UCPD.

de patrones y 4) clasificación. Para ello se emplean los datos obtenidos de la Unidad de Quemados Críticos descritos en la Sección 1.2.4.

En el primer paso, el preprocesamiento se llevó a cabo mediante el empleo de dos métodos de discretización diferentes para los atributos continuos. Un método fue la discretización de atributos realizada por un experto (se pueden consultar los intervalos establecidos para cada variable en la Sección 1.2.4.1). Este método proporcionó a los patrones una mayor interpretabilidad, ya que se expresan en el lenguaje clínico. El otro método es la discretización de preservación de correlación no supervisada (Unsupervised Correlation Preserving Discretization, UCPD), ya que proporcionó la mejor clasificación en comparación con varios algoritmos de discretización automática (véase Capítulo 5).

En el segundo paso, usamos el algoritmo FaSPIP [50] para descubrir patrones secuenciales multivariantes. Consideramos los soportes para minar patrones que van desde 16 % al 6 % con el fin de encontrar el mayor soporte que genera el menor número de patrones con los mejores resultados de clasificación. Esto, por lo tanto, nos permitió obtener desde un pequeño número de patrones interesantes a miles de ellos (consulte la Tabla 7.2). Los mejores resultados no se produjeron con los soportes más bajos, lo que parece implicar que no hay un sobreajuste.

El tercer paso consistió en reducir el número de patrones encontrados para seleccionar únicamente los que serían relevantes para la clasificación. Si el soporte utilizado en el paso anterior es bajo, el número de patrones frecuentes aumenta acusadamente: el fenómeno de la explosión de patrones es una desventaja importante del uso de patrones como predictores para clasificadores.

Decidimos utilizar un experimento de referencia, como base para compararlo con nuestros métodos propuestos. Por lo tanto, empleamos la propiedad de la frecuencia (debido a que es regularmente utilizada para medir el grado de interés) para seleccionar patrones discriminatorios. Con este fin, seleccionamos sólo patrones JEP (Jumping Emerging Patterns) que no son comunes en el subconjunto de no sobrevivientes y sobrevivientes, lo que nos permite

Discretización	Long. patrón	Base JEP	Experim. 1a DOR (< 0.08, > 16)	Experim. 1b DOR (< 0.04, > 32)	Experim. 2 Diferencia DOR	Experim. 3 No Sup. DOR	Experim. 4 Dif. + no sup. DOR
Experto	2	0	5 (0.0 %)	0	289 (2.7 %)	76 (3.5 %)	39 (5.6 %)
	3	41 (0.8 %)	187 (1.3 %)	49 (0.8 %)	2,063 (19.4 %)	461 (21.0 %)	198 (28.2 %)
	4	542 (11.0 %)	1,610 (11.2 %)	552 (9.5 %)	3,912 (36.7 %)	857 (39.0 %)	299 (42.7 %)
	5	1,377 (27.9 %)	4,176 (29.0 %)	1,545 (26.6 %)	3,004 (28.2 %)	612 (27.9 %)	140 (20.0 %)
	6	1,518 (30.8 %)	4,811 (33.4 %)	1,960 (33.8 %)	1,155 (10.8 %)	175 (8.0 %)	23 (3.3 %)
	7	987 (20.0 %)	2,698 (18.7 %)	1,190 (20.5 %)	212 (2 %)	14 (0.6 %)	2 (0.3 %)
	8	372 (7.5 %)	785 (5.4 %)	407 (7.0 %)	20 (0.2 %)	0	0
	9	84 (1.7 %)	139 (1.0 %)	85 (1.5 %)	0	0	0
	10	10 (0.2 %)	13 (0.1 %)	10 (0.2 %)	0	0	0
TOTAL		4,931	14,424	5,798	10,655	2,195	701

Tabla 7.3: Número (y porcentaje) de patrones interesantes por longitud (de 2 a 10), para un soporte del 8 % con discretización experta y selección de todos los patrones cuando sea posible

eliminar el comportamiento común o la evolución de un paciente que no es discriminatoria.

Por último, el cuarto paso consistió en construir un modelo de clasificación con la restricción de que tenía que ser interpretable. Queríamos obtener un modelo con un pequeño número de patrones que fuera fácil de interpretar para el médico. En este caso, usamos un clasificador basado en reglas o en árboles de decisión.

Por un lado, usamos JRIP (la implementación de RIPPER en WEKA) para el aprendizaje de reglas. Por otro lado, elegimos el árbol de decisiones J48 implementado por WEKA para el algoritmo C4.5. Se pueden ampliar los motivos de la elección de estos clasificadores consultando la Sección 4.3.4.

En todos los casos, configuramos los clasificadores con el mismo número mínimo de elementos en cada hoja o regla al 2 % de las instancias. La precisión, sensibilidad, especificidad y AUC se calcularon mediante una validación cruzada de 10 iteraciones.

A continuación, se muestran los resultados del experimento de referencia y los resultados de nuestras cuatro propuestas diferentes utilizando el DOR. El número de patrones generados en el subconjunto de supervivientes y en el conjunto de no supervivientes con diferentes soportes se muestra en la Tabla 7.2. También estudiamos la longitud de los patrones producidos (consulte la Tabla 7.3). De manera general, un patrón corto es más simple y más general (cubre a más pacientes). Por el contrario, un patrón largo es más específico (cubre menos pacientes) y es más difícil de entender. Por lo tanto, es más difícil construir un clasificador con patrones cortos.

En la discusión, exploramos tres aspectos: rendimiento de la clasificación, número y longitud de los patrones seleccionados e interpretabilidad del clasificador.

Clasificador	Discretización	Soporte patrón	Núm. patr.	Total long.	Media long.	Sensib.	Especif.	Precisión	AUC
J48	Experto	10 %	7	33	4.71	100.00 %	43.68 %	89.46 %	0.709
		8 %	17	84	4.94	**100.00 %**	**56.32 %**	**91.83 %**	**0.782**
		6 %	16	80	5	100.00 %	44.83 %	89.68 %	0.720
	UCPD	16 %	8	29	3.63	100.00 %	52.87 %	91.18 %	0.763
		14 %	**10**	**37**	**3.7**	**100.00 %**	**66.67 %**	**93.76 %**	**0.853**
		12 %	12	48	4	100.00 %	59.77 %	92.47 %	0.796
JRIP	Experto	10 %	8	37	4.63	100.00 %	40.23 %	88.82 %	0.704
		8 %	15	79	5.27	**100.00 %**	**58.62 %**	**92.26 %**	**0.777**
		6 %	18	87	4.83	100.00 %	44.83 %	89.68 %	0.729
	UCPD	16 %	7	34	4.86	100.00 %	47.13 %	90.11 %	0.711
		14 %	10	35	3.5	**100.00 %**	**73.56 %**	**95.05 %**	**0.866**
		12 %	12	51	4.25	100.00 %	62.07 %	92.90 %	0.833

Tabla 7.4: Resultados del experimento base de referencia con JEP.

7.3.1. Experimento base de referencia: usando JEP

En el experimento de referencia buscamos patrones discriminatorios, una de las técnicas más importantes utilizadas en minería de datos [80], donde los patrones son podados utilizando solamente las propiedades de soporte. Para ello elegimos los Jumping Emerging Patterns (JEP), lo que significa que seleccionamos patrones encontrados sólo en los sobrevivientes y patrones que ocurrieron exclusivamente en los no sobrevivientes. La utilización de este tipo de patrones realizada en el Capítulo 4 mostró que los resultados de las pruebas de clasificación son comparables a las puntuaciones de gravedad de las quemaduras que utilizan actualmente los médicos, produciendo los mejores resultados de clasificación. Además, de esta forma no es necesario establecer un umbral a priori de manera manual, que podría arrojar resultados diferentes.

La Tabla 7.4 muestra los resultados de los experimentos realizados utilizando dos algoritmos de discretización y variando el soporte del patrón.

Como se observa, los patrones JEP permiten alcanzar una sensibilidad del 100 %, pero la especificidad tiene valores más bajos. Esto se debe al hecho de que el conjunto de datos está desbalanceado con una mayoría de sobrevivientes, y los patrones cubren sólo aquellos pacientes que sobrevivirán o aquellos que morirán. Es necesario lograr una mayor especificidad para predecir a los no supervivientes, por lo que la especificidad más alta se resalta en la Tabla 7.4 como mejor resultado de referencia.

La discretización experta es preferida por los médicos, ya que se basa principalmente en valores de rangos de referencia. Pero hay que tener en cuenta que es posible mejorar los resultados mediante el uso de una discretización automática, como UCPD (tal y como se

mostró en el Capítulo 5).

Cuando se utiliza la discretización experta, la especificidad más alta (58,62 %) se obtiene utilizando el clasificador JRIP con un soporte del 8 %.

Este clasificador requiere de 15 patrones, con una longitud total de 79 ítems y una longitud media de 5.27 ítems por patrón. Como ejemplo, mostramos uno de estos 15 patrones utilizados por el clasificador que se encuentra en el subconjunto de no supervivientes. Para cada variable, el subíndice $_i$ marca el intervalo de discretización i donde para $i = 0$ es el intervalo más bajo:

$$< BAL_4 < BIC_1 < DIUR_2 < BE_0 \text{ (10 defunciones, 0 supervivientes)}$$

También hay un patrón interesante que aparece en los cinco experimentos para el subconjunto de no sobrevivientes:

$$< DIUR_3 < INC_0 < INC_0 < DIUR_3 \text{ (10 defunciones, 0 supervivientes)}$$

Por lo tanto, sería posible interpretar este patrón como ”un paciente morirá si su diuresis es muy alta un día, y durante los próximos dos días hay un ingreso de líquidos bajo con una diuresis muy alta al día siguiente”.

7.3.2. Experimento 1: utilizando el DOR

En este experimento, calculamos el DOR para cada patrón como se muestra en la Sección 7.2. En el lenguaje clínico, un $DOR > 1$ implica que la exposición al patrón es un factor de riesgo. Por el contrario, un $DOR < 1$ implica que el patrón es un factor de protección y la selección de un DOR con un valor muy bajo sugiere por lo tanto un menor riesgo de enfermedad asociada con la exposición. Un valor de $DOR = 1$ significa que el patrón no discrimina entre los pacientes con la afección y aquellos que no la tienen.

Por lo tanto, la selección de patrones con un valor alto o un valor bajo para el DOR generará patrones discriminatorios. Es necesario establecer manualmente un umbral para el valor del DOR para elegir los patrones. Hemos llevado a cabo dos experimentos. En el primer experimento (1a), hemos seleccionado los patrones con un valor DOR superior a 16 o inferior a 0.08, y en el segundo experimento (1b), hemos escogido valores más exigentes, que eran el doble o la mitad del valor DOR anteriores, es decir, con un valor DOR superior a 32 o inferior a 0.04. Esto nos permitió reducir el número de patrones (consulte la Tabla 7.2) y obtuvimos una serie de patrones en este experimento (1b) que fue similar a los alcanzados en el experimento de referencia. En estos experimentos utilizando el DOR, la longitud mayoritaria de los patrones seleccionados es de 6 ítems (consulte la Tabla 7.3), que vuelve a ser igual al experimento base de referencia.

La Tabla 7.5 muestra el rendimiento de clasificación de los dos experimentos utilizando los métodos de discretización Experto y UCPD con diferentes soportes de patrón. La discretización experta permite obtener una mayor especificidad del 62.07 % utilizando tanto como clasificador J48 (experimento 1a) como JRIP (experimento 1b), frente a la especificidad del 56.32 % con J48 o del 58.62 % con JRIP del experimento anterior de referencia (consulte la Tabla 7.4), y peores resultados que cuando se utiliza la discretización UCPD.

Clasificador	Discretización	Soporte patrón	Núm. patr.	Total long.	Media long.	Sensib.	Especif.	Precisión	AUC
J48	Experto	10 %	**13**	67	5.15	**90.21 %**	**62.07 %**	**84.95 %**	**0.766**
		8 %	18	89	4.94	88.62 %	58.62 %	83.01 %	0.759
		6 %	16	80	5	91.80 %	47.13 %	83.44 %	0.702
	UCPD	16 %	8	29	3.62	100.00 %	52.87 %	91.18 %	0.763
		14 %	11	43	3.91	**100.00 %**	**62.07 %**	**92.90 %**	**0.787**
		12 %	12	48	4	100.00 %	59.77 %	92.47 %	0.796
JRIP	Experto	10 %	**10**	**46**	4.6	**91.27 %**	**55.17 %**	**84.52 %**	**0.716**
		8 %	12	58	4.83	93.12 %	54.02 %	85.81 %	0.720
		6 %	14	67	4.79	94.44 %	52.87 %	86.67 %	0.706
	UCPD	16 %	8	33	4.13	100.00 %	41.38 %	89.03 %	0.716
		14 %	12	47	3.92	**100.00 %**	**62.07 %**	**92.90 %**	**0.828**
		12 %	12	46	3.83	100.00 %	59.77 %	92.47 %	0.816

(a) DOR ($< 0.08, > 16$).

Clasificador	Discretización	Soporte patrón	Núm. patr.	Total long.	Media long.	Sensib.	Especif.	Precisión	AUC
J48	Experto	10 %	10	49	4.9	93.65 %	50.57 %	85.59 %	0.710
		8 %	17	84	4.94	**94.18 %**	**55.17 %**	**86.88 %**	**0.767**
		6 %	16	80	5	95.50 %	37.93 %	84.73 %	0.656
	UCPD	16 %	8	29	3.62	100.00 %	52.87 %	91.18 %	0.763
		14 %	11	43	3.91	**100.00 %**	**62.07 %**	**92.90 %**	**0.787**
		12 %	12	48	4	100.00 %	59.77 %	92.47 %	0.796
JRIP	Experto	10 %	11	50	4.55	97.09 %	44.83 %	87.31 %	0.704
		8 %	14	67	4.79	**95.50 %**	**62.07 %**	**89.25 %**	**0.801**
		6 %	16	87	5.44	98.15 %	48.28 %	88.82 %	0.715
	UCPD	16 %	7	26	3.71	100.00 %	47.13 %	90.11 %	0.727
		14 %	11	45	4.09	**100.00 %**	**60.92 %**	**92.69 %**	**0.792**
		12 %	14	55	3.93	100.00 %	60.92 %	92.69 %	0.822

(b) DOR ($< 0.04, > 32$).

Tabla 7.5: Resultados de los Experimentos 1a y 1b usando DOR.

Si elegimos la discretización experta, con un clasificador JRIP y los valores más altos del DOR (véase Tabla 7.5b), obtenemos tal y como hemos dicho anteriormente una especificidad mayor que con JEP (62.07 %), pero una sensibilidad más baja (95.50 %). Esto se puede

explicar de la siguiente manera: si miramos alguno de los 14 patrones utilizados en este clasificador utilizando DOR, podemos encontrar un ejemplo de un patrón corto con solo 3 elementos:

$$BIC_1 < BAL_4 < PH_1 \text{ (72.30 DOR) (14 defunciones, 1 superviviente)}$$

Este patrón, con un valor DOR de 72.30, en general servirá para clasificar a un grupo de pacientes que morirán, aunque con errores mínimos, ya que un paciente sobrevive.

Si observamos el patrón que se selecciona en todos los experimentos ($DIUR_3 < INC_0 < INC_0 < DIUR_3$), se incluye también en este experimento porque tiene un valor DOR de 98.05, y es necesario recordar que todos los pacientes en este patrón morirán (10 defunciones, 0 supervivientes). Este tipo de patrón JEP por lo tanto produce una buena especificidad, y un 100 % de sensibilidad (no hay errores de clasificación).

7.3.3. Experimento 2: utilizando el diferencial del DOR entre un patrón y sus extensiones

Un patrón secuencial p_i, de una longitud específica (l), en un punto en el tiempo (t), tiene un valor DOR, $DOR(p_i)$. En cada extensión de este patrón (l+1), que podría ser una S-extensión (en la próxima vez, t+1) o una I-extensión (al mismo tiempo, t) (véase la representación de los patrones temporales en la Sección 3.3.2), habrá n patrones $(p_{i1}, p_{i2}, \ldots, p_{in})$ que son hijos de super-patrón p_i con diferentes valores DOR, $DOR(p_{ij})$, $j \in [1, n]$. En este experimento, elegimos sólo los patrones que tienen una diferencia de valor DOR entre el super-patrón y sus extensiones más alto que un umbral γ, es decir $DOR(p_i) - DOR(p_{ij}) > \gamma$.

Para una mejor interpretación del DOR, dado un patrón p_i, calculamos la probabilidad del factor de riesgo $R(p_i)$ y la probabilidad del factor de protección $P(p_i)$ como se muestra en las ecuaciones 7.5 y 7.6.

$$R(p_i) = DOR(p_i)/(DOR(p_i) + 1) \tag{7.5}$$

$$P(p_i) = 1 - R(p_i) \tag{7.6}$$

En nuestro experimento, por lo tanto, seleccionamos los patrones con dos condiciones: a) cuando la diferencia entre la probabilidad del factor de riesgo $R(p_i)$ es mayor que el 25 % o b) cuando la diferencia entre la probabilidad del factor de protección $P(p_i)$ es mayor que un 30 %. Elegimos un valor umbral más bajo para $R(p_i)$ porque queremos obtener una especificidad más alta minando más patrones que sean representativos de los no sobrevivientes. En

este experimento obtuvimos patrones con una alta calidad que produjeron grandes cambios en la evolución de los pacientes.

Además, utilizamos dos estrategias alternativas para seleccionar patrones: manteniendo todas las extensiones con una diferencia en el valor DOR que es superior a un umbral o bien explorando solamente las extensiones con una búsqueda del mejor primero (best-first search), en cuyo caso seleccionamos sólo la extensión más prometedora con la diferencia DOR más alta entre todas las extensiones. Las Tablas 7.6a y 7.6b muestran los resultados alcanzados utilizando ambas estrategias.

Con respecto al número de patrones seleccionados (consulte la Tabla 7.2), cuando hemos elegido la mejor extensión, solo hemos reducido el número total de patrones en menos de un tercio porque la mayoría de los patrones solo tienen una o dos extensiones.

