Universidad de las Ciencias Informáticas

Facultad 3

Comparación de clasificadores asociativos

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autor: Ali Corrales Figueroa

Tutores: Ing. Guillermo M. Negrín Ortiz

MSc. Julio César Díaz Vera

La Habana, 2019

DECLARACIÓN DE AUTORÍA

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste se firma la presente a los ____ días del mes de _____ del año _____.

_____

Firma del Autor

Ali Corrales Figueroa

_____

Firma del Tutor

Ing. Guillermo M. Negrín Ortiz

_____

Firma del Tutor

MSc. Julio Cesar Diaz Vera

DATOS DE CONTACTO

MSc. Julio Cesar Díaz Vera

Universidad Martha Abreu, Villa Clara, Cuba.

Correo electrónico: jcdiaz@uci.cu

Ing. Guillermo Manuel Negrín Ortiz

Universidad de las Ciencias Informáticas, La Habana, Cuba.

Correo electrónico: gmnegrin@uci.cu

DEDICATORIA

A mis padres,

por su ejemplo de perseverancia y dedicación,

por su estimulante optimismo y confianza,

por su amor sin límites

AGRADECIMIENTOS

RESUMEN

Recientemente, los estudios de minería de datos se están llevando a cabo con éxito para estimar varios parámetros en una variedad de dominios. Las técnicas de minería de datos han atraído la atención de la industria de la información y de la sociedad en general, debido a la gran cantidad de datos y la inminente necesidad de convertirlos en conocimiento útil. Sin embargo, el uso efectivo de los datos en algunas áreas todavía está en desarrollo, como es el caso de los deportes, que en los últimos años ha presentado un ligero crecimiento; en consecuencia, muchas organizaciones deportivas han empezado a ver que hay una gran cantidad de conocimiento inexplorado en los datos extraídos por ellos. Por lo tanto, este artículo presenta una revisión sistemática de la minería de datos deportivos. Con respecto a los años 2010 a 2018, se encontraron 31 tipos de investigación en este tópico. En función de estos estudios, presentamos el panel actual, los temas, la base de datos utilizada, las propuestas, los algoritmos y las oportunidades de investigación. Nuestros hallazgos brindan una mejor comprensión de los potenciales de minería de datos deportivos, además de motivar a la comunidad científica a explorar este tema oportuno e interesante.

Palabras clave: Reglas de asociación, Reglas de clasificación, Clasificadores Asociativos, Minería de Datos

Índice de contenidos

Índice de tablas

Índice de figuras

INTRODUCCIÓN

La clasificación es una tarea presente en cualquier actividad humana. El desarrollo tecnológico vertiginoso que experimenta la sociedad, apoyada en las Tecnologías de la Información y las Comunicaciones (TIC) hacen que la tarea de clasificación sea un foco de estudio dentro de la inteligencia artificial (Aldape Pérez, 2017). En minería de datos la clasificación es una tarea de reconocimiento automático de patrones. La misma permite la identificación o selección de rasgos o características en los objetos, procesos, fenómenos o conceptos para ubicarlos en una determinada clase de manera automática. Todo lo anterior se hace de manera inductiva, de lo simple a lo complejo, de lo concreto a lo abstracto y con una buena dosis de ensayo y error, por lo que constituye una tarea de aprendizaje supervisado. Varias son las áreas en las que se ha aplicado la clasificación como parte de la inteligencia artificial. Entre estas, se ha utilizado para comprender el motivo de hospitalización de pacientes (García Floriano, y otros, 2015), para mejorar el desempeño de sistemas de toma de decisiones (Itzamá López, y otros, 2014), para determinar la relevancia de la información que puedan aportar los clientes para producir revisiones, entre otros (CGTI, 2019).

Entre los principales enfoques utilizados para llevar a cabo la clasificación se encuentran: Los enfoques estadístico-probabilísticos que se basan específicamente en el teorema de Bayes. Los clasificadores basados en métricas que también usan un enfoque probabilístico, se basan en el concepto de métrica y en las propiedades de los espacios métricos para realizar la clasificación. También existen los árboles de decisión, basados en la teoría de grafos, máquina de soporte vectorial que utiliza un entrenamiento para maximizar el margen de los patrones en el límite de las clases, algoritmos genéticos, basados en la evolución humana, las redes neuronales basadas en la emulación del comportamiento de esta unidad básica del cerebro humano y los clasificadores basados en reglas.

Una forma de representar el conocimiento es precisamente el uso de reglas de asociación. Por su simplicidad son ampliamente utilizadas y permiten describir un

patrón haciendo uso de una expresión *if-then.* (Si ocurre A entoces ocurre B) el significado semántico de esta expresión apunta al hecho de que los elementos A y B ocurren de manera correlacionada en las instancias de una base de datos.

Ejemplo.

$$Pan, Leche \rightarrow Mantequilla, Queso$$

$$Pañales, Hombre \rightarrow Cerveza$$

Una regla de clasificación es una regla de asociación con una clase específica como consecuente (Slimani, 2015). Siendo la clase una distinción o categoría que toma un atributo que es objeto de estudio (Springer, 2018).