Si estudiamos la longitud de los patrones (consulte la Tabla 7.3), mientras que en los experimentos anteriores la longitud de los patrones es de aproximadamente de 6 ítems, en este experimento (y en los que siguen) la mayoría de los patrones tienen una longitud de alrededor de 4 ítems, y es posible encontrar bastantes más patrones con una longitud más corta. Hay que tener en cuenta que la distribución de patrones por longitud ha cambiado. En este experimento tenemos patrones más generales que son más cortos. Esto produce peores resultados de clasificación cuando utilizamos la discretización experta con un clasificador JRIP. Es bien sabido que la discretización experta generalmente funciona peor ya que no se basa en una teoría estadística o de la información que ha sido diseñada específicamente para fines de clasificación [22]. Esto también ocurre en casi todos los siguientes experimentos.

Sin embargo, los resultados obtenidos con la discretización UCPD son similares, e incluso con la clasificación JRIP y la búsqueda del mejor primero, necesitamos el menor número de ítems y de patrones de todos los experimentos: sólo se requieren 5 patrones con una longitud total de 20 ítems para alcanzar la especificidad del 56.32 %.

El árbol de clasificación J48 utilizado para clasificar con discretización experta y soporte del 8 %, utilizando la búsqueda del mejor primero para la mejor extensión de patrón, permite alcanzar una especificidad del 62.07 % y requiere de 21 patrones, con una longitud media de 4.19 ítems por patrón. Este promedio es el valor más bajo de todos los experimentos realizados utilizando el clasificador J48 con discretización experta. Dentro de estos 21 patrones, podemos encontrar dos patrones con solo dos ítems, que se utilizan para clasificar a los sobrevivientes:

$$DIUR_3 < BE_2 \text{ (40.23 \% PROTECCIÓN) (43 defunciones, 150 supervivientes)}$$
$$INC_2 = PH_3 \text{ (43.58 \% PROTECCIÓN) (35 defunciones, 176 supervivientes)}$$

El primero, $DIUR_3 < BE_2$, es interesante porque si tiene la siguiente extensión: $DIUR_3 <$

Clasificador	Discretización	Soporte patrón	Núm. patr.	Total long.	Media long.	Sensib.	Especif.	Precisión	AUC
J48	Experto	10 %	28	100	3.57	89.42 %	49.43 %	81.94 %	0.662
		8 %	21	89	4.24	**86.51 %**	**62.07 %**	**81.94 %**	**0.773**
		6 %	18	84	4.67	96.30 %	44.83 %	86.67 %	0.694
	UCPD	16 %	21	81	3.86	93.65 %	49.43 %	85.38 %	0.677
		14 %	15	56	3.73	94.97 %	56.32 %	87.74 %	0.759
		12 %	12	52	4.33	**100.00 %**	**58.62 %**	**92.26 %**	**0.788**
JRIP	Experto	10 %	4	13	3.25	90.74 %	31.03 %	79.57 %	0.620
		8 %	8	25	3.13	86.77 %	29.89 %	76.13 %	0.600
		6 %	3	7	2.33	89.68 %	29.89 %	78.49 %	0.594
	UCPD	16 %	10	37	3.70	92.86 %	24.14 %	80.00 %	0.594
		14 %	11	41	3.73	94.18 %	33.33 %	82.80 %	0.674
		12 %	8	26	**3.25**	**96.03 %**	**62.07 %**	**89.68 %**	**0.831**

(a) Manteniendo todas las extensiones de los patrones.

Clasificador	Discretización	Soporte patrón	Núm. patr.	Total long.	Media long.	Sensib.	Especif.	Precisión	AUC
J48	Experto	10 %	20	73	3.65	89.15 %	44.83 %	80.86 %	0.642
		8 %	21	88	**4.19**	**87.57 %**	**62.07 %**	**82.80 %**	**0.783**
		6 %	18	84	4.67	97.35 %	43.68 %	87.31 %	0.710
	UCPD	16 %	21	81	3.86	93.65 %	49.43 %	85.38 %	0.675
		14 %	15	56	3.73	94.71 %	56.32 %	87.53 %	0.760
		12 %	12	52	4.33	**100.00 %**	**57.47 %**	**92.04 %**	**0.764**
JRIP	Experto	10 %	18	59	3.28	89.15 %	27.59 %	77.63 %	0.582
		8 %	5	17	3.4	90.48 %	21.84 %	77.63 %	0.569
		6 %	8	29	3.62	91.53 %	31.03 %	80.22 %	0.623
	UCPD	16 %	9	31	3.44	91.01 %	28.74 %	79.35 %	0.618
		14 %	19	71	3.74	94.18 %	34.48 %	83.01 %	0.683
		12 %	**5**	**20**	4	**97.09 %**	**56.32 %**	**89.46 %**	**0.767**

(b) Usando la búsqueda del mejor primero para seleccionar la mejor extensión del patrón.

Tabla 7.6: Resultados del Experimento 2 usando el diferencial del DOR.

$BE_2 < PH_4$ (78.85 % PROTECCIÓN)(5 defunciones, 70 supervivientes), donde el pH es muy alto al día siguiente, la tasa de supervivencia del paciente aumenta un 38.62 %.

Además, hemos descubierto un patrón con el que clasificar los no supervivientes que también se puede encontrar en los clasificadores de árbol de decisión tipo J48 de los experimentos posteriores, y que no se seleccionó en los algoritmos de clasificación utilizados en los experimentos anteriores:

$$p_{i1} = BIC_1 < BIC_2 < PH_1 \text{ (98.87 \% RIESGO)(9 defunciones, 0 supervivientes)}$$

Este patrón tiene un valor DOR de $DOR(p_{i1}) = 87.12$, con una probabilidad de riesgo de $R(p_{i1}) = 98.87\,\%$. Ha sido seleccionado porque su super-patrón $p_i = BIC_1 < BIC_2$ (44

defunciones, 111 supervivientes) tiene un valor DOR de $DOR(p_i) = 2.46$, con una probabilidad de riesgo de $R(p_i) = 71.1\,\%$. Esto significa que hay un aumento en el riesgo del $R(p_{i1}) - R(p_i) = 27.77\,\%$, que es superior al umbral fijado del 25 %.

7.3.4. Experimento 3: utilizando la no superposición del Intervalo de Confianza (IC) del DOR

En este experimento, hemos seleccionado patrones basados en la no superposición del 95 % del IC del DOR (como se indica en [73]). Además, solo se han incluido en la salida patrones cuyo IC no incluye el valor 1 (como se realizó en [116]). Todos los patrones son, por lo tanto, un factor protector o un factor de riesgo, pero no ambos o indeterminados.

La Tabla 7.7a muestra los resultados obtenidos cuando mantenemos todas las extensiones de patrón, mientras que la Tabla 7.7b muestra los resultados obtenidos cuando solo se elige la mejor extensión del patrón mediante la búsqueda del mejor primero.

También obtenemos un número reducido de patrones con respecto al experimento anterior (consulte la Tabla 7.2), y una ventaja de este experimento es que este número no depende de un valor de umbral que se tenga que establecer.

En general, el rendimiento de la clasificación es similar al de los experimentos anteriores, aunque con la clasificación JRIP utilizando la discretización experta, obtenemos mejores resultados al seleccionar solo el mejor hijo.

El árbol de clasificación J48 utilizado para clasificar con discretización experta, y un soporte del 8 %, utilizando la búsqueda del mejor primero para seleccionar solamente la mejor extensión de patrón, nos permite obtener una especificidad del 58,62 % y una sensibilidad mayor que el experimento anterior: se requieren 16 patrones.

Uno de los patrones más cortos que encontramos en el árbol de clasificación J48 es:

$$PH_4 < PH_4 < BE_1 \text{ (6 defunciones, 1 superviviente)}$$

Este patrón tiene un valor DOR de 27.93 en el intervalo (6.71, 116.26). Su super-patrón $PH_4 < PH_4$ (14 defunciones, 109 supervivientes) tiene un valor DOR de 0.47 en el intervalo (0.26, 0.87). Hay que tener en cuenta que el IC de estos dos patrones no se superpone.

7.3.5. Experimento 4: utilizando el diferencial del DOR con la no superposición del IC

La última propuesta consiste en utilizar los dos enfoques anteriores juntos (Experimentos 2 y 3), lo que significa que podamos los patrones basados en la superposición del IC del DOR,

Clasificador	Discretización	Soporte patrón	Núm. patr.	Total long.	Media long.	Sensib.	Especif.	Precisión	AUC
J48	Experto	10 %	10	41	4.1	93.92 %	48.28 %	85.38 %	0.721
		8 %	16	77	4.81	**94.97 %**	**58.62 %**	**88.17 %**	**0.741**
		6 %	18	90	5	96.56 %	56.32 %	89.03 %	0.768
	UCPD	16 %	18	70	3.89	97.35 %	57.47 %	89.89 %	0.794
		14 %	11	43	3.91	**99.74 %**	**62.07 %**	**92.69 %**	**0.803**
		12 %	11	47	4.27	100.00 %	57.47 %	92.04 %	0.786
JRIP	Experto	10 %	11	37	3.36	93.65 %	41.38 %	83.87 %	0.682
		8 %	13	60	4.62	91.80 %	33.33 %	80.86 %	0.641
		6 %	7	30	4.29	96.56 %	42.53 %	86.45 %	0.722
	UCPD	16 %	6	23	3.83	96.30 %	41.38 %	86.02 %	0.727
		14 %	9	33	3.67	98.94 %	56.32 %	90.97 %	0.803
		12 %	14	58	4.14	**96.30 %**	**60.92 %**	**89.68 %**	**0.793**

(a) Manteniendo todas las extensiones de los patrones.

Clasificador	Discretización	Soporte patrón	Núm. patr.	Total long.	Media long.	Sensib.	Especif.	Precisión	AUC
J48	Experto	10 %	10	41	4.1	94.18 %	51.72 %	86.24 %	0.742
		8 %	16	77	4.81	**94.71 %**	**58.62 %**	**87.96 %**	**0.739**
		6 %	18	90	5	96.83 %	55.17 %	89.03 %	0.758
	UCPD	16 %	16	68	4.25	96.30 %	55.17 %	88.60 %	0.798
		14 %	13	51	3.92	**100.00 %**	**62.07 %**	**92.90 %**	**0.795**
		12 %	11	45	4.09	100.00 %	60.92 %	92.69 %	0.812
JRIP	Experto	10 %	6	20	3.33	94.44 %	48.28 %	85.81 %	0.735
		8 %	16	62	3.88	95.24 %	41.38 %	85.16 %	0.700
		6 %	12	51	**4.25**	**95.77 %**	**52.87 %**	**87.74 %**	**0.747**
	UCPD	16 %	16	66	4.13	95.50 %	40.23 %	85.16 %	0.695
		14 %	12	44	3.67	97.88 %	54.02 %	89.68 %	0.747
		12 %	15	60	4	**99.21 %**	**55.17 %**	**90.97 %**	**0.788**

(b) Usando una búsqueda del mejor primero para seleccionar unicamente la mejor extensión del patron.

Tabla 7.7: Resultados del Experimento 3 utilizando la no superposición del IC del DOR.

y también la diferencia entre la probabilidad del factor de riesgo (o protección). En ambos casos mantenemos los mismos umbrales.

En este experimento hemos reducido sustancialmente el número de patrones generados (consulte la Tabla 7.2). Por ejemplo, en el caso de discretización experta y un soporte del 8 % (manteniendo todas las extensiones de patrón), obtenemos sólo 701 patrones con este experimento, que es una disminución del 68 % respecto al Experimento 3 (no superposición del DOR, con 2195 patrones) y una disminución del 85.8 % con respecto al Experimento base de referencia (con 4931 patrones).

Es necesario tener en cuenta que si el número de patrones es demasiado bajo, normalmen-

te no logramos un buen resultado de clasificación. Pero con este experimento, por ejemplo, con un 8 % de soporte, discretización experta y el clasificador J48, con solo 504 patrones, hemos obtenido un resultado similar a los anteriores, utilizando solo 13 patrones en el clasificador, con una sensibilidad del 96.30 % y una especificidad del 57.47 % en la búsqueda del mejor primero para obtener la mejor extensión de patrón (consulte la Tabla 7.8). Este es el número más bajo de patrones requeridos para la discretización experta, utilizando como clasificador J48, con una longitud total de sólo 55 ítems.

El rendimiento de la clasificación es similar al de los experimentos anteriores (consulte la Tabla 7.8a y la Tabla 7.8b).

Ahora analicemos un patrón seleccionado en este experimento y en todos los experimentos anteriores $DIUR_3 < INC_0 < INC_0 < DIUR_3$ (10 defunciones, 0 supervivientes). Tiene un valor DOR de 98.05 en el intervalo (24.21, 397.18), con una probabilidad de riesgo del 98.99 %. Su super-patrón $DIUR_3 < INC_0 < INC_0$ tiene un valor DOR de 2.07 en el intervalo (1.20, 3.57) con una probabilidad de riesgo de 67.39 %, lo que significa que no hay superposición en el IC, y que además existe un aumento en la probabilidad de riesgo del 31.6 %.

7.4. Discusión

Hemos propuesto diferentes formas de utilizar el DOR como un simple indicador de rendimiento diagnóstico, realizando una clasificación de la supervivencia de los pacientes en una Unidad de Quemados Críticos mediante el estudio de su evolución diaria utilizando patrones secuenciales multivariantes. Ahora discutimos los factores que tenemos que considerar para obtener un compromiso principalmente entre la interpretabilidad y el rendimiento de la clasificación.

En relación con la interpretabilidad, un modelo es más interpretable que otro modelo si sus decisiones son más fáciles de comprender para un ser humano que las decisiones del otro modelo. En este sentido, los métodos presentados muestran tres ventajas: 1) la legibilidad e interpretabilidad del contenido de los patrones, 2) la reducida longitud de los patrones, y 3) el pequeño conjunto de patrones significativos seleccionados para construir el clasificador.

De estas tres ventajas, la más directa para el médico es que los propios patrones tienen una interpretación en un lenguaje fácilmente entendible en lenguaje clínico, de forma que el médico no tiene que perder tiempo buscando una correspondencia entre lo que lee en el patrón y su manera habitual de trabajar. Además, la definición de los patrones proporciona no solo información estática del paciente en el momento del ingreso, como es habitual, sino que los patrones también proporcionan la evolución del paciente. Por ejemplo, un patrón como

Clasificador	Discretización	Soporte patrón	Núm. patr.	Total long.	Media long.	Sensib.	Especif.	Precisión	AUC
J48	Experto	10 %	13	42	3.23	94.18 %	44.83 %	84.95 %	0.672
		8 %	**13**	**55**	4.23	**95.50 %**	**55.17 %**	**87.96 %**	**0.743**
		6 %	17	78	4.59	97.88 %	47.13 %	88.39 %	0.711
	UCPD	16 %	20	74	3.7	94.97 %	50.57 %	86.67 %	0.761
		14 %	7	28	4	98.41 %	58.62 %	90.97 %	0.804
		12 %	12	50	4.17	**100.00 %**	**65.52 %**	**93.55 %**	**0.820**
JRIP	Experto	10 %	4	13	3.25	93.12 %	29.89 %	81.29 %	0.622
		8 %	12	40	3.33	94.44 %	29.89 %	82.37 %	0.625
		6 %	20	74	3.7	91.80 %	39.08 %	81.94 %	0.668
	UCPD	16 %	7	24	3.43	94.44 %	27.59 %	81.94 %	0.632
		14 %	6	23	3.83	97.35 %	32.18 %	85.16 %	0.653
		12 %	16	63	3.94	**98.68 %**	**59.77 %**	**91.40 %**	**0.795**

(a) Manteniendo todas las extensiones de los patrones.

Clasificador	Discretización	Soporte patrón	Núm. patr.	Total long.	Media long.	Sensib.	Especif.	Precisión	AUC
J48	Experto	10 %	10	35	3.5	95.50 %	41.38 %	85.38 %	0.694
		8 %	**13**	**55**	4.23	**96.30 %**	**57.47 %**	**89.03 %**	**0.770**
		6 %	16	75	4.69	98.41 %	50.57 %	89.46 %	0.739
	UCPD	16 %	20	74	3.7	93.92 %	50.57 %	85.81 %	0.758
		14 %	7	28	4	96.83 %	58.62 %	89.68 %	0.808
		12 %	12	50	4.17	**100.00 %**	**59.77 %**	**92.47 %**	**0.812**
JRIP	Experto	10 %	6	21	3.5	92.59 %	25.29 %	80.00 %	0.597
		8 %	14	43	3.07	91.80 %	29.89 %	80.22 %	0.614
		6 %	15	57	3.8	92.59 %	29.89 %	80.86 %	0.626
	UCPD	16 %	10	37	3.7	96.83 %	35.63 %	85.38 %	0.671
		14 %	10	36	3.6	98.68 %	32.18 %	86.24 %	0.673
		12 %	15	59	3.93	**98.68 %**	**50.57 %**	**89.68 %**	**0.759**

(b) Usando la búsqueda del mejor primero para seleccionar la mejor extensión del patrón.

Tabla 7.8: Resultados del Experimento 4 usando el diferencial del DOR y la no superposición del IC.

$DIUR_3 < INC_0 < INC_0 < DIUR_3$ conduce al médico a los factores clínicos relacionados con el patrón: una diuresis alta y entradas muy bajas, en cuatro días diferentes.

Respecto a la segunda ventaja, si estudiamos la longitud de los patrones finalmente seleccionados (consulte la Tabla 7.3), se observará que la mayoría de los patrones del experimento base de referencia (mediante JEP) y la mayoría de los patrones del primer experimento (Usando DOR) tienen una longitud de 6 ítems, mientras que la mayoría de los patrones de los experimentos posteriores tienen una longitud de 4 ítems. Podemos observar que la distribución de patrones por longitud ha cambiado, con un mayor número de patrones más cortos en los últimos experimentos. En los posteriores Experimentos 2, 3 y 4 hemos observado que,

por un lado, el clasificador es menos preciso. Por otro lado, los patrones más cortos son más fáciles de entender, más generales y describen bien a la población, pero cubren simultáneamente a sobrevivientes y no sobrevivientes.