Ejemplo:

Dado que una persona está solicitando un préstamo a un determinado banco, este hace el siguiente análisis para otorgárselo o no, teniendo como resultado con las siguientes entradas,

$$Ingreso[Alto], Deuda[No] \rightarrow Préstamo[Si]$$

El préstamo sería positivo, asumiendo que las entradas toman valores diferentes tendríamos,

$$Ingreso[Bajo], Deuda[Si] \rightarrow Préstamo[No]$$

Los clasificadores que utilizan reglas de asociación de clasificación se denominan clasificadores asociativos. Los clasificadores asociativos son muy utilizados (Gredos, 2018) (García Floriano, y otros, 2015) (Slimani, 2015), debido a que su estructura facilita la comprensión del modelo por parte de los especialistas, ya que el ser humano comúnmente utiliza reglas para hacer inferencias lo que favorece la

clasificación. De esta forma este modelo es favorecida sobre las variantes que usan modelos de caja negra.

La clasificación asociativa se divide en dos partes. La primera es la extracción de todas las reglas de asociación relacionadas con las clases y la segunda es la selección de un subconjunto de reglas para construir el clasificador.

La clasificación en si no es un proceso sencillo pues no existe un clasificador ideal en el que cada vez que se reciba un patrón haya una regla en la base de conocimiento que permita clasificarla unívocamente, esto hace que dentro de esta base de conocimiento para un patrón a clasificar exista más de una regla a la que se le pueda otorgar su clase. Por lo tanto, uno de los momentos en los que más se le dedica tiempo es precisamente en romper esos empates, pues de esto depende en gran medida la correctitud de la clasificación. En algunos dominios es más importante ganar en cantidad de patrones clasificados mientras que en otros es más importante que, aunque no se llegue a clasificar el total de elementos, la clasificación sea precisa.

Para ejemplificar esto último se tienen dos dominios, el primero de un supermercado que desea clasificar clientes de acuerdo a su volumen de compra. En este caso es más importante clasificar la mayor cantidad de clientes, aunque haya clientes que no estén clasificados dentro de la categoría correcta. Por otro lado, si se intentara clasificar la aptitud de pacientes en una unidad de trasplantes no sería tan importante cubrir todos los patrones sino lograr una clasificación mucho más precisa para otorgarle el órgano al paciente correcto.

Existen diferentes alternativas para medir el desempeño de un clasificador remarcando que ante diferentes situaciones y dominios no se usan las mismas. Para determinar la precisión de la clasificación se utilizan los métodos siguientes: Sensibilidad, Específica, Precisión, Valor de predicción Negativa, Error de clasificación, Prevalencia. El enfoque de esta investigación es la creación de un marco para la comparación de clasificadores asociativos utilizando diferentes métricas.

Aunque gran cantidad de algoritmos se han desarrollado con diferentes variantes de pre-procesamiento de los datos, de manera general los autores no acostumbran a publicar detalles de implementación de sus experimentos y se hace muy difícil replicar las condiciones en las que estos tienen éxito. Finalmente, se llega a la conclusión que la experimentación bajo estas condiciones es ineficiente.

En esta investigación se plantea la implementación de varios algoritmos para establecer una comparación, teniendo en cuenta la misma entrada con el mismo pre-procesamiento, utilizando varias métricas. De esta manera se garantizan los mecanismos de inclusión de nuevos algoritmos de manera tal que sea más eficiente la ejecución de nuevos experimentos.

Con motivo de esta situación y para guiar el desarrollo de la investigación se plantea el siguiente **problema a resolver**: ¿Cuál es el desempeño de los principales clasificadores asociativos con respecto a la precisión de la clasificación en diferentes tipos de datasets[1]?

Tomando en cuenta el problema antes propuesto se define como **objeto de estudio**: los clasificadores asociativos.

De esta forma se determina como **objetivo general**: Determinar el desempeño de diferentes clasificadores asociativos ante distintos datasets.

Se desglosan del objetivo general los siguientes **objetivos específicos**:

1. Establecer el marco conceptual para el desarrollo de la investigación.
2. Definir un mecanismo para la comparación de clasificadores asociativos.
3. Comparar los clasificadores asociativos.

Campo de acción:

las técnicas de comparación de los clasificadores asociativos.

1Colección de conjuntos de información relacionados que se componen de elementos separados pero que pueden ser manipulados como una unidad por una computadora.

Para dar cumplimiento a los objetivos propuestos se definen las siguientes **tareas de la investigación**:

1. Recopilación de la bibliografía referente al tema.
2. Selección de la bibliografía.
3. Análisis de la bibliografía.
4. Selección de los elementos esenciales que componen los clasificadores asociativos.
5. Implementación de los clasificadores seleccionados.
6. Ejecución de las implementaciones sobre datasets.
7. Presentación de los resultados.

Por lo que se definen las siguientes **preguntas de la investigación**:

- ¿Qué es un clasificador asociativo?
- ¿Cuáles son los tipos de clasificadores que existen?
- ¿Cómo se mide la precisión de un clasificador asociativo?
- ¿Qué datasets son los más utilizados para evaluar clasificadores asociativos?
- ¿Cómo clasificar un datasets?
- ¿Qué test estadístico se aplican en la evaluación?
- ¿Qué clasificador asociativo tiene mejor desempeño?

Como **posible resultado** una vez finalizado el presente trabajo, se tendrá un modelo de desempeño de diferentes clasificadores asociativos.