En general, estos patrones más cortos producen peores resultados de clasificación cuando utilizamos la discretización experta con un clasificador JRIP. Por un lado, la discretización experta generalmente funciona peor, ya que no se basa en una teoría estadística o de la información que ha sido diseñada específicamente para fines de clasificación, y por otro lado, JRIP proporciona el mejor rendimiento en términos de la complejidad de la estructura del árbol, mientras que J48 produce una alta precisión de clasificación (como explican los autores en [89]). Con patrones más cortos, sin embargo, es más fácil interpretar el significado de los patrones y explicar su comportamiento.

Con relación a la tercera ventaja, podríamos por tanto decir que es preferible un modelo que nos permita conseguir un buen resultado de clasificación con un número reducido de patrones (y en consecuencia de ítems).

En la Tabla 7.2 se puede ver que obtuvimos el menor número de patrones con el Experimento 4 (Utilizando el diferencial del DOR y la no superposición del IC). Estos patrones están simultáneamente restringidos por estas dos condiciones, lo cual genera que se seleccione un pequeño número de patrones, a los que incluso podría ser interesante llevar a cabo una revisión manual y un estudio de los mismos (aunque eso está fuera del alcance de este capítulo).

El experimento base de referencia (usando JEP) y el Experimento 3 (no superposición del IC del DOR) no dependen de un valor umbral y también obtenemos un número razonablemente pequeño de patrones. Sin embargo, dependiendo del valor del umbral que se establezca en los otros experimentos (Experimentos 1, 2 y 4), se producirán cambios en el número de patrones finalmente seleccionados. Por lo tanto, hemos realizado 2 variaciones en el Experimento 1 (usando DOR), restringiendo el valor mínimo del DOR necesario para seleccionar patrones (consulte la Tabla 7.5), lo que significa que hemos sido capaces de reducir significativamente y de manera adecuada el número de patrones seleccionados.

Cuando trabajamos con datos desbalanceados, como es habitual en los dominios médicos, es necesario hacer resaltar la clasificación correcta de la clase minoritaria en comparación con otros casos generales. Por lo tanto, es necesario comprobar la mayor especificidad para elegir el mejor resultado de clasificación, que en nuestros experimentos se produce mediante el uso de la discretización automática UCPD con JEP como una medida discriminatoria clásica basada en frecuencias. Los patrones JEP se han utilizado generalmente para construir clasificadores precisos, mientras que la discretización UCPD explota la estructura de correlación subyacente en los datos para obtener los intervalos discretos y asegurarse de

que se conservan las correlaciones inherentes.

Además, hemos demostrado que esta discretización automática realiza generalmente mejores clasificaciones que la discretización experta. Pero los médicos prefieren usar un rango de referencia de discretización para valores fisiológicos y de laboratorio. Esto significa que, por ejemplo, prefieren utilizar el intervalo (7.35, 7.45) como valor normal para el pH, tal y como normalmente se entiende en medicina. La interpretabilidad de los resultados de la clasificación mediante la discretización experta es, por lo tanto, un factor predominante en nuestra elección.

De esta manera, si consideramos únicamente la discretización del experto, el mejor resultado de clasificación se logra en el Experimento 1b (Usando el DOR), con una especificidad del 62.07 % y un valor AUC de 0.801 (véase la Tabla 7.5b). En este experimento obtuvimos simultáneamente patrones encontrados tanto en los supervivientes como en los no supervivientes basados únicamente en el valor DOR de cada patrón.

El modelo de clasificación que es más fácil de comprender y tiene alta especificidad requiere sólo 5 patrones (con una longitud total de 20 ítems), y se logra con la discretización UCPD y un clasificador JRIP en el Experimento 2b (usando el diferencial del DOR) utilizando la búsqueda del mejor primero para seleccionar la mejor extensión. Obtiene una especificidad del 56.32 % y un valor AUC de 0.767 (véase la Tabla 7.6b). Si tenemos en cuenta sólo la discretización experta, con un clasificador J48 necesitamos al menos 13 patrones (con una longitud total de 55 ítems) para obtener una especificidad del 57.47 % y un valor de AUC de 0.770 (véase Tabla 7.8b) en el Experimento 4b (Usando el diferencial del DOR y la no superposición).

7.5. Conclusiones

En esta investigación, hemos conseguido un modelo con el que predecir la supervivencia de los pacientes teniendo en cuenta dos aspectos: la relevancia de la evolución temporal de los pacientes como parte del modelo y un modelo interpretable para los médicos. Hemos logrado los aspectos anteriores a) utilizando patrones secuenciales multivariantes en modelos de clasificación que pueden ser fácilmente comprendidos por los expertos, b) utilizando un número reducido de patrones, y c) utilizando un lenguaje que es bien conocido por los médicos en lo que respecta tanto a la discretización de valores como a las medidas de interés de los patrones.

La principal contribución de este trabajo es la propuesta y evaluación de cuatro formas de emplear la Razón de Probabilidades Diagnóstica (Diagnostic Odds Ratio, DOR) para reducir el número de patrones y seleccionar sólo los más discriminatorios, ya que la explosión de

patrones es el principal problema de los clasificadores basados en patrones. Hemos comparado estas cuatro propuestas con un experimento base de referencia utilizando patrones JEP (Jumping Emerging Patterns). Esta es, hasta donde sabemos, la primera vez que algunos de estos enfoques han sido propuestos y comparados en la literatura científica.

Las principales conclusiones obtenidas son las siguientes:

- Con respecto a seleccionar el menor número de patrones, la mejor opción es la de usar un DOR diferencial y no superpuesto (como en el Experimento 4). A medida que hemos aumentado las restricciones aplicadas, hemos reducido significativamente el número de patrones, logrando así patrones más generales, simples e interesantes. Con la discretización experta y un soporte del 10 %, sólo se encuentran por ejemplo 198 patrones (usando la búsqueda del mejor primero para seleccionar la mejor extensión), y, muy curiosamente, estos patrones cubren a todos los pacientes que no sobrevivieron. A pesar de no estar dentro del alcance de este capítulo, sería interesante para un médico llevar a cabo una interpretación manual de estos patrones.

- Con respecto a la precisión, los mejores resultados de clasificación se producen utilizando patrones JEP junto con la discretización UCPD. Los JEP se han utilizado ampliamente para construir clasificadores precisos y para producir mejores resultados cuando utilizamos una discretización basada en la teoría estadística o de la información que está específicamente destinada a la clasificación. Sin embargo, como requerimos patrones interpretables, que sean fáciles de entender para el médico, se debe utilizar una discretización por rangos de referencia creada por un experto. Por lo tanto, si consideramos sólo la discretización experta, la mayor especificidad se alcanza utilizando sólo el DOR para seleccionar los patrones (como se ha realizado en el Experimento 1). A pesar de los esfuerzos realizados para reducir la cantidad y la longitud de los patrones en los Experimentos 2, 3 y 4, en los que hemos comparado cada patrón con sus extensiones, el clasificador construido es menos preciso. Los patrones más cortos son más fáciles de entender, más generales, y describen bien a la población, pero al mismo tiempo cubren a sobrevivientes y no sobrevivientes.

- Respecto a la interpretabilidad, podemos observar que la discretización tiene un gran impacto en el rendimiento de la clasificación a expensas de la interpretabilidad, ya que se requieren patrones cada vez más largos. Con la discretización UCPD, solo requerimos 5 patrones (con una longitud total de 20 ítems) para construir un conjunto de reglas y obtener una especificidad del 56.32 % cuando usamos el DOR diferencial (consulte Experimento 2). Con la discretización experta, necesitamos al menos 13 patrones (con

una longitud total de 55 ítems) para obtener una especificidad del 57.47 % utilizando un DOR diferencial y no superpuesto para seleccionar los patrones (ver Experimento 4).

- Proponemos además la posible inclusión de una poda temprana dentro del algoritmo de descubrimiento de patrones, al realizar alguna operación relacionada con el DOR entre un patrón secuencial y sus extensiones. En esta poda se podría integrar la búsqueda del mejor primero (best-first search) con el fin de extraer directamente un menor número de patrones secuenciales para la clasificación. De esta manera, utilizando por ejemplo la discretización experta con un soporte del 8 % se consiguen reducir los patrones seleccionados, en un -17.6 % en el experimento 2 (al pasar de 10655 a 8781), del -15.4 % en el experimento 3 (de 2195 a 1856 patrones) y del -28.1 % en el experimento 4 (de 701 a 504 patrones). Este tipo de búsqueda aporta una mayor especificidad al utilizarse discretización experta en el experimento 3, proporcionando además mejoras puntuales en los otros experimentos.

Capítulo 8

Descubriendo novedosos patrones secuenciales multivariantes (JDORSP) en la evolución clínica temporal de los pacientes usando la razón de probabilidades diagnóstica

En el ámbito de la medicina, el abrumador número de patrones y el valioso pero limitado tiempo de los médicos para validarlos sigue siendo un factor limitante. Los esfuerzos actuales en la definición de medidas de interés de los patrones se centran en reducir el número de patrones y cuantificar su relevancia (utilidad/provecho). Sin embargo, a pesar de que la dimensión temporal desempeña un papel clave en los registros médicos, se han hecho pocos esfuerzos en la extracción de conocimiento temporal sobre la evolución del paciente a partir de patrones secuenciales multivariantes.

En este capítulo, basado en el artículo enviado a *Data Mining and Knowledge Discovery* [23], proponemos un método para extraer Patrones Secuenciales de Salto de la Razón de Probabilidades Diagnóstica (Jumping Diagnostic Odds Ratio Sequential Patterns, JDORSP). El objetivo de este método es seleccionar un pequeño subconjunto de patrones que representen el estado de un paciente con un factor de protección estadísticamente significativo (es decir, un patrón asociado con pacientes que sobreviven) y aquellas extensiones cuya evolución cambia repentinamente el estado clínico del paciente, y los patrones se convierten en un factor de riesgo estadísticamente significativo (es decir, un patrón asociado a un paciente que no sobrevive), o viceversa. También proponemos evaluar el conocimiento temporal obtenido

teniendo en cuenta dos criterios basados en el dominio: sorpresividad y relevancia para el problema clínico. Además, definimos una medida del grado de sorpresividad (SUR), para ordenar los patrones más interesantes.

8.1. Introducción

En la minería de patrones, es común utilizar la significación estadística de un patrón para reducir el gran número de patrones que se generan inicialmente. Muchos de estos patrones son irrelevantes o evidentes, y no proporcionan nuevo conocimiento al experto del dominio. Para aumentar la utilidad, relevancia y aprovechamiento de los patrones descubiertos, se utilizan diferentes medidas de interés para reducir el número de patrones [40].

En el dominio clínico se han introducido un gran número de indicadores cuantitativos específicos del rendimiento de las pruebas, como la especificidad y la sensibilidad, los valores predictivos, las relaciones de verosimilitud, el área bajo la curva característica de funcionamiento del receptor (AUC), y muchos más [96]. Pero hay un único indicador de rendimiento diagnóstico, llamado Razón de Probabilidades Diagnóstica (Diagnostic Odds Ratio, DOR), que está estrechamente relacionado con los indicadores existentes, facilita el metaanálisis formal de estudios sobre el rendimiento de las pruebas diagnósticas, y que se deriva de la regresión logística [49].

En este artículo definimos los JDORSP "Jumping Diagnostic Odds Ratio Sequential Patterns", y mostramos cómo usarlos con el fin de extraer conocimiento temporal sobre la evolución de los pacientes en la Unidad de Quemados Críticos.

Estos nuevos patrones (JDORSP) se generan mediante una nueva forma de usar el DOR como medida para reducir significativamente el número de patrones en el dominio clínico y obtener sólo aquellos patrones secuenciales que proporcionen un conocimiento sólido basado en la definición de los factores de riesgo y protección.

La idea es extraer conocimiento de un pequeño número de patrones secuenciales que representan el estado de un paciente con un factor de protección estadísticamente significativo y cuyas extensiones (o evolución) de repente cambian el estado clínico del paciente, y los patrones se convierten en un factor de riesgo estadísticamente significativo (o viceversa).

Junto con la introducción de JDORSP, evaluamos el conocimiento temporal en el dominio que proporcionan estos nuevos patrones mediante dos parámetros: sorpresividad y relevancia para el dominio. Además, evaluaremos su relevancia para construir un modelo de clasificación.

Los antecedentes de este tema concreto pueden verse en el Capítulo 3, concretamente en la Sección 3.3.2 sobre minería de patrones secuenciales, o en la Sección 3.5.3 sobre la

minería de patrones discriminatorios. Respecto a las medidas de interés para seleccionar patrones, se puede ampliar información en la Sección 3.5.2. Igualmente recomendamos al lector consultar la razón de probabilidades diagnóstica (Diagnostic Odds Ratio, DOR) como medida de interés en el dominio clínico, en la Sección 3.5.2.4 y en la Sección 7.2.

8.2. Patrones secuenciales de salto DOR (Jumping DOR Sequential Patterns, JDORSP)

Con la nueva forma propuesta de utilizar DOR para reducir drásticamente el número de patrones secuenciales, elegimos sólo el i-ésimo patrón p_i con longitud (l) ítems, en un punto específico en el tiempo (t), que tiene un factor de protección estadísticamente significativo, y las n extensiones ($p_{i1}, p_{i2}, ..p_{in}$) de p_i, de forma que cada una de ellas, con longitud (l+1) ítems, podría ser una S-extensión (en la próxima vez, es decir en t+1) o una I-extensión (al mismo tiempo, t), y tener un factor de riesgo estadísticamente significativo, y viceversa.

Así, por ejemplo, la Figura 8.1 muestra primeramente el patrón número 14 ($BAL_4 < BAL_4 < DIUR_2$) (extraído de la Tabla 8.8 mostrada en la Sección 8.8), con una longitud de 3 ítems, en tres días diferentes A, B y C, con un DOR de 1.68 en el intervalo (1.05, 2.69), teniendo un factor de riesgo estadísticamente significativo, y a continuación sus 2 extensiones, con un factor de protección estadísticamente significativo y una longitud 4 ítems, en primer lugar la S-extensión número 14A ($BAL_4 < BAL_4 < DIUR_2 < PH_4$), en otro día D, donde Día A < Día B < Día C < Día D, con un DOR de 0.18 en el intervalo (0.05, 0.67) y en segundo lugar la I-extensión número 14B ($BAL_4 < BAL_4 < DIUR_2 = PH_4$), en el mismo día C, con un DOR de 0.23 en el intervalo (0.08, 0.69). Tenga en cuenta que los días de la secuencia no son necesariamente consecutivos.

Pattern	Day A	Day B	Day C	Day D
14	BAL_4	BAL_4	$DIUR_2$	
14A	BAL_4	BAL_4	$DIUR_2$	PH_4
14B	BAL_4	BAL_4	$DIUR_2$ PH_4	

Figura 8.1: Ejemplo de extensiones de los patrones.

De esta manera obtenemos patrones secuenciales con la siguiente interpretación o significado: en un punto del tiempo podemos decir que el paciente sobrevivirá (con un factor

estadísticamente significativo de protección DOR), y de repente, algo sucede (generalmente al día siguiente), y una extensión del patrón tiene un factor estadísticamente significativo de riesgo DOR (o viceversa). Los llamamos Jumping Diagnostic Odds Ratio Sequential Patterns (JDORSP).

8.3. Medida de priorización de los patrones: grado de sorpresividad

Con el fin de hacer una ordenación de los patrones, priorizándolos de acuerdo con su interés, definimos una medida que se basa en la diferencia del DOR entre un patrón y sus extensiones. Cuanto mayor sea la diferencia, más sorprendente será el patrón.

Cuando un patrón tiene varias extensiones, se debe realizar una medida agregada. En nuestro caso, queremos priorizar los patrones más sorprendentes, por lo que definimos el grado de sorpresividad, SUR, como el máximo de la diferencia en el valor absoluto de un patrón con cualquiera de sus extensiones. Sea p un patrón, y sea P_x el conjunto con todas las extensiones del patrón p, definimos formalmente SUR en la Ecuación 8.1

$$SUR = \max_{\{x \in P_x\}}(|dor(p) - dor(x)|) \tag{8.1}$$

8.4. Proceso de descubrimiento del conocimiento en tres pasos

En el Capítulo 4 con el fin de construir modelos para predecir la mortalidad en la Unidad de Quemados Críticos, definimos un proceso de descubrimiento del conocimiento de 4 pasos. En el primer paso se realiza el preprocesamiento previo de la base de datos, para posteriormente utilizar una técnica de descubrimiento de patrones secuenciales que muestran la evolución de los pacientes. A continuación, proponemos un posprocesamiento de los patrones con el fin de reducir el número de patrones descubiertos. Por último, con el fin de obtener modelos interpretables, los patrones restantes se utilizan para construir modelos de clasificación en forma de reglas o árboles de decisión.

En este experimento hemos empleado los tres primeros pasos, porque sólo queremos obtener un número reducido de patrones secuenciales con un comportamiento médico específico, y no planteamos como objetivo que los patrones sean buenas variables predictoras en la clasificación. Sin embargo, intentaremos crear un clasificador y mostraremos los resultados de la clasificación utilizando estos patrones.

Punto de corte	INC	DIUR	BAL	BIC	pH	BE
Primero	2.3	0.5	-2.0	17.0	7.20	-4.0
Segundo	3.66	1.0	10.5	21.0	7.30	-2.0
Tercero	5.78	1.9	20.4	25.0	7.35	2.0
Cuarto			52.22	29.0	7.45	4.0
Quinto					7.50	
Sexto					7.60	

Tabla 8.1: Puntos de corte de cada atributo usando la discretización experta.

Punto de corte	INC	DIUR	BAL	BIC	pH	BE
Primero	0.2303	1.4941	0.5701	16.5	7.2767	-7.8833
Segundo	0.2913	1.7749	4.7804	17.6667	7.29	-6.63333
Tercero	0.3203	2.0857	7.4763	19.6167	7.3667	-4.3167
Cuarto	0.5588	2.1697	13.1833	23.0333	7.3983	-0.2333

Tabla 8.2: Puntos de corte de cada atributo usando la discretización UCPD.

En el Paso 0 “Discretización de atributos temporales” hemos utilizado para cada atributo la discretización generada por un experto basada en valores de referencia clínica y además la Discretización de Preservación de la Correlación No Supervisada (Unsupervised Correlation Preserving Discretization, UCPD [87]). Estas medidas se seleccionan de acuerdo con el Capítulo 5, donde se comparan diferentes métodos de discretización.