Descripción de los capítulos

El presente trabajo de diploma está estructurado en 3 capítulos de la siguiente forma:

Capítulo 1: Se presentan los conceptos sobre los que se cimienta el trabajo de investigación realizado. Primeramente, se exponen los conceptos de regla de asociación, clasificador asociativo y clase. Se define una planificación para la ejecución de la investigación. Finalmente se introducen elementos de conocimiento previo para analizar su efecto en la comparación de clasificadores asociativos.

Capítulo 2: Se procede a la comparación de los clasificadores asociativos, comenzando con la extracción de las reglas de clasificación a través de un algoritmo de extracción y aplicándole un respectivo pre-procesamiento. Finalizando con la implementación de los clasificadores asociativos.

Capítulo 3: Validación de los datos obtenidos en el capítulo anterior con la implementación de las métricas definidas para establecer el experimento de comparación.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA

Introducción

En este capítulo se presentan los elementos sobre los que se cimienta la investigación, mediante la revisión sistemática de la literatura, con el propósito de identificar investigaciones utilizando los clasificadores asociativos, así como el método y la planificación que se tuvo en cuenta para la revisión de dicha literatura. Se presenta el panorama actual de la investigación, las bases de datos utilizadas, los criterios de búsquedas, así como los criterios de inclusión, exclusión y calidad tenidos en cuenta para la selección final de los artículos para el estudio que se persigue en la presente investigación.

1.1 Método de revisión sistemática

Revisión sistemática es el proceso en el que se realiza una revisión de aspectos cuantitativos y cualitativos de estudios primarios, con el objetivo de resumir la información existente respecto de un tema en particular.

1.2 Planificación de la investigación

En esta etapa se abordan la definición de las preguntas científicas, la intervención de la especificación de intereses, la identificación de bases de datos, la definición de palabras clave, las estrategias de búsqueda, los criterios de inclusión, exclusión y calidad de los artículos (Brereton, y otros, 2007). Por lo tanto, los siguientes problemas se definieron como preguntas de investigación (PI):

PI1. ¿Cuál es el panorama de las investigaciones actuales?
PI2. ¿Cuáles son las técnicas más utilizadas?
PI3. ¿Cuál es la distribución temporal de los artículos?
PI4. ¿Cuáles son los dataset analizados?
PI5. ¿Cuál fue el resultado de las investigaciones?

Por lo general, los criterios de inclusión, exclusión y calidad se determinan después de la definición de las preguntas de investigación (Smith, y otros, 2011).

Por lo tanto, en la tabla que se presenta a continuación se establecen los criterios utilizados en este estudio.

Tabla 1: Criterios de inclusión, exclusión y calidad

	Criterios
Inclusión	• Estudios en idioma Inglés • Relacionados con los clasificadores asociativos • Estudios relevantes en las diferentes esferas de la sociedad
Exclusión	• No cumplen con los criterios de calidad • Redactados en idioma diferente al inglés • No aplican a los clasificadores asociativos • Publicados antes del 2014
Calidad	• Estudios con buenos resultados • Estudios con diferentes propósitos y resultados

Fuente: Elaboración propia

Los aspectos de este proceso pueden incluir decisiones sobre el tipo de revisiones que deben incluirse en la investigación, que se utiliza para gestionar los criterios de selección en un subconjunto de estudios primarios (Menezes, y otros, 2016). Por lo tanto, para garantizar la calidad de esta revisión, cada artículo encontrado se analizará de acuerdo con el Autor, Titulo, Dataset, Algoritmo y Validación. Además, esta investigación utilizó las siguientes bases de datos electrónicas para encontrar los artículos relacionados con la clasificación de reglas de asociación:

- Biblioteca digital ACM: dl.acm.org
- Biblioteca digital IEEE Xplore: ieeexplore.ieee.org
- Science Direct: www.sciencedirect.com
- Académico semántico: www.semanticscholar.org

Finalmente, el método seleccionado para buscar en estas bases de datos fue la recuperación booleana. Esencialmente, divide un espacio de búsqueda, identificando un subconjunto de documentos en una colección, de acuerdo con los criterios de consulta (Karimi, y otros, 2010). En el caso de la presente

investigación, la clave es la siguiente cadena: (*association rules AND associative classifier OR class rules) AND (model reduction OR redundancy*). Una vez sentadas las bases en esta etapa, se procede a realizar el plan de ejecución de la presente investigación.

1.3 Plan de ejecución

Esta etapa incluye cinco pasos (Brereton, y otros, 2007):

1. Realizar la búsqueda en las bases de datos seleccionadas.
2. Comparación de resultados de búsquedas para excluir trabajos repetidos.
3. La aplicación de los criterios de inclusión, exclusión y calidad.
4. Evaluación de todos los estudios que pasaron la revisión inicial.
5. Síntesis de datos.

La Figura 1 muestra cómo se llevaron a cabo los pasos descritos anteriormente para la revisión sistemática de la literatura. Básicamente, la primera fase consistió en ejecutar las cadenas de búsqueda en todas las bases de datos, en la que se obtuvo como resultado de la búsqueda 991 artículos. Por lo tanto, para ayudar a la revisión y lograr una mejor precisión y confiabilidad, se usó la herramienta StArt (Estado del arte a través de revisiones sistemáticas). Esta herramienta tiene el propósito de apoyar a los investigadores en su análisis sistemático (Fabbr, y otros, 2014). La misma utiliza las extensiones BibTeX (archivo de formato bibliográfico utilizado en los documentos de LaTeX) para realizar estos análisis. Por lo tanto, estas extensiones también se extrajeron de las bases de datos mencionadas anteriormente. Es importante tener en cuenta que los archivos BibTeX se exportaron sin ningún filtro, lo que explica el número de investigaciones devueltas.

A partir de entonces, se eliminaron los trabajos publicados antes de 2014, que arrojaron una cantidad de 538 títulos (453 artículos rechazados). Sin embargo, para refinar la búsqueda y eliminar los artículos que estaban fuera del alcance de esta revisión, se aplicó un análisis cuidadoso en los títulos, palabras clave y resúmenes, de acuerdo con los criterios de exclusión (ver tabla 1). Se eliminaron 413 trabajos, quedando un conjunto de preselección de 42 títulos para análisis de calidad.

Finalmente, después de la preselección de los trabajos, se realizó una síntesis de los datos, con el objetivo de aplicar una evaluación basada en los criterios de calidad establecidos. De ese modo, de los 42 artículos, 27 fueron eliminados, ya que no describían los métodos o técnicas aplicadas, lo que llevó a un conjunto final de 15 artículos con información relevante sobre los clasificadores asociativos (ver Figura 1).

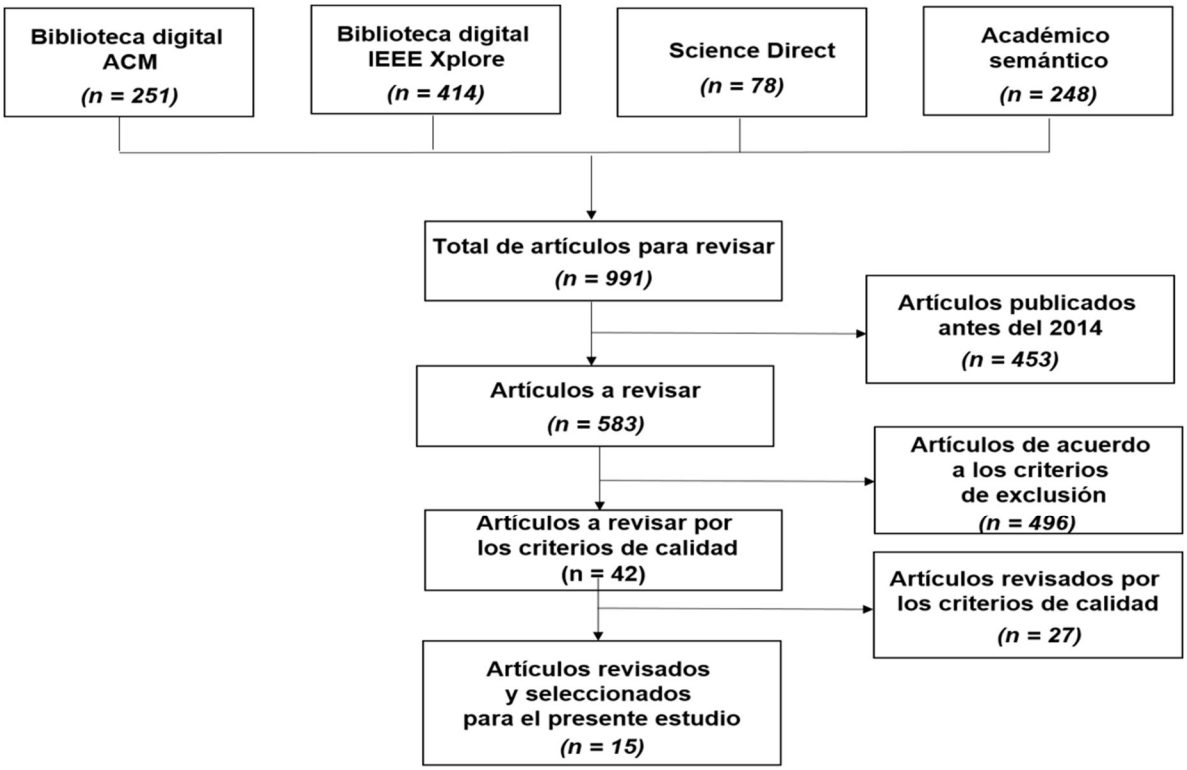


Figura 1: Flujo para la revisión sistemática de la literatura

Fuente: Elaboración propia

A continuación, se muestra un resumen de los 15 artículos seleccionados para el estudio que propone la presente investigación:

No. 1	**Título:** Un clasificador asociativo que adopta reglas ponderadas difusas basadas en la ganancia de información
Autores: Yi Chai, Guixia Kang, Ningbo Zhang, Yanyan Guo y Jingning Wang **Dataset:** Real con datos médicos **Breve descripción:** El documento apunta a proponer un enfoque para construir un clasificador asociativo basado en la regla de asociación ponderada difusa	

(IGWFAC), siendo este el algoritmo propuesto en la investigación. El mismo emplea una estrategia de selección de atributos basada en la ganancia de información para determinar el grado de importancia de los atributos y asigna las ponderaciones correspondientes de manera que a los atributos más importantes se les preste más atención. Los resultados experimentales muestran que hay una mejora en la precisión de la clasificación y una reducción en la redundancia de reglas.