Por un lado, la discretización de rangos de referencia realizada por un experto (consulte la Tabla 8.1) se ha determinado a partir de una variedad de fuentes y, por otro, los puntos de corte automáticos calculados por el método de discretización UCPD se muestran en la Tabla 8.2. La discretización experta es preferida por los médicos, ya que se basa principalmente en valores de rangos de referencia, siendo necesario hacer una interpretación manual de los patrones. Para una mejor comprensión de los patrones, los intervalos de discretización experta se muestran en la Tabla 1.2 de la Sección 1.2.4.1, donde además se explica la discretización experta aplicada a cada variable.

Por ejemplo, si $_i$ marca el intervalo de discretización i donde $i = 0$ es el intervalo más bajo, cuando encontramos el patrón $INC_0 < DIUR_1 < PH_2$ usando la discretización experta, significa que la secuencia temporal comienza con entrantes de fluidos administrados menores de 2.3 (INC_0), le sigue la diuresis entre 0.5 y 1 ($DIUR_1$), y más tarde el pH se encuentra entre 7.3 y 7.35 (PH_2).

En el Paso 1 “Minería de patrones secuenciales multivariantes” usamos el algoritmo FaSPIP [50]. Hemos considerado diferentes soportes de reglas desde el 16 % al 6 % para generar los patrones, tal y como se hizo en el Capítulo 5, donde comparamos diferentes algoritmos

de discretización, tratando de encontrar el soporte más alto que genera menos patrones con los mejores resultados de clasificación. De esta manera podremos comparar el número de patrones obtenidos y observar la reducción realizada en el número de patrones.

Y finalmente en el Paso 2 "Posprocesamiento", vamos a seleccionar los patrones basados en la definición del factor de riesgo o del factor de protección del DOR.

Cuando calculamos el DOR de cada patrón, también podemos calcular el intervalo de confianza para inferir si la asociación es estadísticamente significativa o no. Los intervalos de confianza son una ayuda importante para interpretar la importancia clínica y la significancia estadística de la razón de probabilidades diagnóstica (véase la Sección 7.2).

En este experimento hemos elegido un intervalo de confianza del 95 %, que muestra si un DOR es estadísticamente significativo [58], de forma que:

- Cuando todo el intervalo de confianza del 95 % es menor que 1, el DOR es estadísticamente significativo, y la exposición se considera protectora en la población del estudio.
- Cuando todo el intervalo de confianza del 95 % es mayor que 1, el DOR es estadísticamente significativo, y la exposición se considera de riesgo en la población del estudio.
- Cuando el intervalo de confianza del 95 % se superpone a DOR = 1, se dice que el DOR no es estadísticamente significativo en la población del estudio. Esto puede reflejar una verdadera ausencia de una relación entre la exposición y la enfermedad.

Para calcular el DOR, los problemas potenciales asociados con sensibilidades y especificidades del 100 % se resuelven añadiendo 0.5 a todas las celdas de la tabla de contingencia diagnóstica.

8.5. Experimentos

Para realizar los experimentos vamos a emplear los datos obtenidos de los 465 pacientes de la Unidad de Quemados Críticos descritos en la sección 1.2.4. Elegiremos solamente los patrones secuenciales que tienen un DOR con un factor estadístico de protección y todos los patrones consecutivos con un factor estadístico de riesgo, y viceversa (a los que llamaremos Jumping DOR Sequential Patterns).

Para comparar el número de patrones generados, proponemos dos experimentos de referencia para seleccionar patrones discriminatorios que tampoco necesitan umbrales especificados por el usuario: Usando patrones JEP (Jumping Emerging Patterns) y usando la no superposición de los intervalos de confianza del DOR [73].

A continuación se explican detalladamente cada uno de estos experimentos.

8.5.1. Experimento de referencia 1: utilizando patrones JEP

En nuestro primer experimento de referencia, ya que queremos extraer reglas exclusivas de los supervivientes como de los no supervivientes, extraemos patrones JEP del subconjunto de supervivientes y del subconjunto de no supervivientes eliminando el comportamiento común de la evolución del paciente que no es discriminatorio. Se puede encontrar más información de los patrones JEP utilizados en la Sección 3.5.3.1. Además, este experimento de referencia esta descrito más detalladamente en el Capítulo 4.

8.5.2. Experimento de referencia 2: utilizando la no superposición del intervalo de confianza del DOR

En el segundo experimento de referencia, seleccionamos aquellas extensiones de los patrones que tengan un cambio estadísticamente significativo en el DOR, tal y como se indica en [73]. El DOR entre dos patrones es significativamente diferente si sus intervalos de confianza del 95 % no se superponen. Además, solo se han incluido reglas en las que al calcular su intervalo, este no cruza el valor 1 (como se ha realizado en [116]). De esta manera todas las reglas serán estadísticamente significativas.

En el Capítulo 7 se amplía más información sobre cómo se puede utilizar el DOR en diferentes formas para seleccionar patrones secuenciales multivariantes.

8.5.3. Experimento: utilizando patrones JDORSP

Por último, hemos seleccionado los Jumping DOR Sequential Patterns que han sido propuestos previamente. Con el fin de reducir aún más el número de patrones, también hemos realizado una búsqueda de los mejores primero, para elegir sólo la extensión de patrón más prometedora con el DOR más alto para cada patrón.

Como comprobaremos posteriormente, los patrones obtenidos tienen la más alta calidad de todos los experimentos que se han realizado anteriormente, ya que se genera una cantidad muy reducida de patrones, por lo que pueden ser revisados manualmente por un experto con el fin de evaluar su posible relevancia clínica.

Discre-tización	Soporte patrones	Patrones iniciales Viven+Mueren	Referencia 1 JEP	Referencia 2 No superp. DOR		Experimento JDORSP	
				todos	mejor	todos	mejor
Experto	10 %	46,041+83,015	391	858	746	83	76
	8 %	88,084+241,866	4,931	2,195	1,856	163	146
	6 %	224,952+492,504	47,113	4,545	3,803	303	273
UCPD	16 %	238,337+49,947	2,179	1,529	1,415	269	212
	14 %	396,238+68,654	7,556	2,296	2,052	357	280
	12 %	647,943+137,546	22,940	6,418	5,228	680	552

Tabla 8.3: Número de patrones interesantes seleccionados después de la minería en el subconjunto de supervivientes y en el conjunto de no supervivientes, para la discretización de Experto y UCPD

8.5.4. Resultados de los experimentos

La Tabla 8.3 muestra el número de patrones discriminatorios que se han seleccionado después de procesar los dos experimentos de referencia y nuestra nueva propuesta utilizando diferentes algoritmos de discretización (Experto y UCPD) y variando el soporte de las reglas dependiendo de la discretización.

Como explicamos anteriormente, tanto con nuestra nueva propuesta, como con la no superposición del DOR, hemos realizado adicionalmente dos tipos de experimentos, ya que cuando calculamos la diferencia del DOR entre un patrón y sus extensiones para elegir aquellos patrones discriminatorios, podemos seleccionar "todas" las extensiones de los patrones o elegir sólo la "mejor" extensión del patrón utilizando una búsqueda del mejor primero para seleccionar el valor más alto del DOR. De esta manera somos capaces de reducir ligeramente el número de patrones. En nuestro caso pocos patrones tienen dos o más extensiones, pero este criterio es válido en general.

Como podemos ver en la Tabla 8.3, hay en general una gran reducción en el número de patrones. Así, para la discretización experta y el 8 % de soporte, de los 329,950 patrones iniciales (88,084 patrones para supervivientes + 241,866 patrones para no supervivientes), obtenemos 4,931 patrones utilizando JEP (Jumping Emerging Patterns) (es decir, -98.5 % respecto de los patrones originales). Con la no superposición del DOR (todos) obtenemos 2,195 patrones (-55.5 % respecto del uso de JEP) y utilizando nuestra nueva propuesta Jumping DOR Sequential Patterns (todos) obtenemos exclusivamente 163 patrones (-92.57 % respecto del uso de la no superposición de DOR).

Además, podemos observar, que cuando usamos la discretización UCPD, la reducción en el número de patrones es similar.

En la Tabla 8.4, vemos el número de patrones seleccionados inicialmente y el número de

sus extensiones cuando utilizamos el DOR para minar los patrones (en el Experimento 2 de referencia y con el nuevo Experimento). Por lo tanto, si continuamos con la discretización experta y un soporte del 8 % (todos), podemos ver que usando la no superposición del DOR, obtenemos 2,195 patrones, donde 928 son patrones iniciales y 1,267 son sus extensiones. Si estudiamos detenidamente estas extensiones, podemos observar que cuando un patrón inicial tiene un DOR de protección, las extensiones pueden transformarse a riesgo (41 patrones) o continuar con un factor de protección (21 patrones). Además, el patrón inicial puede tener un factor de riesgo, y sus extensiones tener un factor de riesgo (1,156 patrones) o un factor de protección (49 patrones).

Como podemos apreciar, con la discretización experta y el uso de la no superposición del DOR, es habitual comenzar con patrones iniciales que son factor de riesgo para continuar con extensiones que también constituyen factor de riesgo. Consideramos que estos patrones son menos interesantes para originar nuevo conocimiento clínico, y que aquellos patrones en los que hay un cambio significativo son los interesantes con el fin de obtener patrones sorprendentes.

Por lo tanto, con nuestra nueva propuesta, Jumping DOR Sequential Patterns (JDORSP), seleccionamos sólo aquellos patrones en los que se produce un cambio sorpresivo del DOR, es decir, patrones que inicialmente tienen un factor de protección y sus extensiones tienen un factor de riesgo (41 patrones si continuamos con el ejemplo anterior) o los patrones que inicialmente tienen un factor de riesgo y sus extensiones tienen un factor de protección (49 patrones).

De esta manera, podemos ver en la misma Tabla 8.4 que, con JDORSP, de los 163 patrones seleccionados, hay 73 patrones iniciales y 90 extensiones (41 con factor de riesgo y 49 con factor de protección).

8.6. Discusión

Una vez que hemos obtenido un pequeño número de patrones secuenciales que representan un cambio abrupto en la evolución del paciente, lo interesante es que un experto evalúe manualmente cada uno de los patrones obtenidos, y trate de explicar su comportamiento. Podemos observar en la Tabla 8.5 algunos de los 38 patrones descubiertos más interesantes (y sus 45 extensiones) utilizando un soporte del 10 % con discretización experta. La tabla completa está en la Sección 8.8 donde en las últimas columnas evaluamos el nivel de interés (sorpresividad y relevancia) de cada patrón, usando una escala de importancia de 1 a 5 (nada, poca, moderada, importante, mucha). En dicha sección, en primer lugar mostramos en la Tabla 8.8 los patrones descubiertos que inicialmente están en riesgo, y sus extensiones tienen

Minería usando DOR	Discretización	Soporte regla	Número de patrones			Número de extensiones de patrones inic. Protec. a ext.		inic. Riesgo a ext.	
			Inicial	Extensión	Total	Riesgo	Protec.	Riesgo	Protec.
Referencia 2 No Superposición DOR	Experto	10 %	373	485	858	14	21	419	31
		8 %	**928**	**1,267**	**2,195**	**41**	**21**	**1,156**	**49**
		6 %	1,901	2,644	4,545	86	21	2,456	81
	UCPD	16 %	707	822	1,529	135	483	176	28
		14 %	1,024	1,272	2,296	186	593	462	31
		12 %	2,600	3,818	6,418	363	675	2,739	41
Experimento JDORSP	Experto	10 %	38	45	83	14	0	0	31
		8 %	**73**	**90**	**163**	**41**	**0**	**0**	**49**
		6 %	136	167	303	86	0	0	81
	UCPD	16 %	106	163	269	135	0	0	28
		14 %	140	217	357	186	0	0	31
		12 %	276	404	680	363	0	0	41

Tabla 8.4: Número de patrones interesantes iniciales con el número de sus extensiones, indicando el tipo de salto asociado a cada extensión, para la discretización Experta y UCPD, incluyendo todos los hijos.

un factor de protección y en la Tabla 8.9 los patrones que inicialmente tienen un factor de protección y después están en riesgo.

También hemos calculado la diferencia absoluta en el valor DOR entre el patrón inicial y cada extensión. Creemos que este valor, llamado "SUR" por nosotros (y definido anteriormente en la Sección 8.3), podría ser un indicador de la importancia de la extensión del patrón en términos de sorpresividad y relevancia.

Un ejemplo de patrón interesante que se encuentra en la Tabla 8.5 (extraído de la Tabla 8.9) es el patrón número 34 ($PH_3 < PH_3 < PH_3$). Este patrón tiene un factor de protección estadísticamente significativo, con un DOR de 0.59 en el intervalo (0.37, 0.94). Este patrón se produce en 279 pacientes, de los que 43 mueren (15.41 %).

Este patrón tiene la extensión número 34A: $PH_3 < PH_3 < PH_3 < BAL_4$ con un valor DOR de 4.06 en el intervalo (1.85, 8.92), por lo tanto, tiene un factor de riesgo estadísticamente significativo. Este patrón se encuentra en 24 pacientes, donde mueren 11 de ellos (45.83 %). El *grado de sorpresividad* (SUR) es 3.47, y se calcula como el valor absoluto de 0.59 menos 4.06.

Por lo tanto, podemos observar que si el nivel de pH es normal en tres días consecutivos, los pacientes generalmente sobrevivirán, pero si en un cuarto día el balance de fluidos es muy alto, entonces los pacientes tienen un riesgo mucho mayor de muerte.

Si usamos un soporte más bajo, podemos descubrir patrones donde estos cambios ocurren drásticamente. En la Tabla 8.6 podemos ver los 10 principales patrones descubiertos de Riesgo a Protección y los 10 principales patrones descubiertos de Protección al Riesgo, orde-

Núm.	Patrón y extensiones	SUR	DOR	Intervalo	Paci.	% Muerte	Signif.
3	$BAL_4 < BIC_2 < BIC_2$	1.81	2.16	(1.35, 3.45)	156	26.92 % (42)	RIES.
3A	$BAL_4 < BIC_2 < BIC_2 < PH_4$	1.81	0.35	(0.13, 0.95)	50	8 % (4)	PROT.
13	$INC_3 < BE_2 < BE_2$	1.51	1.76	(1.10, 2.81)	178	24.16 % (43)	RIES.
13A	$INC_3 < BE_2 < BE_2 < PH_4$	1.51	0.25	(0.08, 0.76)	50	6 % (3)	PROT.
18	$BAL_4 < BIC_2 < BE_2$	1.44	1.82	(1.14, 2.90)	185	24.32 % (45)	RIES.
18A	$BAL_4 < BIC_2 < BE_2 < PH_4$	1.44	0.38	(0.15, 0.96)	57	8.77 % (5)	PROT.
22	$BAL_4 < BE_2$	1.37	1.88	(1.12, 3.17)	296	21.96 % (65)	RIES.
22A	$BAL_4 < BE_2 < INC_2$	1.37	0.51	(0.27, 0.98)	102	11.76 % (12)	PROT.
22B	$BAL_4 < BE_2 < PH_4$	1.37	0.51	(0.29, 0.90)	139	12.23 % (17)	PROT.

(a) De Riesgo a Protección

Núm.	Patrón y extensiones	SUR	DOR	Intervalo	Paci.	% Muerte	Signif.
34	$PH_3 < PH_3 < PH_3$	3.47	0.59	(0.37, 0.94)	279	15.41 % (43)	PROT.
34A	$PH_3 < PH_3 < PH_3 < BAL_4$	3.47	4.06	(1.85, 8.92)	24	45.83 % (11)	RIES.
34B	$PH_3 < PH_3 < PH_3 = BE_1$	1.86	2.45	(1.03, 5.83)	23	34.78 % (8)	RIES.
35	$BIC_3 < PH_3$	2.92	0.58	(0.36, 0.93)	249	14.86 % (37)	PROT.
35A	$BIC_3 < PH_3 < BAL_4$	2.92	3.50	(1.61, 7.60)	26	42.31 % (11)	RIES.
38	$DIUR_2 < PH_3 < BIC_3$	1.77	0.61	(0.38, 0.98)	211	14.69 % (31)	PROT.
38A	$DIUR_2 < PH_3 < BIC_3 < BAL_3$	1.77	2.38	(1.19, 4.76)	39	33.33 % (13)	RIES.

(b) De Protección a Riesgo

Tabla 8.5: Ejemplo de algunos de los patrones más sorpresivos descubiertos utilizando el proceso de minería JDORSP (todos) con soporte al 10 % y discretización experta (extraídos de las tablas de la Sección 8.8).

nados por SUR, utilizando un soporte del 6 % con discretización experta (mejor), extraídos de los 273 patrones JDORSP originales descubiertos.

Podemos observar que en algunos de estos patrones hay un cambio drástico en sus propiedades de frecuencia. De esta manera hemos descubierto patrones secuenciales donde todos los pacientes finalmente pueden acabar muriendo o viviendo. A estos patrones los llamamos JDORSP Extremos.

Como ejemplo, el patrón 39 ($BIC_1 < BE_2 < BE_2$) donde el bicarbonato es bajo y después del exceso de base es normal 2 días consecutivos tiene un factor de riesgo estadísticamente significativo (DOR = 2.65). Pero, si al día siguiente, representado por el patrón 39A ($BIC_1 < BE_2 < BE_2 < PH_4$), el pH es un poco más alto entonces tenemos un factor de protección estadísticamente significativo (DOR = 0.09), donde además sobreviven absolutamente todos los pacientes (sobreviven 22 pacientes de los 22 pacientes que tienen este patrón).