No. 2	**Título:** Algoritmo distribuido de clase múltiple basado en clasificación de clase utilizando RIPPER

Autores: Aruna Govada, Varsha S. Thomas, Ipsita Samal y Sanjay K. Saha

Dataset: Sintético con datos generales

Breve descripción: Se propone un algoritmo distribuido de clasificación basado en reglas de múltiples clases (DiRUC) que implementa repetidas podas incrementales para la reducción de la producción (RIPPER) y luego se fusiona en un nivel global de manera distribuida. El algoritmo primero construye los conjuntos de reglas locales para los datos distribuidos y luego, en cada iteración, los modelos locales se envían de una ubicación a otra. Finalmente, el modelo global se construye mediante la fusión eficiente de estos modelos locales y se pone a disposición en cada sitio para una mayor predicción de las etiquetas de clase. El análisis de rendimiento del algoritmo se realiza para los cinco conjuntos de datos con diferentes parámetros y el resultado muestra que el enfoque propuesto DiRUC supera el rendimiento normal de RIPPER.

No. 3	**Título:** Aprendizaje inductivo basado en la teoría de conjuntos rudos para la toma de decisiones médicas

Autores: Ahmad Taher Azar, Nidhal Bouaynaya y Robi Polikar

Dataset: Real con datos médicos

Breve descripción: Este artículo propone un algoritmo que utiliza el aprendizaje inductivo y la teoría de conjuntos aproximados (ILRS) para analizar los datos

clínicos disponibles en un archivo de paciente (registros). Se demuestra que el mismo es capaz de reducir el número de características disponibles en un conjunto de núcleo más pequeño que describe con precisión el sistema de información. También que se puede evaluar cuantitativamente el nivel de dependencia de la patología considerada, o característica de decisión, en un conjunto dado de características o atributos de condición. Además, que es capaz de hacer frente a información incierta e incompleta. Se considera un estudio de caso de un sistema de información incompleta obtenido durante la canulación de las arterias pelis radial y dorsal. Se muestra también cómo el ILRS logra eliminar la redundancia y determinar los atributos de condición más significativos para un conjunto dado de atributos de decisión de datos contaminados con incertidumbre. Una clasificación de múltiples clases con relaciones de preferencia se presenta a través de un conjunto de reglas de decisión. A diferencia del análisis estadístico de los datos clínicos, la confiabilidad del algoritmo ILRS propuesto es independiente del tamaño de los datos.

No. 4	**Título:** ¿Cuáles son las diferencias entre clasificadores bayesianos y clasificadores de información mutua?

Autores: Bao-Gang Hu

Dataset: Sintético con datos generales

Breve descripción: En este artículo, se examinan clasificaciones binarias con o sin opción de rechazo, tanto los clasificadores de información bayesiana como los de información mutua. Las reglas de decisión generales se derivan de los clasificadores bayesianos con distinciones sobre tipos de error y tipos de rechazo. Se lleva a cabo un análisis formal para revelar la redundancia de parámetros de los términos de costo cuando se aplican las clasificaciones de abstención. La redundancia implica un problema intrínseco de no consistencia para interpretar los términos de costo. Si no se dan datos a los términos de costo, demostramos la debilidad de los clasificadores bayesianos en clasificaciones con desequilibrio de clase. Por el contrario, los clasificadores de información mutua pueden proporcionar una solución objetiva a partir de los datos dados, que muestra un

equilibrio razonable entre los tipos de error y los tipos de rechazo. Se dan ejemplos numéricos del uso de dos tipos de clasificadores para confirmar las diferencias, incluidos los casos extremadamente desequilibrados en las clases. Finalmente, se resume brevemente los clasificadores bayesianos y de información mutua en términos de sus ventajas y desventajas de aplicación, respectivamente.

No. 5	**Título:** Función de transferencia local aproximada para la detección de trastornos cardíacos mediante el uso de sonidos cardíacos

Autores: Aboul Ella Hassanien, Mostafa a. Salama y Jan Platos

Dataset: Real con datos médicos

Breve descripción: Se aplicaron clasificadores alternativos a los mismos datos para comparación, incluida la Máquina de vectores de soporte (SVM), Red Bayesiana Oculta Naive (HNB), Red Bayesiana (BN), Árbol Bayesiano Naive (NBT), Árbol de Decisión (DT), Optimización mínima secuencial (SMO), tabla de decisión (DT), bosque de rotación (RoF) y bosque aleatorio (RF); sin embargo, su desempeño para los mismos problemas de diagnóstico fue menor que la función de transferencia local aproximada propuesta.

No. 6	**Título:** Reglas FCBF: un método de selección de características para conjuntos de datos de múltiples etiquetas

Autores: Shima Kashef

Dataset: Sintético con datos generales

Breve descripción: En este documento, se introduce un nuevo algoritmo de selección de características de etiquetas múltiples basado en el método de selección de características del filtro basado en correlación rápida (FCBF), que es un enfoque de filtro para conjuntos de datos de etiqueta única. La principal contribución de este documento corresponde al paso en el que las características efectivas y útiles se distinguen de las redundantes en el método FCBF. Para hacerlo, se implementan tres reglas y cuando una de estas reglas no se cumple, la característica no se elimina. El método propuesto, junto con los tres métodos

de selección de características de etiqueta múltiple propuestos recientemente, se aplican en 6 conjuntos de datos de etiqueta múltiple estándar para evaluación. Los resultados obtenidos indican la gran capacidad del algoritmo propuesto para encontrar el mejor subconjunto de características, en comparación con otros algoritmos.

No. 7	**Título:** Un enfoque basado en la minería para la enumeración eficiente de estructuras algebraicas
Autores: Majid Ali Khan, Nazeeruddin Mohammad y Shahabuddin Muhammad **Dataset:** Real con datos de administración de red **Breve descripción:** Las estructuras algebraicas son estructuras matemáticas bien estudiadas en álgebra abstracta con aplicaciones en muchos campos de la seguridad informática, como la criptografía y la autenticación. En este artículo, se presenta un enfoque basado en la minería para la ruptura de simetría en estructuras algebraicas. El enfoque reduce el número de estructuras redundantes al identificar reglas basadas en patrones recurrentes en las estructuras conocidas anteriormente. Estas reglas se utilizan como restricciones en un solucionador de restricciones líder (herramientas o de Google). Cuando se aplica a un bucle de IP, una clase especial de estructuras algebraicas, las reglas reducen el número de soluciones redundantes que resultan en una mejora significativa del tiempo.	

No. 8	**Título:** Un conjunto basado en la programación genética multi-objetivo para la selección y clasificación simultáneas de características
Autores: Kaustuv Nag y Nikhil R. Pal **Dataset:** Reales con datos genéticos **Breve descripción:** Se presenta un algoritmo integrado para la selección simultánea de funciones (FS) y el diseño de diversos clasificadores utilizando una programación genética multiobjetiva (GP) de estado estable, que minimiza tres objetivos: 1) falsos positivos (FP); 2) falsos negativos (FNs); y 3) el número de nodos de hoja en el árbol. Nuestro método divide un problema de clase C en problemas de clasificación binario. Se desarrolla c conjuntos de programas	

genéticos para crear c conjuntos. Se probó la propuesta en ocho conjuntos de microarrays y 11 conjuntos de datos de texto con un número diverso de clases (de 2 a 44), un gran número de funciones (de 2000 a 49151) y una alta relación de características a muestra (de 1.03 a 273.1). Se comprobó además en un esquema de GP con dos objetivos que no utiliza ninguna estrategia de reducción de tamaño de regla y FS; y con cuatro métodos de clasificación en combinación con seis algoritmos de selección de características y un conjunto completo de características. Se demostró que el esquema funciona mejor para 380 de 474 combinaciones de conjuntos de datos, algoritmo y método FS.

No. 9	**Título:** Método de fusión de elementos de situación de seguridad de red basado en ontología

Autores: Cheng Si, Hongqi Zhang, Yongwei Wang y Jiang Liu

Dataset: Sintético con datos de seguridad de red

Breve descripción: Como la investigación actual no puede resolver el problema de hacer que los elementos de situación de seguridad de red heterogéneos de múltiples fuentes se describan de manera uniforme, se propone un método de fusión de elementos de situación de seguridad de red basado en ontología. En primer lugar, se construye un modelo de fusión que contiene un entorno de red, vulnerabilidad de red, ataque de red, incidente de seguridad de red y sensor como clase clave. En segundo lugar, tres reglas de fusión que contienen agregación de alertas, verificación de alertas y reconstrucción de la sesión de ataque se formulan mediante el uso del Lenguaje de reglas web mejorado con consultas semánticas. Finalmente, el ejemplo de aplicación muestra que el método puede hacer que los elementos de la situación se describan de manera uniforme.

No. 10	**Título:** Post-minería de reglas de asociación: Técnicas para la eficacia extracción de conocimiento

Autores: Yanchang Zhao, Chengqi Zhang y Longbing Cao

Dataset: Sintético con datos generales

Breve descripción: Examina el post-análisis y post-minería de las reglas de asociación para obtener conocimiento útil de un gran número de reglas descubiertas y presenta una vista sistemática del tema anterior. Introduce una investigación actualizada sobre la extracción de conocimiento útil de un gran número de reglas de asociación descubiertas, y cubre interés, post-minería, selección de reglas, resumen, representación y visualización de reglas de asociación, así como nuevas formas de reglas de asociación y nuevas tendencias de minería de reglas de asociación.

No. 11	**Título:** Un estudio experimental de tres fórmulas de clasificación de reglas diferentes en la clasificación asociativa

Autores: Neda Abdelhamid, Aladdin Ayesh y Fadi Thabtah

Dataset: Reales con datos genéticos

Breve descripción: Este documento investiga el impacto de la clasificación de reglas antes de construir el clasificador en la minería de clasificación asociativa. Se predice que la clasificación de reglas puede desempeñar un papel importante en la determinación de la precisión de los clasificadores y también puede considerarse un paso previo a la ejecución de las reglas. Se han utilizado dieciséis conjuntos de datos diferentes del repositorio de datos DCI en los experimentos, y las bases de las comparaciones son la tasa de error y el número de reglas. Los resultados revelan que la clasificación de reglas desempeña un papel importante en la determinación del subconjunto de reglas que se utilizará en el paso de predicción y, de hecho, afecta el poder predictivo de dicho subconjunto.