También debemos examinar patrones donde el cambio de frecuencia es muy alto (y no sólo del 100 %). Si observamos el patrón 51 ($PH_4 < PH_4$) donde el pH es ligeramente alto 2 días, tiene un factor de protección estadísticamente significativo, con un DOR = 0.47, donde

Núm.	Patrón y extensiones	SUR	DOR	Intervalo	Paci.	% Muerte	Signif.
39	$BIC_1 < BE_2 < BE_2$	2.56	2.65	(1.61, 4.36)	100	32 % (32)	RIES.
39A	$BIC_1 < BE_2 < BE_2 < PH_4$	2.56	0.09	(0.01, 0.90)	22	0 % (0)	PROT.
40	$DIUR_3 = BAL_4 < DIUR_3 = BIC_2$	2.50	2.58	(1.42, 4.67)	56	33.93 % (19)	RIES.
40A	$DIUR_3 = BAL_4 < DIUR_3 = BIC_2 < PH_4$	2.50	0.08	(0.01, 0.74)	25	0 % (0)	PROT.
41	$INC_3 = DIUR_3 < DIUR_3 = BIC_2$	2.41	2.50	(1.38, 4.53)	57	33.33 % (19)	RIES.
41A	$INC_3 = DIUR_3 < DIUR_3 = BIC_2 < PH_4$	2.41	0.09	(0.01, 0.90)	22	0 % (0)	PROT.
42	$INC_3 = DIUR_3 = BAL_4 < BIC_2$	2.32	2.48	(1.40, 4.39)	64	32.81 % (21)	RIES.
42A	$INC_3 = DIUR_3 = BAL_4 < BIC_2 < PH_4$	2.32	0.16	(0.03, 0.97)	26	3.85 % (1)	PROT.
43	$DIUR_3 = BAL_4 < BAL_4$	2.18	2.34	(1.39, 3.94)	88	30.68 % (27)	RIES.
43A	$DIUR_3 = BAL_4 < BAL_4 < PH4$	2.18	0.16	(0.03, 0.97)	26	3.85 % (1)	PROT.
44	$INC_3 = DIUR_3 < BIC_2 = PH_3$	2.14	2.30	(1.30, 4.06)	67	31.34 % (21)	RIES.
44A	$INC_3 = DIUR_3 < BIC_2 = PH_3 < PH_4$	2.14	0.16	(0.03, 0.97)	26	3.85 % (1)	PROT.
45	$BIC_1 < BE_2$	2.03	2.26	(1.42, 3.61)	157	27.39 % (43)	RIES.
45A	$BIC_1 < BE_2 < INC_2$	2.03	0.23	(0.06, 0.87)	37	5.41 % (2)	PROT.
46	$INC_3 = DIUR_3 < BIC_2$	1.98	2.09	(1.24, 3.51)	94	28.72 % (27)	RIES.
46A	$INC_3 = DIUR_3 < BIC_2 < PH_4$	1.98	0.11	(0.02, 0.58)	37	2.7 % (1)	PROT.
47	$BE_0 < DIUR_2 < BAL_0$	1.97	2.06	(1.20, 3.53)	83	28.92 % (24)	RIES.
47A	$BE_0 < DIUR_2 < BAL_0 = BE_3$	1.97	0.09	(0.01, 0.90)	22	0 % (0)	PROT.
48	$BAL_4 < BE_2 < BIC_2 = BE_2$	1.93	2.13	(1.32, 3.42)	139	27.34 % (38)	RIES.
48A	$BAL_4 < BE_2 < BIC_2 = BE_2 < PH_4$	1.93	0.20	(0.05, 0.73)	40	4.76 % (2)	PROT.

(a) De Riesgo a Protección

Núm.	Patrón y extensiones	SUR	DOR	Intervalo	Paci.	% Muerte	Signif.
49	$BIC_2 < DIUR_2 = BAL_0$	45.55	0.49	(0.27, 0.88)	128	11.72 % (15)	PROT.
49A	$BIC_2 < DIUR_2 = BAL_0 < PH_1$	45.55	46.04	(8.24, 257.18)	5	100 % (5)	RIES.
50	$DIUR_2 < DIUR_2 = BAL_0$	45.47	0.57	(0.34, 0.97)	156	13.46 % (21)	PROT.
50A	$DIUR_2 < DIUR_2 = BAL_0 < PH_1$	45.47	46.04	(8.24, 257.18)	5	100 % (5)	RIES.
51	$PH_4 < PH_4$	27.46	0.47	(0.26, 0.87)	123	11.38 % (14)	PROT.
51A	$PH_4 < PH_4 < BE_1$	27.46	27.93	(6.71, 116.26)	7	85.71 % (6)	RIES.
52	$DIUR_2 = BAL_0$	27.41	0.52	(0.32, 0.87)	183	13.11 % (24)	PROT.
52A	$DIUR_2 = BAL_0 < PH_1$	27.41	27.93	(6.71, 116.26)	7	85.71 % (6)	RIES.
53	$DIUR_2 < DIUR_2 < BAL_0$	13.36	0.57	(0.34, 0.94)	176	13.64 % (24)	PROT.
53A	$DIUR_2 < DIUR_2 < BAL_0 = PH_1$	13.36	13.93	(3.97, 48.84)	8	75 % (6)	RIES.
54	$INC_3 = DIUR_2 < BAL_0$	12.13	0.53	(0.30, 0.92)	143	12.59 % (18)	PROT.
54A	$INC_3 = DIUR_2 < BAL_0 = BE_0$	12.13	12.66	(4.34, 36.95)	11	72.73 % (8)	RIES.
55	$DIUR_2 < INC_2 < BAL_0$	11.03	0.43	(0.22, 0.82)	107	10.28 % (11)	PROT.
55A	$DIUR_2 < INC_2 < BAL_0 = BE_0$	11.03	11.43	(3.04, 43.26)	7	71.43 % (5)	RIES.
56	$BAL_0 < BIC_3$	10.99	0.47	(0.27, 0.82)	153	11.76 % (18)	PROT.
56A	$BAL_0 < BIC_3 = PH2$	10.99	11.46	(3.04, 43.26)	7	71.43 % (5)	RIES.
57	$DIUR_2 < BE_2 < PH_3 < BIC_3$	10.99	0.47	(0.24, 0.91)	100	11 % (11)	PROT.
57A	$DIUR_2 < BE_2 < PH_3 < BIC_3 < BAL_3$	10.99	11.46	(3.04, 43.26)	7	71.43 % (5)	RIES.
58	$BIC_3 < DIUR_2$	10.97	0.49	(0.31, 0.79)	237	13.5 % (32)	PROT.
58A	$BIC_3 < DIUR_2 = BE_1$	10.97	11.46	(3.04, 43.26)	7	71.43 % (5)	RIES.

(b) De Protección a Riesgo

Tabla 8.6: Listado de los 10 patrones más sorpresivos descubiertos usando el proceso de minería JDORSP (mejor) con soporte al 6 % y discretización experta.

sólo mueren 14 pacientes de 123 (11.38 %). Pero, si al día siguiente, representado por el patrón 51A ($PH_4 < PH_4 < BE_1$), el exceso de base es bajo, entonces tenemos un factor de riesgo estadísticamente significativo, donde 6 de 7 pacientes morirán (85.71 %).

Adicionalmente hemos construido un modelo de clasificación utilizando el reducido nú-

mero de patrones JDORSP descubiertos, con la restricción adicional de que tiene que ser interpretable. El objetivo es obtener un modelo con un pequeño número de patrones y que sea fácil de interpretar por el médico. Para ello utilizamos un clasificador basado en reglas o en árboles de decisión como clasificadores interpretables representativos (realizamos una evaluación más completa de la clasificación utilizando como medida de interés el DOR en el Capítulo 7).

Por un lado, usamos JRIP (la implementación de RIPPER en WEKA) para el aprendizaje de reglas, y por otro lado, elegimos el árbol de decisiones J48 implementado por WEKA para el algoritmo C4.5. En la Sección 4.3.4 se explican con más detalle los motivos por los que se han elegido estos algoritmos de clasificación.

En todos los casos, configuramos los clasificadores con el mismo número mínimo de elementos en cada hoja o regla al 2 % de las instancias. La precisión, sensibilidad, especificidad y AUC se calcularon mediante una validación cruzada de 10 iteraciones.

En la Tabla 8.7 proporcionamos los resultados de clasificación de los experimentos de referencia (utilizando patrones JEP y la no superposición del intervalo de confianza del DOR) y los resultados utilizando los nuevos patrones JDORSP, con un soporte del 8 % y discretización experta. Los resultados de clasificación de los experimentos de referencia se explican con más detalle en el Capítulo 7, pero en general con JEP e incluso con la no superposición del intervalo de confianza del DOR, obtenemos una precisión superior al 90 % con una especificidad cercana al 60 %. Sin embargo, con los patrones JDORSP, la precisión es sólo del 80 % y la especificidad es demasiado baja, alrededor del 33 %.

También debemos recordar que para la construcción del modelo de clasificación, inicialmente teníamos 4,931 patrones usando JEP, 2,195 patrones utilizando la no superposición del DOR, y sólo 163 patrones con JDORSP. Aunque estos 163 patrones cubren a todos los pacientes, debemos observar que de sólo estos 163 patrones el clasificador tiene que elegir los mejores para clasificar a los 87 pacientes que finalmente van a morir (de un total de 465 pacientes).

También hemos estudiado el número y la longitud de los patrones utilizados en el clasificador. Un patrón corto es más simple y más general (cubre a más pacientes), sin embargo, un patrón largo es más específico (cubre menos pacientes) y es más difícil de entender. Cuando intentamos construir un clasificador con JDORSP, estos patrones son los más cortos, y por lo tanto, demasiado generales. Además, el clasificador no es capaz de elegir patrones más específicos porque el número de patrones a utilizar es demasiado bajo, lo que provoca que obtengamos resultados de clasificación inferiores a los obtenidos en otros capítulos.

Finalmente, para evaluar el nivel de interés (sorpresividad y relevancia) de los nuevos patrones secuenciales descubiertos (JDORSP), los dos médicos optaron por estudiar la relación

Clasificador	Patrones iniciales	Método	Núm. patr.	Total long.	Media long.	Sensib.	Especif.	Precisión	AUC
J48	4,931	JEP	17	84	4.94	**100.00 %**	**56.32 %**	**91.83 %**	**0.782**
	2,195	No superp.	16	77	4.81	**94.97 %**	**58.62 %**	**88.17 %**	**0.741**
	163	JDORSP	23	71	3.08	91.53 %	33.33 %	80.65 %	0.605
JRIP	4,931	JEP	15	79	5.27	**100.00 %**	**58.62 %**	**92.26 %**	**0.777**
	2,195	No superp.	13	60	4.62	91.80 %	33.33 %	80.86 %	0.641
	163	JDORSP	13	37	2.84	90.21 %	33.33 %	79.57 %	0.617

Tabla 8.7: Resultados del experimento de clasificación, con soporte de patrón del 8 % y discretización experta, utilizando JEP, la no superposición del intervalo de confianza del DOR (todos) y JDORSP (todos) en el paso de posprocesamiento.

entre las variables relacionadas con la reanimación (aporte de líquidos y balance de fluidos), las variables relacionadas con la perfusión tisular (pH de la sangre arterial, concentración de bicarbonato y exceso de base) y la mortalidad en la Unidad de Quemados Críticos. Esta decisión se tomó por varias razones. En primer lugar, se trata de variables susceptibles de alteración y, si se demuestra una asociación con el desenlace de interés, se podría hipotetizar una relación causal y, de demostrarse, dichas variables podrían utilizarse como dianas terapéuticas. En segundo lugar, porque esas variables están relacionadas con los esfuerzos de reanimación destinados a restaurar la perfusión de órganos después del trauma. Los líquidos infundidos (para restaurar la perfusión de los órganos y la producción de orina), la producción de orina (el objetivo más inmediato de la reanimación) y el balance de fluidos (la diferencia entre los líquidos administrados y los líquidos perdidos por la orina y otras pérdidas corporales), son variables que resumen los cambios asociados con la principal intervención terapéutica inmediatamente después del trauma, es decir, la reposición de líquidos. Podría decirse que los patrones sobre la evolución del paciente pueden tener al menos dos usos diferentes. En primer lugar, pueden utilizarse para establecer objetivos terapéuticos o resultados a conseguir en el tratamiento de los pacientes y, en segundo lugar, pueden utilizarse como medio de seguimiento y anticipación de la aparición de riesgos en el paciente.

Los médicos evaluaron si los patrones secuenciales agregan nuevo conocimiento (sorpresividad) y si son clínicamente relevantes porque pueden implicar algo interesante que revisar (relevancia). Los patrones serán buenos si son relevantes. Si son sorprendentes, se podría identificar una posible línea de interés para la investigación, mientras que si no son novedosos, sino confirmatorios, se podría concluir que el método podría utilizarse para otros campos no explorados. Se utilizó una escala del 1 al 5, donde 1 es muy bajo y 5 es muy alto. La relevancia aumenta cuando se convierte un factor de riesgo en un factor de protección, como cuando se corrige una alteración, o cuando, tras varias determinaciones anormales, una sola determinación corregida cambia el pronóstico.

Al analizar las tablas en la Sección 8.8, se notará que los patrones encontrados son altamente relevantes, con una relevancia promedio de 4.8. La novedad de las extensiones es mayor que la de los patrones padres, tanto globalmente como en los dos tipos de patrones. En cuanto a la novedad que aportan, las extensiones de los patrones que pasan de factor de riesgo a factor de protección es muy alta (4.9) respecto al factor global, que es 3.55, o las extensiones de los patrones que pasan a ser de riesgo (3.36). En este caso, se observará que lo más interesante es que después de varios días de estar en riesgo, se produce un cambio y los patrones empiezan a tener un factor de protección. Este cambio no sería de extrañar en otros patrones de menor duración.

Los patrones secuenciales (de riesgo o protección) identificados aquí son de gran interés clínico, ya que algunos son muy sorprendentes o relevantes (puntuaciones cercanas a 5). Por ejemplo, el patrón 1, que muestra que un balance de fluidos muy positivo se asocia con un mal pronóstico, es relevante, ya que indica que los médicos deben tener en cuenta este cambio al realizar el pronóstico de los pacientes con quemaduras (y tal vez para afinar la administración de líquidos durante la reanimación). Sin embargo, no es sorprendente, ya que el paradigma actualmente aceptado propone que la administración excesiva de líquidos podría conducir a la formación de un edema excesivo y, por lo tanto, estar asociado con un mal pronóstico. Sin embargo, el patrón 1A, que muestra que un balance de fluidos fuertemente positivo seguido de un exceso de base y bicarbonato dentro del rango normal, seguido a su vez por un pH más bien en el rango alcalótico, es protector, es bastante sorprendente. Esto se debe a que documenta que los efectos nocivos de un balance de fluidos positivo parecen compensarse si el pH se normaliza posteriormente (o incluso se desplaza hacia el rango alcalótico). Este patrón también es relevante, ya que informa de la fisiopatología aun incompletamente conocida de los cambios posteriores al trauma y su impacto en el pronóstico.

8.7. Conclusiones

En este capítulo proponemos un método novedoso para obtener un subconjunto reducido de patrones temporales sorprendentes y novedosos para representar la evolución temporal del estado clínico del paciente, llamados “Jumping Diagnostic Odds Ratio Sequential Patterns (JDORSP)”. Utilizamos la Razón de Probabilidades Diagnóstica (Diagnostic Odds Ratio, DOR) para seleccionar patrones secuenciales que representen un cambio drástico en la evolución, es decir, patrones que se convierten en un factor de protección cuando extendemos un patrón que era un factor de riesgo, o viceversa. Hasta donde sabemos es la primera vez que el DOR y los patrones secuenciales se utilizan de esta manera.

Hemos evaluado la idoneidad de nuestro método en pacientes de una Unidad de Quema-

dos Críticos, y hemos obtenido las siguientes conclusiones:

- Destacamos la drástica reducción de patrones con respecto al estado actual de la técnica (patrones JEP o la no superposición del intervalo de confianza del DOR). Esta notable reducción es particularmente útil para poder realizar posteriormente una revisión manual de los patrones encontrados por expertos médicos.
- Los patrones JDORSP proporcionan nuevo conocimiento, siendo patrones locales que representan un subconjunto de toda la población. Esta característica implica que un clasificador que utilice sólo estos patrones como covariables no consigue una alta precisión.
- Hemos evaluado la sorpresividad y relevancia de estos nuevos patrones con los médicos, y el hecho más interesante encontrado es la alta sorpresividad (4.9 sobre 5) de las extensiones de los patrones que se convierten en factor de protección, es decir, los pacientes que se recuperan a los pocos días de estar en alto riesgo de morir.