No. 12	**Título:** Una encuesta de algoritmo de clasificación asociativa

Autores: Sohil Gambhir

Dataset: Sintético con datos generales

Breve descripción: En el campo de la minería de datos, se procesa una gran cantidad de datos para obtener una pequeña cantidad de datos útiles. Para optimizar la eficiencia se combinan dos métodos clásicos de minería de datos, a

saber, la minería de reglas de asociación y la minería de reglas clásica. El nuevo método se llama clasificación asociativa. Este trabajo es una encuesta de los principales métodos de clasificación asociativa. Después de este estudio se puede hacer una mejor comparación de varios métodos de clasificación asociativa.

No. 13	**Título:** Clasificación asociativa con reglas positivas y negativas estadísticamente significativas

Autores: Jundong Li y Osmar R. Zaïane

Dataset: Sintético con datos generales

Breve descripción: Se propone un nuevo clasificador asociativo que se basa en reglas de asociación de clasificaciones positivas y negativas que muestran dependencias estadísticamente significativas. Los resultados experimentales en conjuntos de datos del mundo real muestran que el método logra un rendimiento competitivo o incluso mejor que los conocidos clasificadores basados en reglas y asociativos en términos de precisión de clasificación y eficiencia computacional.

No. 14	**Título:** Clasificación mejorada basada en la regla de minería

Autores: Bangaru Veera Balaji y M. Tech Final

Dataset: Sintético con datos generales

Breve descripción: La minería y clasificación de reglas de asociación son dos técnicas importantes de la minería de datos en el proceso de descubrimiento de conocimiento. La integración de estas dos técnicas es un importante foco de investigación y tiene muchas aplicaciones en la minería de datos. La integración de estas dos técnicas ha producido nuevos enfoques llamados Minería de Reglas de Asociación de Clase o Técnica de Clasificación Asociativa. Estos dos enfoques combinados proporcionan una mejor precisión de clasificación al clasificar los datos. En este documento, se propone implementar dos nuevos algoritmos CPAR (Clasificación basada en la regla de asociación predictiva) y CMAR (Clasificación basada en reglas de asociación de clases múltiples) que combina las ventajas de

la clasificación asociativa y la clasificación tradicional basada en reglas. En lugar de generar un gran número de reglas candidatas como en la clasificación asociativa, CPAR adopta un algoritmo codicioso para generar reglas directamente a partir de los datos de entrenamiento. Además, CPAR genera y prueba más reglas que los clasificadores tradicionales basados en reglas para evitar perder reglas importantes. Para evitar el ajuste excesivo, CPAR utiliza la precisión esperada para evaluar cada regla y utiliza las mejores reglas k en la predicción. CMAR aplica una estructura de árbol de CR para almacenar y recuperar las reglas de asociación extraídas de manera eficiente, y las reglas de poda de manera efectiva basadas en la confianza, la correlación y la cobertura de la base de datos. La clasificación se realiza en base a un análisis de χ2 ponderado utilizando múltiples reglas de asociación fuertes. Estos extensos experimentos muestran que CMAR es consistente, altamente efectivo en la clasificación de varios tipos de bases de datos y tiene una mejor precisión de clasificación promedio en comparación con FOIL (Aprendizaje Inductivo de Primer Orden) y PRM (Minería de Reglas Predictiva). Los algoritmos propuestos son superiores en términos de requisitos de memoria, complejidad de tiempo y eliminan las estructuras de datos intermedios en la implementación.

No. 15	**Título:** Reglas de poda y afinación para clasificadores asociativos

Autores: Osmar R. Za¨ıane y Maria-Luiza Antonie

Dataset: Sintético con datos generales

Breve descripción: La integración de reglas de clasificación y asociación supervisadas para construir modelos de clasificación no es nueva. Una ventaja importante es que los modelos son legibles y se pueden editar. Sin embargo, es de conocimiento general que la minería de reglas de asociación generalmente produce un gran número de reglas que derrotan el propósito de un modelo legible por humanos. La eliminación de reglas innecesarias sin poner en peligro la precisión de la clasificación es primordial pero muy desafiante. En este trabajo se estudian estrategias para la poda de reglas de clasificación en el caso de clasificadores asociativos.