8.8. Patrones JDORSP descubiertos (con 10 % de soporte y discretización experta)

Núm.	Patrón y extensiones	SUR	DOR	DOR Intervalo	Pacientes	% Muerte	Signif.	Sorpr.	Relev.
1	$BAL_4 < BE_2 < BIC_2 = BE_2$	1.93	2.13	(1.32, 3.42)	139	27.34 % (38)	RIES.	2	4
1A	$BAL_4 < BE_2 < BIC_2 = BE_2 < PH_4$	1.93	0.20	(0.05, 0.73)	42	4.76 % (2)	PROT.	5	5
2	$INC_3 < BIC_2 = BE_2 < BE_2$	1.88	2.09	(1.29, 3.36)	136	27.21 % (37)	RIES.	4	4
2A	$INC_3 < BIC_2 = BE_2 < BE_2 < PH_4$	1.88	0.21	(0.06, 0.78)	40	5 % (2)	PROT.	5	5
3	$BAL_4 < BIC_2 < BIC_2$	1.81	2.16	(1.35, 3.45)	156	26.92 % (42)	RIES.	2	4
3A	$BAL_4 < BIC_2 < BIC_2 < PH_4$	1.81	0.35	(0.13, 0.95)	50	8 % (4)	PROT.	5	5
4	$BAL_4 < BE_2 < BIC_2$	1.77	1.94	(1.21, 3.12)	146	26.03 % (38)	RIES.	2	4
4A	$BAL_4 < BE_2 < BIC_2 < PH_4$	1.77	0.17	(0.05, 0.63)	47	4.26 % (2)	PROT.	5	5
5	$DIUR_3 = BAL_4 < BIC_2$	1.74	1.84	(1.09, 3.11)	97	26.8 % (26)	RIES.	2	4
5A	$DIUR_3 = BAL_4 < BIC_2 < PH_4$	1.74	0.10	(0.02, 0.48)	42	2.38 % (1)	PROT.	5	5
6	$BAL_4 < BIC_2$	1.72	2.18	(1.29, 3.69)	289	22.84 % (66)	RIES.	3	4
6A	$BAL_4 < BIC_2 < PH_4$	1.72	0.46	(0.26, 0.84)	132	11.36 % (15)	PROT.	5	5
7	$DIUR_3 < BIC_2 = PH_3$	1.71	1.80	(1.10, 2.95)	120	25.83 % (31)	RIES.	5	4
7A	$DIUR_3 < BIC_2 = PH_3 < PH_4$	1.71	0.09	(0.02, 0.44)	45	2.22 % (1)	PROT.	5	5
8	$DIUR_3 = BAL_4 < DIUR_3$	1.68	1.78	(1.06, 3.00)	99	26.26 % (26)	RIES.	2	4
8A	$DIUR_3 = BAL_4 < DIUR_3 < PH_4$	1.68	0.10	(0.02, 0.54)	39	2.56 % (1)	PROT.	5	5
9	$BAL_4 < BIC_2 = BE_2 < BE_2$	1.65	1.91	(1.19, 3.06)	152	25.66 % (39)	RIES.	2	5
9A	$BAL_4 < BIC_2 = BE_2 < BE_2 < PH_4$	1.65	0.26	(0.08, 0.78)	49	6.12 % (3)	PROT.	5	5
10	$BAL_4 < BIC_2 = BE_2 < BIC_2$	1.63	1.73	(1.05, 2.86)	118	25.42 % (30)	RIES.	2	5
10A	$BAL_4 < BIC_2 = BE_2 < BIC_2 < PH_4$	1.63	0.10	(0.02, 0.50)	41	2.44 % (1)	PROT.	5	5
11	$BAL_4 < BIC_2 < BIC_2 = BE_2$	1.62	1.92	(1.19, 3.11)	133	26.32 % (35)	RIES.	3	5
11A	$BAL_4 < BIC_2 < BIC_2 = BE_2 < PH_4$	1.62	0.30	(0.10, 0.94)	43	6.98 % (3)	PROT.	5	5
12	$BAL_4 < BE_2 < BE_2$	1.61	1.90	(1.19, 3.03)	202	24.26 % (49)	RIES.	2	5
12A	$BAL_4 < BE_2 < BE_2 < PH_4$	1.61	0.29	(0.11, 0.78)	58	6.9 % (4)	PROT.	5	5
13	$INC_3 < BE_2 < BE_2$	1.51	1.76	(1.10, 2.81)	178	24.16 % (43)	RIES.	3	5

Núm.	Patrón y extensiones	SUR	DOR	DOR Intervalo	Pacientes	% Muerte	Signif.	Sorpr.	Relev.
13A	$INC_3 < BE_2 < BE_2 < PH_4$	1.51	0.25	(0.08, 0.76)	50	6 % (3)	PROT.	5	5
14	$BAL_4 < BAL_4 < DIUR_2$	1.50	1.68	(1.05, 2.69)	172	23.84 % (41)	RIES.	2	5
14A	$BAL_4 < BAL_4 < DIUR_2 < PH_4$	1.50	0.18	(0.05, 0.67)	45	4.44 % (2)	PROT.	5	5
14B	$BAL_4 < BAL_4 < DIUR_2 = PH_4$	1.45	0.23	(0.08, 0.69)	54	5.56 % (3)	PROT.	5	5
15	$BIC_2 < BIC_2$	1.46	1.92	(1.17, 3.14)	263	22.81 % (60)	RIES.	3	5
15A	$BIC_2 < BIC_2 < PH_4$	1.46	0.46	(0.24, 0.87)	110	10.91 % (12)	PROT.	5	5
16	$INC_3 < BIC_2 = BE_2$	1.46	1.88	(1.18, 2.99)	203	24.14 % (49)	RIES.	3	5
16A	$INC_3 < BIC_2 = BE_2 < INC_2$	1.46	0.42	(0.20, 0.88)	82	9.76 % (8)	PROT.	5	5
17	$BAL_4 < DIUR_2$	1.45	1.90	(1.05, 3.43)	343	20.99 % (72)	RIES.	2	5
17A	$BAL_4 < DIUR_2 = PH_4$	1.45	0.45	(0.25, 0.82)	134	11.19 % (15)	PROT.	5	5
17B	$BAL_4 < DIUR_2 < PH_4$	1.44	0.46	(0.26, 0.81)	147	11.56 % (17)	PROT.	2	5
18	$BAL_4 < BIC_2 < BE_2$	1.44	1.82	(1.14, 2.90)	185	24.32 % (45)	RIES.	2	5
18A	$BAL_4 < BIC_2 < BE_2 < PH_4$	1.44	0.38	(0.15, 0.96)	57	8.77 % (5)	PROT.	5	5
19	$BAL_4 < BAL_4 = PH_3$	1.44	1.76	(1.09, 2.83)	149	24.83 % (37)	RIES.	2	5
19A	$BAL_4 < BAL_4 = PH_3 < PH_4$	1.44	0.32	(0.12, 0.86)	54	7.41 % (4)	PROT.	5	5
20	$BAL_4 = BIC_2 < BIC_2$	1.42	1.70	(1.06, 2.73)	161	24.22 % (39)	RIES.	2	5
20A	$BAL_4 = BIC_2 < BIC_2 < PH_4$	1.42	0.28	(0.12, 0.69)	72	6.94 % (5)	PROT.	5	5
21	$INC_3 < BAL_4 < DIUR_2$	1.42	1.63	(1.01, 2.64)	150	24 % (36)	RIES.	2	5
21A	$INC_3 < BAL_4 < DIUR_2 < PH_4$	1.42	0.21	(0.06, 0.78)	40	5 % (2)	PROT.	5	5
22	$BAL_4 < BE_2$	1.37	1.88	(1.12, 3.17)	296	21.96 % (65)	RIES.	2	5
22A	$BAL_4 < BE_2 < INC_2$	1.37	0.51	(0.27, 0.98)	102	11.76 % (12)	PROT.	5	5
22B	$BAL_4 < BE_2 < PH_4$	1.37	0.51	(0.29, 0.90)	139	12.23 % (17)	PROT.	5	5
23	$BAL_4 < BAL_3$	1.35	1.71	(1.06, 2.73)	228	22.81 % (52)	RIES.	2	5
23A	$BAL_4 < BAL_3 < PH_4$	1.35	0.36	(0.17, 0.76)	91	8.79 % (8)	PROT.	5	5
24	$INC_1 < BIC_2$	1.35	1.64	(1.004, 2.68)	131	24.43 % (32)	RIES.	4	5
24A	$INC_1 < BIC_2 < BIC_3$	1.35	0.29	(0.09, 0.88)	45	6.67 % (3)	PROT.	5	5
25	$INC_3 < BAL_4 < PH_3$	1.33	1.61	(1.002, 2.57)	166	23.49 % (39)	RIES.	2	5
25A	$INC_3 < BAL_4 < PH_3 < PH_4$	1.33	0.28	(0.09, 0.86)	46	6.52 % (3)	PROT.	5	5

Núm.	Patrón y extensiones	SUR	DOR	DOR Intervalo	Pacientes	% Muerte	Signif.	Sorpr.	Relev.
26	$DIUR_3 < BIC_2$	1.31	1.60	(1.003, 2.57)	171	23.39 % (40)	RIES.	4	5
26A	$DIUR_3 < BIC_2 < PH_4$	1.31	0.29	(0.12, 0.70)	71	7.04 % (5)	PROT.	5	5
27	$BAL_4 < BIC_2 = PH_3$	1.29	1.63	(1.02, 2.60)	216	22.69 % (49)	RIES.	2	5
27A	$BAL_4 < BIC_2 = PH_3 < PH_4$	1.29	0.34	(0.16, 0.71)	95	8.42 % (8)	PROT.	5	5
28	$BAL_4 < BIC_2 = PH_3 = BE_2$	1.17	1.60	(1.003, 2.57)	171	23.39 % (40)	RIES.	2	5
28A	$BAL_4 < BIC_2 = PH_3 = BE_2 < PH_4$	1.17	0.43	(0.20, 0.91)	80	10 % (8)	PROT.	5	5

Tabla 8.8: Listado de los 59 patrones JDORSP descubiertos, desde el riesgo hasta la protección, con un soporte del 10 % y discretización experta (todos).

Núm.	Patrón y extensiones	SUR	DOR	DOR Intervalo	Pacientes	% Muerte	Signif.	Sorpr.	Relev.
29	$INC_3 = DIUR_2 < BAL_0$	12.13	0.53	(0.30, 0.92)	143	12.59 % (18)	PROT.	2	5
29A	$INC_3 = DIUR_2 < BAL_0 = BE_0$	12.13	12.66	(4.34, 36.95)	11	72.73 % (8)	RIES.	3	5
30	$DIUR_2 < DIUR_2 < BAL_0$	12.09	0.57	(0.34, 0.94)	176	13.64 % (24)	PROT.	4	5
30A	$DIUR_2 < DIUR_2 < BAL_0 = BE_0$	12.09	12.66	(4.34, 36.95)	11	72.73 % (8)	RIES.	3	5
31	$DIUR_2 < BAL_0$	10.22	0.58	(0.36, 0.93)	275	15.27 % (42)	PROT.	2	5
31A	$DIUR_2 < BAL_0 = BE_0$	10.22	10.80	(4.42, 26.39)	16	68.75 % (11)	RIES.	3	5
31B	$DIUR_2 < BAL_0 = BIC_1$	5.43	6.01	(2.54, 14.20)	18	55.56 % (10)	RIES.	3	5
31C	$DIUR_2 < BAL_0 = BE_1$	2.51	3.09	(1.27,7.50)	20	40.00 % (8)	RIES.	3	5
32	$DIUR_2 = PH_3$	5.69	0.59	(0.36, 0.96)	346	16.47 % (57)	PROT.	3	5
32A	$DIUR_2 = PH_3 < PH_1$	5.69	6.28	(2.40, 16.44)	14	57.14 % (8)	RIES.	3	5
32B	$DIUR_2 = PH_3 < BE_0$	2.04	2.63	(1.10, 6.30)	22	36.36 % (8)	RIES.	3	5
33	$BIC_3 = PH_3 < PH_3$	3.67	0.58	(0.35, 0.95)	192	14.06 % (27)	PROT.	4	5
33A	$BIC_3 = PH_3 < PH_3 < BAL_4$	3.67	4.25	(1.78, 10.11)	19	47.37 % (9)	RIES.	5	5
34	$PH_3 < PH_3 < PH_3$	3.47	0.59	(0.37, 0.94)	279	15.41 % (279)	PROT.	2	5
34A	$PH_3 < PH_3 < PH_3 < BAL_4$	3.47	4.06	(1.85, 8.92)	24	45.83 % (11)	RIES.	3	5
34B	$PH_3 < PH_3 < PH_3 = BE_1$	1.86	2.45	(1.03, 5.83)	23	34.78 % (8)	RIES.	3	5
35	$BIC_3 < PH_3$	2.92	0.58	(0.36, 0.93)	249	14.86 % (37)	PROT.	2	5
35A	$BIC_3 < PH_3 < BAL_4$	2.92	3.50	(1.61, 7.60)	26	42.31 % (11)	RIES.	5	5
36	$BIC_3 = PH_3$	2.19	0.39	(0.24, 0.62)	285	12.98 % (37)	PROT.	2	5
36A	$BIC_3 = PH_3 < PH_2$	2.19	3.09	(1.27, 7.50)	20	40 % (8)	RIES.	3	5
37	$PH_3 < BIC_3$	1.99	0.46	(0.29, 0.74)	309	14.56 % (45)	PROT.	2	5
37A	$PH_3 < BIC_3 = PH_2$	1.99	2.45	(1.03, 5.83)	23	34.78 % (8)	RIES.	3	5
38	$DIUR_2 < PH_3 < BIC_3$	1.77	0.61	(0.38, 0.98)	211	14.69 % (31)	PROT.	3	5
38A	$DIUR_2 < PH_3 < BIC_3 < BAL_3$	1.77	2.38	(1.19, 4.76)	39	33.33 % (13)	RIES.	4	5

Tabla 8.9: Listado de los 24 patrones JDORSP descubiertos, desde la protección al riesgo, con un soporte del 10 % y discretización experta (todos).

Capítulo 9

Conclusiones y trabajo futuro

9.1. Conclusiones

En esta tesis hemos aplicado el proceso de descubrimiento del conocimiento en el contexto de las Unidades de Quemados Críticos, minando patrones secuenciales multivariantes que se han utilizado bien para construir un modelo para predecir la mortalidad de los pacientes o bien para extraer información sobre la posible evolución de las variables del paciente (y por lo tanto, la respuesta al tratamiento).

En relación a la hipótesis de partida (Sección 2.1), se ha demostrado que extraer información comprimida y de fuerte contraste entre dos conjuntos de datos secuenciales puede ser útil para conocer la evolución clínica de los pacientes, y en la construcción de modelos de clasificación secuenciales.

En el Capítulo 4 se trabajó fundamentalmente para conseguir el primer y el segundo objetivo. Así, hemos podido comprobar respecto a la predicción de la mortalidad de los pacientes quemados que solamente utilizando los atributos estáticos no se consigue obtener buenos resultados en la clasificación. Además, también nos hemos encontrado con el problema del drástico aumento del número de patrones cuando se baja el soporte, siendo imprescindible aplicar alguna medida para mejorar la calidad de los patrones.

En nuestro caso no se ha realizado una integración de la minería de patrones con la clasificación, por lo que para construir el clasificador necesitamos primero extraer un conjunto completo de patrones secuenciales dado un soporte mínimo, y después seleccionar una serie de patrones discriminatorios con los que crear un clasificador. De hecho, la consideración más importante en la clasificación con secuencias no es la de encontrar el conjunto completo de reglas, sino la de descubrir los patrones más discriminatorios.

La solución que inicialmente hemos propuesto para seleccionar aquellos patrones fre-

cuentes discriminatorios a utilizar en la clasificación consiste en recurrir a la propiedad de la frecuencia, escogiendo patrones emergentes o patrones JEP, y así poder obtener un número reducido de patrones, que al mismo tiempo ayuden a realizar una clasificación efectiva. Además, hemos *mejorado aún más la calidad de estos patrones*, extrayendo representaciones comprimidas maximales y cerradas de los conjuntos de ítems con alta utilidad.

Los resultados de las pruebas de clasificación muestran que *nuestro enfoque supera a las puntuaciones de gravedad* de quemaduras utilizadas actualmente por los médicos siguiendo la puntuación de Brier, y hasta donde sabemos, este sería el primer trabajo donde patrones secuenciales multivariantes se utilizan como predictores en la UCI.

En el Capítulo 5 se trabajó fundamentalmente para conseguir el tercer objetivo. Hemos comprobado que en el proceso de minería de datos, el número de patrones generados depende en gran medida de la discretización elegida. Una buena discretización en el dominio médico debe generar un pequeño número de patrones y dar lugar a una buena precisión en la clasificación. Si encontramos pocos patrones, estos son más significativos y más robustos, y por lo tanto es más fácil encontrar un significado médico simple para un patrón específico. El mejor rendimiento conseguido en nuestros experimentos con la clasificación se ha obtenido con la discretización automática UCPD y Fusinter. También obtenemos un resultado aceptable con la discretización experta, superando a muchos algoritmos de discretización automática. Podemos concluir que la discretización automática produce cortes arbitrarios que generalmente no se corresponden con el conocimiento clínico y complica la comprensión y la interpretación. Los resultados mostrados podrían ayudar en el futuro a reducir el número de candidatos para elegir un discretizador, en el caso de que se quisiera realizar una clasificación con patrones secuenciales en el dominio clínico.

Hay que tener también en consideración que aunque se han evaluado una amplia gama de algoritmos de discretización, sin embargo, pocos estudios han examinado la discretización de datos clínicos, y ninguno de ellos, que conozcamos, ha tratado con patrones secuenciales específicamente.

En el Capítulo 6 se trabajó fundamentalmente para seguir mejorando el segundo objetivo, de forma que se ha conseguido reducir aún más el número de patrones secuenciales, para que los patrones predictores usados en la clasificación tengan la mayor relevancia médica posible y se encuentren uniformemente distribuidos por toda la base de datos de pacientes. Con este fin, se ha realizado una evaluación de la consistencia, mediante una validación en varias particiones estratificadas, explotando las propiedades basadas en soporte para descartar aquellos patrones no significativos. Hemos comprobado también que con los patrones JEP y seleccionando los patrones consistentes sin restringir el desbalanceo entre las clases se obtienen los mejores resultados de clasificación. También hemos mostrado que el uso de

patrones JEP minimales mejora en general los resultados de clasificación con respecto a los patrones cerrados y maximales, además de generar también un conjunto más reducido de patrones

En el Capítulo 7 se trabajó fundamentalmente para lograr el cuarto objetivo, puesto que proponemos y evaluamos cuatro formas de emplear una métrica estadística, la Razón de Probabilidades Diagnóstica (Diagnostic Odds Ratio, DOR), como medida de interés para reducir el número de patrones y seleccionar sólo los más discriminatorios, ya que la explosión de patrones es el principal problema de la minería de patrones para usarlos como clasificadores. Después de comparar estas cuatro propuestas con un experimento de referencia basado en las propiedades de la frecuencia utilizando patrones JEP, se puede concluir que respecto a seleccionar un menor número de patrones, la mejor opción es la de usar un DOR diferencial y no superpuesto, ya que que a medida que hemos aumentado las restricciones aplicadas, hemos reducido significativamente el número de patrones, logrando así patrones más generales, simples e interesantes. Respecto a la interpretabilidad, podemos concluir que la discretización tiene un gran impacto en el rendimiento de la clasificación a expensas de la interpretabilidad. Con respecto a la precisión, los mejores resultados de clasificación se producen utilizando patrones JEP junto con la discretización UCPD, pero si consideramos únicamente la discretización experta, la mayor especificidad se alcanza utilizando el valor del DOR para seleccionar los patrones. Esta es, hasta donde sabemos, la primera vez que algunos de estos enfoques han sido propuestos y comparados en la literatura científica.

Además, una ventaja que se ha encontrado es que se podría realizar una poda temprana integrada en el algoritmo de descubrimiento de patrones, usando como medida de interés alguna operación relacionada con el DOR entre un patrón secuencial y sus extensiones. Dentro de esta poda temprana se podría incluir la selección de aquellos patrones con las mejores extensiones con el DOR. La realización de nuevos algoritmos de descubrimiento de patrones secuenciales queda fuera del objetivo de esta tesis.

Y por último en el Capítulo 8 se trabajó fundamentalmente para lograr el cuarto y el quinto objetivo, aunque realmente se podría considerar este último objetivo como general para todos los capítulos. De manera que proponemos un nuevo método para obtener un subconjunto reducido de patrones temporales sorprendentes y novedosos para representar la evolución temporal del estado clínico de los pacientes, a los que llamamos “Jumping Diagnostic Odds Ratio Sequential Patterns (JDORSP)”. Utilizamos la Razón de Probabilidades Diagnóstica (DOR) para seleccionar patrones secuenciales que representen un cambio drástico en la evolución del paciente, es decir, patrones que se convierten en un factor de protección cuando extendemos un patrón que era un factor de riesgo, o viceversa.

Hasta donde sabemos es la primera vez en la que el DOR y los patrones secuenciales

se utilizan de esta manera. Destacamos la drástica reducción de patrones con respecto al estado actual de la técnica (patrones JEP o la no superposición del intervalo de confianza del DOR). Esta notable reducción es particularmente útil para poder realizar posteriormente por expertos médicos una revisión manual de la sorpresividad y relevancia de los patrones descubiertos. Así el hecho más interesante encontrado es la alta sorpresividad (4.9 sobre 5) en los patrones secuenciales que inicialmente tienen un factor de riesgo, y sus extensiones se convierten en un factor de protección, es decir, pacientes que se recuperan a los pocos días de estar en alto riesgo de morir.

9.2. Trabajo futuro

El uso de una representación de los patrones secuenciales mediante intervalos en vez de puntos puede reducir el número de datos a procesar, aunque su procesamiento sea algo más complejo. Consideramos de gran interés determinar cuál sería la opción más eficiente y la más interpretable. Además, queremos utilizar técnicas de submuestreo o sobremuestreo para equilibrar las proporciones entre clases, tratando de predecir correctamente más instancias negativas. Asimismo, incluiremos otro paso para la selección de variables con el fin de reducir el número de patrones o evaluaremos si pudiera ser mejor hacer una discretización diaria para cada variable en lugar de la actual discretización global.

Otra línea adicional de investigación que nos gustaría explorar sería emplear de manera diferente el DOR y utilizar otras métricas epidemiológicas como medidas de interés, tal y como podría ser el riesgo relativo.

Bibliografía

[1] Charu C. Aggarwal. *Data Mining*. Springer International Publishing, 2015. ISBN 978-3-319-38116-9. doi:10.1007/978-3-319-14142-8.

[2] Charu C. Aggarwal y Jiawei Han, eds. *Frequent Pattern Mining*. Springer International Publishing, 2014. doi:10.1007/978-3-319-07821-2.

[3] R. Agrawal y R. Srikant. Fast algorithms for mining association rules in large databases. En *International Conference on Very Large Databases (VLDB)*, págs. 487–499. 1994.

[4] R. Agrawal y R. Srikant. Mining sequential patterns. En *Proceedings of the Eleventh International Conference on Data Engineering*, págs. 3–14. IEEE Comput. Soc. Press, 1995. doi:10.1109/icde.1995.380415.

[5] Rakesh Agrawal, Tomasz Imielinski, y Arun Swami. Mining association rules between sets of items in large databases. En *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD 93*. ACM Press, 1993. doi:10.1145/170035.170072.

[6] Mohammed Ibrahim Al-Twijri, Jose Maria Luna, Francisco Herrera, y Sebastian Ventura. Course recommendation based on sequences: An evolutionary search of emerging sequential patterns. *Cognitive Computation*, 2022. doi:10.1007/s12559-022-10015-5.

[7] J. Alcalá-Fdez, L. Sánchez, S. García, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, J. C. Fernández, y F. Herrera. KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 13(3):307–318, 2008. doi:10.1007/s00500-008-0323-y.

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

[9] F.J. García Amigueti, F. Herrera Morillas, J.L. García Moreno, R. VelázquezGuisado, y S. Picó Tato. Manejo y reanimación del paciente quemado. *Emergencias y Catástrofes*, 1(4):217–224, 2000.

[10] Bavani Arunasalam y Sanjay Chawla. Cccs: A top-down associative classifier for imbalanced class distribution. En *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 06*. ACM Press, 2006. doi:10.1145/1150402.1150461.

[11] Iyad Batal, Dmitriy Fradkin, James Harrison, Fabian Moerchen, y Milos Hauskrecht. Mining recent temporal patterns for event detection in multivariate time series data. En *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 12*, págs. 280–288. ACM Press, 2012. doi:10.1145/2339530.2339578.

[12] Iyad Batal y Milos Hauskrecht. Constructing classification features using minimal predictive patterns. En *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM 10, págs. 869–878. Association for Computing Machinery, New York, NY, USA, 2010. ISBN 9781450300995. doi:10.1145/1871437.1871549. URL https://doi.org/10.1145/1871437.1871549.

[13] Stephen D. Bay y Michael J. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001. ISSN 1573-756X. doi:10.1023/A:1011429418057. URL https://doi.org/10.1023/A:1011429418057.

[14] Roberto J. Bayardo. Efficiently mining long patterns from databases. *ACM SIGMOD Record*, 27(2):85–93, 1998. doi:10.1145/276305.276313.

[15] Roberto J. Bayardo y Rakesh Agrawal. Mining the most interesting rules. En *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 99*, KDD '99, págs. 145–154. ACM, New York, NY, USA, 1999. ISBN 1-58113-143-7. doi:10.1145/312129.312219. URL http://doi.acm.org/10.1145/312129.312219.

[16] Roberto J. Bayardo, Rakesh Agrawal, y Dimitrios Gunopulos. Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, 4(2):217–240, 2000. ISSN 1573-756X. doi:10.1023/A:1009895914772. URL https://doi.org/10.1023/A:1009895914772.

[17] Björn Bringmann, Siegfried Nijssen, y Albrecht Zimmermann. Pattern-based classification: A unifying perspective. En *From Local Patterns to Global Models: Proceedings of the ECML/PKDD-09 Workshop (LeGo-09)*, págs. 36–50. 2009. URL http://arxiv.org/abs/1111.6191.

[18] Ivan Bruha. Pre- and post-processing in machine learning and data mining. En *Machine Learning and Its Applications (ACAI 1999)*, págs. 258–266. Springer Berlin Heidelberg, 2001. doi:10.1007/3-540-44673-7_13.

[19] Isidoro J. Casanova, Manuel Campos, Jose M. Juarez, Antonio Fernandez-Fernandez-Arroyo, y Jose A. Lorente. Evaluación de consistencia de patrones secuenciales multivariable para predecir la supervivencia de pacientes en la unidad de quemados críticos. En *XVI Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2015)*, págs. 31–40. 2015. ISBN 978-84-608-4099-2.

[20] Isidoro J. Casanova, Manuel Campos, Jose M. Juarez, Antonio Fernandez-Fernandez-Arroyo, y Jose A. Lorente. Using multivariate sequential patterns to improve survival prediction in intensive care burn unit. En *Artificial Intelligence in Medicine (AIME 2015). Lecture Notes in Computer Science, vol 9105*, tomo 9105, págs. 277–286. Springer International Publishing, 2015. ISBN 978-3-319-19551-3. doi:10.1007/978-3-319-19551-3_36. URL https://doi.org/10.1007/978-3-319-19551-3_36.

[21] Isidoro J. Casanova, Manuel Campos, Jose M. Juarez, Antonio Fernandez-Fernandez-Arroyo, y Jose A. Lorente. Impact of discretization with multivariate sequential patterns to do the classification of the survival prediction in intensive care burn unit. En *XVII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2016)*, págs. 847–856. 2016. ISBN 978-84-9012-632-5.

[22] Isidoro J. Casanova, Manuel Campos, Jose M. Juarez, Antonio Fernandez-Fernandez-Arroyo, y Jose A. Lorente. Impact of time series discretization on intensive care burn unit survival classification. *Progress in Artificial Intelligence*, 7(1):41–53, 2018. ISSN 2192-6360. doi:10.1007/s13748-017-0130-8. URL https://doi.org/10.1007/s13748-017-0130-8.

[23] Isidoro J. Casanova, Manuel Campos, Jose M. Juarez, Antonio Gomariz, Bernardo Canovas-Segura, Marta Lorente-Ros, y Jose A. Lorente. Surprising and novel multivariate sequential patterns for temporal evolution in healthcare. *Data Mining and Knowledge Discovery*, under review.

[24] Isidoro J. Casanova, Manuel Campos, Jose M. Juarez, Antonio Gomariz, Marta Lorente-Ros, y Jose A. Lorente. Using the diagnostic odds ratio to select patterns to build an interpretable pattern-based classifier in a clinical domain: Multivariate sequential pattern mining study. *JMIR Medical Informatics*, 10(8):e32319, 2022. ISSN 2291-9694. doi:10.2196/32319. URL https://doi.org/10.2196/32319.

[25] Hong Cheng, Xifeng Yan, Jiawei Han, y Chih-Wei Hsu. Discriminative frequent pattern analysis for effective classification. En *2007 IEEE 23rd International Conference on Data Engineering*, págs. 716–725. 2007. ISSN 1063-6382. doi:10.1109/ICDE.2007.367917.

[26] K.J. Cios, W. Pedrycz, R.W. Swiniarski, y L. Kurgan. *Data Mining: A Knowledge Discovery Approach*. Springer US, 2007. ISBN 978-1-4419-4120-6. doi:10.1007/978-0-387-36795-8.

[27] Ellis J. Clarke y Bruce A. Barton. Entropy and MDL discretization of continuous variables for bayesian belief networks. *International Journal of Intelligent Systems*, 15(1):61–92, 2000. doi:10.1002/(sici)1098-111x(200001)15:1<61::aid-int4>3.0.co;2-o.

[28] William W. Cohen. Fast effective rule induction. En *Proceedings of the Twelfth International Conference on Machine Learning*, págs. 115–123. Elsevier, 1995. doi:10.1016/b978-1-55860-377-6.50023-2.

[29] Gao Cong, Anthony K. H. Tung, Xin Xu, Feng Pan, y Jiong Yang. Farmer: Finding interesting rule groups in microarray datasets. En *Proceedings of the 2004 ACM SIGMOD international conference on Management of data - SIGMOD 04*, págs. 143–154. ACM Press, 2004. doi:10.1145/1007568.1007587.

[30] Alicia L. Culleiton y Lynn M. Simko. Cuidados en los pacientes quemados. *Nursing (Ed. española)*, 31(3):28–36, 2014. doi:10.1016/j.nursi.2014.07.010.

[31] Carlos E. de los Santos. *Guía Básica para el Tratamiento del Paciente Quemado*. libros-electronicos.net, Santo Domingo, República Dominicana, segunda ed$^{\underline{\text{ón}}}$., 1999. ISBN 84-95119-07-2. URL https://www.quemados.com/.

[32] Janez Demšar, Blaž Zupan, Noriaki Aoki, Matthew J. Wall, Thomas H. Granchi, y J. Robert Beck. Feature mining and predictive model construction from severe trauma patient's data. *International Journal of Medical Informatics*, 63(1-2):41–50, 2001. doi:10.1016/s1386-5056(01)00170-8.

[33] Ana Domínguez Ruiz-Huerta. *Estudio retrospectivo sobre requerimientos transfusionales en cirugía precoz del paciente quemado grave: efecto del ácido tranexámico*. phdthesis, Universidad Autónoma de Madrid. Departamento de Cirugía, 2012. URL http://hdl.handle.net/10486/9673.

[34] Guozhu Dong y Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. En *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, págs. 43–52. ACM, New York, NY, USA, 1999. ISBN 1-58113-143-7. doi:10.1145/312129.312191. URL http://doi.acm.org/10.1145/312129.312191.

[35] Guozhu Dong, Jinyan Li, y Xiuzhen Zhang. Discovering jumping emerging patterns and experiments on real datasets. En *9th International Database Conference on Heterogeneous and Internet Databases (IDC)*. 1999.

[36] Guozhu Dong y Jian Pei. *Sequence data mining*. Springer, New York, 2007. ISBN 9780387699363.

[37] Guozhu Dong, Xiuzhen Zhang, Limsoon Wong, y Jinyan Li. CAEP: Classification by aggregating emerging patterns. En Lecture Notes in Computer Science, ed., *Discovery Science*, tomo 1721, págs. 30–42. Springer Berlin Heidelberg, 1999. doi:10.1007/3-540-46846-3_4.

[38] Luisa Fernanda Durango y Francisco Vargas. Manejo médico inicial del paciente quemado. *Iatreia*, 17:54 – 61, 2004. ISSN 0121-0793. URL http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-07932004000100004&nrm=iso.

[39] Themis P. Exarchos, Markos G. Tsipouras, Costas Papaloukas, y Dimitrios I. Fotiadis. A two-stage methodology for sequence classification based on sequential pattern mining and optimization. *Data & Knowledge Engineering*, 66(3):467–487, 2008. ISSN 0169-023X. doi:https://doi.org/10.1016/j.datak.2008.05.007. URL http://www.sciencedirect.com/science/article/pii/S0169023X08000748.

[40] Hongjian Fan. *Efficient mining of interesting emerging patterns and their effective use in classification*. Tesis Doctoral, The Department of Computer Science and Software Engineering, University of Melbourne, 2004.

[41] Gang Fang, Wen Wang, Benjamin Oatley, Brian Van Ness, y Vipin Kumar. Characterizing discriminative patterns. arXiv:1102.4104, 2011.

[42] A.J. Ferreira. *Feature selection and discretization for high-dimensional data*. Tesis Doctoral, Instituto Superior Técnico, Universidade de Lisboa, 2014.

[43] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage Uday Kiran, Yun Sing Koh, y Rincy Thomas. A survey of sequential pattern mining. *Data Science and Pattern Recognition (DSPR)*, 1(1):54–77, 2017.

[44] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Roger Nkambou, Bay Vo, y Vincent S. Tseng, eds. *High-Utility Pattern Mining*. Springer International Publishing, 2019. doi:10.1007/978-3-030-04921-8.

[45] Salvador Garcia, J. Luengo, José Antonio Sáez, Victoria López, y F. Herrera. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):734–750, 2013. doi: 10.1109/tkde.2012.35.

[46] Minos N. Garofalakis, Rajeev Rastogi, y Kyuseok Shim. Spirit: Sequential pattern mining with regular expression constraints. En *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB 99, págs. 223–234. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999. ISBN 1558606157.

[47] Liqiang Geng y Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3), 2006. ISSN 0360-0300. doi:10.1145/1132960.1132963. URL http://doi.acm.org/10.1145/1132960.1132963.

[48] Shameek Ghosh. *Multivariate Sequential Contrast Pattern Mining and Prediction Models for Critical Care Clinical Informatics*. Tesis Doctoral, University of Technology Sydney, 2017. URL http://hdl.handle.net/10453/123204.

[49] Afina S. Glas, Jeroen G. Lijmer, Martin H. Prins, Gouke J. Bonsel, y Patrick M.M. Bossuyt. The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology*, 56(11):1129–1135, 2003. doi:10.1016/s0895-4356(03)00177-x.

[50] A. Gomariz. *Techniques for the Discovery of Temporal Patterns*. Tesis Doctoral, University of Murcia (Spain), University of Antwerp (Belgium), 2014. URL http://hdl.handle.net/10201/38109.

[51] Henrik Grosskreutz y Daniel Paurat. Fast and memory-efficient discovery of the top-k relevant subgroups in a reduced candidate space. En Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, y Michalis Vazirgiannis, eds., *Machine Learning and*

Knowledge Discovery in Databases, págs. 533–548. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-23780-5.

[52] Jiawei Han, Jian Pei, y Yiwen Yin. Mining frequent patterns without candidate generation. *ACM SIGMOD Record*, 29(2):1–12, 2000. doi:10.1145/335191.335372.

[53] David Hand, Heikki Mannila, y Padhraic Smyth. *Principles of data mining*. MIT Press, Cambridge, Mass, 2001. ISBN 026208290x.

[54] Zengyou He, Feiyang Gu, Can Zhao, Xiaoqing Liu, Jun Wu, y Ju Wang. Conditional discriminative pattern mining: Concepts and algorithms. *Information Sciences*, 375:1–15, 2017. doi:10.1016/j.ins.2016.09.047.

[55] Edwin O. Heierman, G. Michael Youngblood, y Diane J. Cook. Mining temporal sequences to discover interesting patterns. En *In: Proceedings of the 2004 International Conference on Knowledge Discovery and Data Mining*. 2004.

[56] Tu Bao Ho, Trong Dung Nguyen, Saori Kawasaki, Si Quang Le, Dung Duc Nguyen, Hideto Yokoi, y Katsuhiko Takabayashi. Mining hepatitis data with temporal abstraction. En *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 03*, págs. 369–377. ACM Press, 2003. doi:10.1145/956750.956793.

[57] Frank Höppner. Time series abstraction methods - a survey. En *Workshop on Knowledge Discovery in Databases*, págs. 777–786. Gesellschaft für Informatik e.V., 2002.

[58] K.H. Jacobsen. *Introduction to Health Research Methods*. Jones & Bartlett Learning, 2016. ISBN 9781284094381.

[59] Mojdeh Jalali-Heravi y Osmar R. Zaïane. A study on interestingness measures for associative classifiers. En *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, págs. 1039–1046. ACM, New York, NY, USA, 2010. ISBN 978-1-60558-639-7. doi:10.1145/1774088.1774306. URL http://doi.acm.org/10.1145/1774088.1774306.

[60] Amín D. Jaskille, Jeffrey W. Shupp, Anna R. Pavlovich, Philip Fidler, Marion H. Jordan, y James C. Jeng. Outcomes from burn injury—should decreasing mortality continue to be our compass? *Clinics in Plastic Surgery*, 36(4):701–708, 2009. doi:10.1016/j.cps.2009.05.003.

[61] Xiaonan Ji, J. Bailey, y Guozhu Dong. Mining minimal distinguishing subsequence patterns with gap constraints. En *Fifth IEEE International Conference on Data Mining (ICDM'05)*, págs. 8 pp.–. 2005. ISSN 1550-4786. doi:10.1109/ICDM.2005.96.

[62] Fernando Jiménez, Gracia Sánchez, y José M. Juárez. Multi-objective evolutionary algorithms for fuzzy classification in survival prediction. *Artificial Intelligence in Medicine*, 60(3):197–219, 2014. doi:10.1016/j.artmed.2013.12.006.

[63] Adem Karahoca, ed. *Advances in Data Mining Knowledge Discovery and Applications*. IntechOpen, 2012. ISBN 9535107488. doi:10.5772/3349. URL https://www.ebook.de/de/product/36326419/advances_in_data_mining_knowledge_discovery_and_applications.html.

[64] Randy Kerber. Chimerge: Discretization of numeric attributes. En *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI'92, pág. 123–128. AAAI Press, 1992. ISBN 0262510634.

[65] Willi Klösgen. Explora: A multipattern and multistrategy discovery assistant. En *Advances in Knowledge Discovery and Data Mining*, págs. 249–271. American Association for Artificial Intelligence, 1996.

[66] Flip Korn, Alexandros Labrinidis, Yannis Kotidis, y Christos Faloutsos. Quantifiable data mining using ratio rules. *The VLDB Journal The International Journal on Very Large Data Bases*, 8(3-4):254–266, 2000. doi:10.1007/s007780050007.

[67] Chang-Hwan Lee. A hellinger-based discretization method for numeric attributes in classification learning. *Knowledge-Based Systems*, 20(4):419–425, 2007. doi:10.1016/j.knosys.2006.06.005.

[68] Neal Lesh, Mohammed J. Zaki, y Mitsunori Ogihara. Mining features for sequence classification. En *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 99*, págs. 342–346. ACM Press, 1999. doi:10.1145/312129.312275.

[69] I-Hui Li, Jyun-Yao Huang, I-En Liao, y Jin-Han Lin. A sequence classification model based on pattern coverage rate. En *Grid and Pervasive Computing, GPC 2013*, págs. 737–745. Lecture Notes in Computer Science, vol 7861. Springer, Berlin, 2013. doi:10.1007/978-3-642-38027-3_81.

[70] J. Li, G. Dong, y K. Ramamohanarao. Jep classifier: Classification by aggregating jumping emerging patterns. Inf. téc., Univ. of Melbourne, 1999.

[71] Jiuyong Li, Ada Wai chee Fu, y Paul Fahey. Efficient discovery of risk patterns in medical data. *Artificial Intelligence in Medicine*, 45(1):77–89, 2009. doi:10.1016/j.artmed.2008.07.008.

[72] Jiuyong Li, Ada Wai-chee Fu, Hongxing He, Jie Chen, Huidong Jin, Damien McAullay, Graham Williams, Ross Sparks, y Chris Kelman. Mining risk patterns in medical data. En *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, págs. 770–775. ACM, New York, NY, USA, 2005. ISBN 1-59593-135-X. doi:10.1145/1081870.1081971. URL http://doi.acm.org/10.1145/1081870.1081971.

[73] Jiuyong Li, Jixue Liu, Hannu Toivonen, Kenji Satou, Youqiang Sun, y Bingyu Sun. Discovering statistically non-redundant subgroups. *Knowledge-Based Systems*, 67:315–327, 2014. doi:10.1016/j.knosys.2014.04.030.

[74] Wenmin Li, Jiawei Han, y Jian Pei. Cmar: accurate and efficient classification based on multiple class-association rules. En *Proceedings 2001 IEEE International Conference on Data Mining*, págs. 369–376. IEEE Comput. Soc, 2001. doi:10.1109/icdm.2001.989541.

[75] Lima. Heuristic discretization method for bayesian networks. *Journal of Computer Science*, 10(5):869–878, 2014. doi:10.3844/jcssp.2014.869.878.

[76] Jessica Lin, Eamonn Keogh, Stefano Lonardi, y Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. En *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery - DMKD03*. ACM Press, 2003. doi:10.1145/882082.882086.

[77] Bing Liu, Wynne Hsu, y Yiming Ma. Integrating classification and association rule mining. En *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, págs. 80–86. AAAI Press, 1998. URL http://dl.acm.org/citation.cfm?id=3000292.3000305.

[78] Bing Liu, Yiming Ma, y Ching-Kian Wong. Classification using association rules: Weaknesses and enhancements. En *Data Mining for Scientific and Engineering Applications*, págs. 591–605. Springer US, 2001. doi:10.1007/978-1-4615-1733-7_30.

[79] Huan Liu, Farhad Hussain, Chew Lim Tan, y Manoranjan Dash. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6(4):393–423, 2002. doi:10.1023/a:1016304305535.

[80] Xiaoqing Liu, Jun Wu, Feiyang Gu, Jie Wang, y Zengyou He. Discriminative pattern mining and its applications in bioinformatics. *Briefings in Bioinformatics*, 16(5):884–900, 2014. doi:10.1093/bib/bbu042.

[81] Xuan Liu, Pengzhu Zhang, y Dajun Zeng. Sequence matching for suspicious activity detection in anti-money laundering. En *Intelligence and Security Informatics*, págs. 50–61. Springer Berlin Heidelberg, 2008. doi:10.1007/978-3-540-69304-8_6.

[82] J. A. Lorente y Esteban A. *Cuidados intensivos del paciente quemado*. Springer, Barcelona, 1998. ISBN 9788407001783.

[83] James Malone, Kenneth Mcgarry, y Chris Bowerman. Performing trend analysis on spatiotemporal proteomics data using differential ratio rules. En *In Proceedings of the 6th EPSRC Conference on Postgraduate Research in Electronics*, págs. 103–105. 2004.

[84] Matteo Mantovani, Carlo Combi, y Milos Hauskrecht. Mining compact predictive pattern sets using classification model. En David Riaño, Szymon Wilk, y Annette Ten Teije, eds., *Artificial Intelligence in Medicine*, págs. 386–396. Springer International Publishing, Cham, 2019. ISBN 978-3-030-21642-9.

[85] D. M. Maslove, T. Podchiyska, y H. J. Lowe. Discretization of continuous features in clinical datasets. *Journal of the American Medical Informatics Association*, 20(3):544–553, 2013. doi:10.1136/amiajnl-2012-000929.

[86] Ken McGarry. A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review*, 20(1):39–61, 2005. ISSN 0269-8889. doi:10.1017/S0269888905000408. URL http://dx.doi.org/10.1017/S0269888905000408.

[87] S. Mehta, S. Parthasarathy, y Hui Yang. Toward unsupervised correlation preserving discretization. *IEEE Transactions on Knowledge and Data Engineering*, 17(9):1174–1185, 2005. doi:10.1109/tkde.2005.153.

[88] MinSal. *Guía Clínica Gran Quemado*, 2007. URL https://www.minsal.cl/portal/url/item/7222d6a3774f3535e04001011f01482e.pdf.

[89] W. Nor Haizan W. Mohamed, Mohd Najib Mohd Salleh, y Abdul Halim Omar. A comparative study of reduced error pruning method in decision tree algorithms. En *2012 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2012)*, págs. 392–397. IEEE, 2012. doi:10.1109/iccsce.2012.6487177.

[90] Robert Moskovitch y Yuval Shahar. Classification-driven temporal discretization of multivariate time series. *Data Mining and Knowledge Discovery*, 29(4):871–913, 2014. doi:10.1007/s10618-014-0380-z.

[91] Fabian Mörchen y Alfred Ultsch. Optimizing time series discretization for knowledge discovery. En *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD05*. ACM Press, 2005. doi:10.1145/1081870.1081953.

[92] Mustafa Nofal y Alaa Al Deen. Classification based on association-rule mining techniques: a general survey and empirical comparative evaluation. *Ubiquitous Computing and Communication (UBICC) Journal*, 5(3), 2010.

[93] Petra Kralj Novak, Nada Lavrač, y Geoffrey I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403, 2009. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1577069.1577083.

[94] D.W. Corne N.R. Daud. Human readable rule induction in medical data mining: A survey of existing algorithms. En *European Computing Conference, Lecture Notes in Electrical Engineering, Vol 27 LNEE*, tomo 27, págs. 787–798. 2009. doi:10.1007/978-0-387-84814-3_79.

[95] M. Ohsaki, H. Abe, S. Tsumoto, H. Yokoi, y T. Yamaguchi. Proposal of medical kdd support user interface utilizing rule interestingness measures. En *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*, págs. 759–764. 2006. ISSN 2375-9232. doi:10.1109/ICDMW.2006.137.

[96] UM Okeh y LN Ogbonna. Statistical evaluation of indicators of diagnostic test performance. *American Journal of BioScience*, 1(4):63, 2013. doi:10.11648/j.ajbio.20130104.13.

[97] Nicolas Pasquier, Yves Bastide, Rafik Taouil, y Lotfi Lakhal. Discovering frequent closed itemsets for association rules. En *Proceedings of the 7th International Conference on Database Theory*, ICDT 99, págs. 398–416. Springer-Verlag, Berlin, Heidelberg, 1999. ISBN 3540654526.

[98] Jian Pei, Jiawei Han, B. Mortazavi-Asl, Jianyong Wang, H. Pinto, Qiming Chen, U. Dayal, y Mei-Chun Hsu. Mining sequential patterns by pattern-growth: the prefixs-

pan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1424–1440, 2004. ISSN 1041-4347. doi:10.1109/TKDE.2004.77.

[99] François Petitjean, Tao Li, Nikolaj Tatti, y Geoffrey I. Webb. Skopus: Mining top-k sequential patterns under leverage. *Data Mining and Knowledge Discovery*, 30(5):1086–1111, 2016. ISSN 1573-756X. doi:10.1007/s10618-016-0467-9. URL https://doi.org/10.1007/s10618-016-0467-9.

[100] Maite Pérez, José Lara, Javier Ibañez, Leopoldo Cagigal, y Carlos Manuel León. *Guía de Actuación ante el Paciente Quemado*. Hospital R.U. Carlos Haya Málaga, Junta de Andalucia, 2006. ISBN MA-0126/2006.

[101] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986. doi:10.1007/bf00116251.

[102] Azulay R., Moskovitch R., Stopel D., Verduijn M., de Jonge E., y Shahar Y. Temporal discretization of medical time series - a comparative study. En *Workshop on Intelligent Data Analysis in Biomedicine and Pharmacology (IDAMAP 2007)*. 2007.

[103] F.J. Ruiz, C. Angulo, y N. Agell. IDD: A supervised interval distance-based method for discretization. *IEEE Transactions on Knowledge and Data Engineering*, 20(9):1230–1238, 2008. doi:10.1109/tkde.2008.66.

[104] Colleen M. Ryan, David A. Schoenfeld, William P. Thorpe, Robert L. Sheridan, Edwin H. Cassem, y Ronald G. Tompkins. Objective estimates of the probability of death from burn injuries. *New England Journal of Medicine*, 338(6):362–366, 1998. doi:10.1056/nejm199802053380604.

[105] D. Kanellopoulos. S. Kotsiantis. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, 32(1):47–58, 2006.

[106] Yuval Shahar. A framework for knowledge-based temporal abstraction. *Artificial Intelligence*, 90(1-2):79–133, 1997. doi:10.1016/s0004-3702(96)00025-2.

[107] N.N. Sheppard, S. Hemington-Gorse, O.P. Shelley, B. Philp, y P. Dziewulski. Prognostic scoring systems in burns: A review. *Burns*, 37(8):1288–1295, 2011. doi:10.1016/j.burns.2011.07.017.

[108] David L. Smith, Bruce A. Cairns, Fuad Ramadan, J. Scott Dalston, Samir M. Fakhry, Robert Rutledge, Anthony A. Meyer, y H. D. Peterson. Effect of inhalation injury,

burn size, and age on mortality: a study of 1447 consecutive burn patients. *The Journal of Trauma: Injury, Infection, and Critical Care*, 37(4):655–659, 1994. doi:10.1097/00005373-199410000-00021.

[109] Ramakrishnan Srikant y Rakesh Agrawal. Mining sequential patterns: Generalizations and performance improvements. En Lecture Notes in Computer Science, ed., *Advances in Database Technology (EDBT 1996)*, tomo 1057, págs. 1–17. Springer Berlin Heidelberg, 1996. doi:10.1007/bfb0014140.

[110] Michael Stacey y Carolyn McGregor. Temporal abstraction in intelligent clinical data analysis: A survey. *Artificial Intelligence in Medicine*, 39(1):1–24, 2007. doi:10.1016/j.artmed.2006.08.002.

[111] Chao-Ton Su y Jyh-Hwa Hsu. An extended chi2 algorithm for discretization of real value attributes. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):437–441, 2005. doi:10.1109/tkde.2005.39.

[112] Pang-Ning Tan, Vipin Kumar, y Jaideep Srivastava. Selecting the right interestingness measure for association patterns. En *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, págs. 32–41. ACM, New York, NY, USA, 2002. ISBN 1-58113-567-X. doi:10.1145/775047.775053. URL http://doi.acm.org/10.1145/775047.775053.

[113] Joyce Tobiasen, John M. Hiebert, y Richard F. Edlich. The abbreviated burn severity index. *Annals of Emergency Medicine*, 11(5):260–262, 1982. doi:10.1016/s0196-0644(82)80096-6.

[114] Tudor Toma, Ameen Abu-Hanna, y Robert-Jan Bosman. Discovery and integration of univariate patterns from daily individual organ-failure scores for intensive care mortality prediction. *Artificial Intelligence in Medicine*, 43(1):47–60, 2008. doi:10.1016/j.artmed.2008.01.002.

[115] Tudor Toma, Robert-Jan Bosman, Arno Siebes, Niels Peek, y Ameen Abu-Hanna. Learning predictive models that use pattern discovery—a bootstrap evaluative approach applied in organ functioning sequences. *Journal of Biomedical Informatics*, 43(4):578–586, 2010. doi:10.1016/j.jbi.2010.03.004.

[116] Giulia Toti, Ricardo Vilalta, Peggy Lindner, Barry Lefer, Charles Macias, y Daniel Price. Analysis of correlation between pediatric asthma exacerbation and exposure to

pollutant mixtures with association rule mining. *Artificial Intelligence in Medicine*, 74:44–52, 2016. doi:10.1016/j.artmed.2016.11.003.

[117] Giulia Toti, Ricardo Vilalta, Peggy Lindner, y Daniel Price. Effect of the definition of non-exposed population in risk pattern mining. En *5th Workshop on Data Mining for Medicine and Healthcare, SDM 2016*. 2016.

[118] Vincent S. M. Tseng y Chao-Hui Lee. CBS: A new classification method by using sequential patterns. En *Proceedings of the 2005 SIAM International Conference on Data Mining (SDM 2005)*, págs. 596–600. Society for Industrial and Applied Mathematics, 2005. doi:10.1137/1.9781611972757.68.

[119] K.B. Irani U.M. Fayyad. Multi-interval discretization of continuous-valued attributes for classification learning. En *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93)*. 1993. URL http://hdl.handle.net/2014/35171.

[120] Matthijs van Leeuwen y Arno Knobbe. Non-redundant subgroup discovery in large and complex data. En Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, y Michalis Vazirgiannis, eds., *Machine Learning and Knowledge Discovery in Databases*, págs. 459–474. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-23808-6.

[121] Florian Verh y Sanjay Chawla. Using significant, positively associated and relatively class correlated rules for associative classification of imbalanced datasets. En *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, págs. 679–684. IEEE, 2007. doi:10.1109/icdm.2007.63.

[122] Geoffrey I. Webb. Discovering significant patterns. *Machine Learning*, 68(1):1–33, 2007. doi:10.1007/s10994-007-5006-x.

[123] Geoffrey I. Webb, Shane Butler, y Douglas Newlands. On detecting differences between groups. En *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, págs. 256–265. ACM, New York, NY, USA, 2003. ISBN 1-58113-737-0. doi:10.1145/956750.956781. URL http://doi.acm.org/10.1145/956750.956781.

[124] Li Wei y Eamonn Keogh. Semi-supervised time series classification. En *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 06*. ACM Press, 2006. doi:10.1145/1150402.1150498.

[125] Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. En *Principles of Data Mining and Knowledge Discovery*, págs. 78–87. Springer Berlin Heidelberg, 1997. doi:10.1007/3-540-63223-9_108.

[126] Qingxiang Wu, David A. Bell, Girijesh Prasad, y Thomas Martin McGinnity. A distribution-index-based discretizer for decision-making with symbolic AI approaches. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):17–28, 2007. doi:10.1109/tkde.2007.250582.

[127] Shanshan Wu, Yanchang Zhao, Huaifeng Zhang, Chengqi Zhang, Longbing Cao, y Hans Bohlscheid. Debt detection in social security by adaptive sequence classification. En Dimitris Karagiannis y Zhi Jin, eds., *Knowledge Science, Engineering and Management, KSEM 2009*, págs. 192–203. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-10488-6.

[128] Zhengzheng Xing, Jian Pei, y Eamonn Keogh. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, 12(1):40–48, 2010. doi:10.1145/1882471.1882478.

[129] Xiaoxin Yin y Jiawei Han. Cpar: Classification based on predictive association rules. En *Proceedings of the 2003 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2003. doi:10.1137/1.9781611972733.40.

[130] Janzekovic Z. A new concept in the early excision and immediate grafting of burns. *The Journal of Trauma: Injury, Infection, and Critical Care*, 10(12):1103–1108, 1970.

[131] M. J. Zaki, S. Parthasarathy, M. Ogihara, y W. Li. New algorithms for fast discovery of association rules. En *Proceedings of the 3rd ACM SIGKDD international conference on Knowledge discovery and data mining, KDD 1997*, págs. 283–286. 1997.

[132] Mohammed J. Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1/2):31–60, 2001. doi:10.1023/a:1007652502315.

[133] Mohammed J. Zaki, Neal Lesh, y Mitsunori Ogihara. Planmine: Predicting plan failures using sequence mining. *Artificial Intelligence Review*, 14(6):421–446, 2000. ISSN 1573-7462. doi:10.1023/A:1006612804250. URL https://doi.org/10.1023/A:1006612804250.

[134] D. A. Zighed, S. Rabaséda, y R. Rakotomalala. FUSINTER: A method for discretization of continuous attributes. *International Journal of Uncertainty,*

Fuzziness and Knowledge-Based Systems, 06(03):307–326, 1998. doi:10.1142/s0218488598000264.