1.4 Análisis de los resultados

En este epígrafe, se presentan los resultados una vez terminado el plan de ejecución. Por lo tanto, se responderán cada una de las preguntas de la investigación definidas para la revisión de los artículos, para ello se utilizará la notación R_ seguido del identificador de cada pregunta:

R_PI1. Panorama de las investigaciones actuales: Los trabajos seleccionados, descritos en el epígrafe anterior, contienen un identificador con el prefijo **No.**, al examinarlos, se proporciona un informe sobre el panorama actual del análisis de datos en cuanto a los clasificadores asociativos o de reglas de asociación. Por lo tanto, para proporcionar una visión general de los tipos temáticos que han sido propuestos en varios artículos, se categorizaron en 9 clases temáticas. Para una mejor comprensión de estas categorías, y los artículos relacionados, en la tabla 2 que se muestra a continuación, se relacionan cada una de estas clases y los identificadores de los artículos relacionados con dicha clase:

Tabla 2: Clases temáticas y artículos relacionados

Clase temática	Artículos relacionados
Selección de atributos basada en la ganancia de información	• No. 1
Implementación de repetidas podas incrementales para la reducción de la producción (RIPPER)	• No. 2
Análisis de datos clínicos de un paciente	• No. 3
Análisis de Datos en el campo del Estudio del Mercado	• No. 4, No. 11, No. 12
Detección de trastornos cardíacos	• No. 5
Propuesta de un nuevo método y/o algoritmo	• No. 6, No. 8, No. 9, No. 13, No. 14

Presentación de un enfoque basado en la minería para la ruptura de simetría en estructuras algebraicas	• No. 7
Presentación de una vista sistemática sobre reglas de asociación	• No. 10
Estudio para la poda de reglas de clasificación	• No. 15

Fuente: Elaboración propia

Como se puede observar, la clase “Propuesta de un nuevo método y/o algoritmo” es la que abarca mayor cantidad de artículos, seguida de “Análisis de Datos en el campo del Estudio del Mercado”, que son aquellos artículos que se enfocan en la comparación entre dos clasificadores asociativos, dando un resultado final de cuál es el adecuado para una determinada situación. Mientras que el resto de los trabajos solamente se basan en aplicar y/o utilizar un clasificador ya existente para resolver un problema determinado.

R_PI2. Técnicas más utilizadas. En la tabla 3 se muestra un resumen de las técnicas identificadas en los artículos, el mismo se basa en las palabras contenidas en los títulos de los artículos seleccionados, así como en el resultado propio de cada investigación. Por lo tanto, como se puede ver en esta tabla, la clasificación aplicada al resultado de la minería de reglas de asociación son las más utilizadas, teniendo en cuenta que las mismas quedan reflejadas en el desarrollo de cada artículo.

Tabla 3: Técnicas más utilizadas

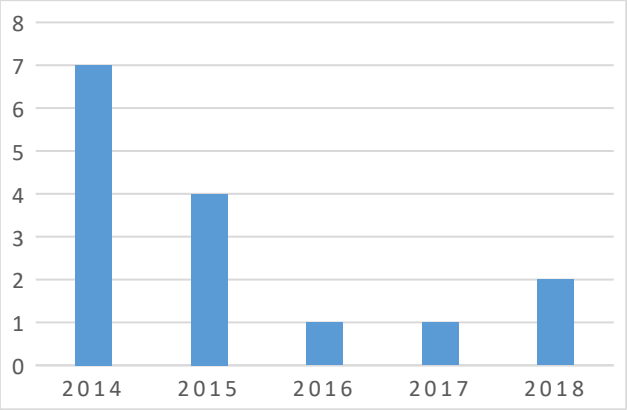
Técnicas identificadas	Artículos relacionados
Regla de asociación ponderada difusa	• No. 1
Algoritmo distribuido de clasificación basado en reglas de múltiples clases (DiRUC)	• No. 2
Aprendizaje inductivo y la teoría de conjuntos aproximados	• No. 3
Máquina de soporte vectorial (SVM), Red Bayesiana Oculta Naive (HNB), Red	• No. 4, No. 5

Bayesiana (BN), Árbol Bayesiano Naive (NBT), Árbol de Decisión (DT), Optimización mínima secuencial (SMO), tabla de decisión (DT), bosque de rotación (RoF) y bosque aleatorio (RF)	
Método de selección de características del filtro basado en correlación rápida	• No. 6
Reglas basadas en patrones recurrentes	• No. 7
Programación genética multiobjetiva (GP)	• No. 8
Ontología	• No. 9
Minería de reglas de asociación y/o minería de reglas clásica	• No. 10, No. 11, No. 12, No. 14, No. 15
Reglas positivas y negativas	• No. 13

Fuente: Elaboración propia

R_PI3. Distribución temporal de los artículos: Al analizar la distribución temporal de los artículos incluidos, se observó que los años 2014 y 2015 reportaron la mayor cantidad de publicaciones, con 7 y 4 artículos respectivamente, que representan el 73,33% de los trabajos revisados. Los años 2016 y 2017 presentaron la misma cantidad de artículos con solo un trabajo en cada año, que representa un 13, 33%, que en su totalidad presentan el mismo número de artículos para el 2018. Esta distribución se muestra en la siguiente tabla.

Tabla 4: Distribución temporal de los artículos



Fuente: Elaboración propia

R_PI4 / R_PI5. Dataset analizados y resultados de las investigaciones: Esta sección presenta las propuestas de solución o los resultados, así como los dataset utilizados por los artículos seleccionados. Es válido destacar que la extracción depende de la calidad y la cantidad de datos de entrada para que los algoritmos produzcan resultados satisfactorios.

La gran mayoría busca patrones para predecir resultados, estrategias, así como al desarrollo de nuevos algoritmos, principalmente, para el apoyo en la toma de decisiones. Otros artículos discuten la efectividad de un clasificador, cualesquiera, generando elementos para determinar qué tan efectiva fue la clasificación de un dataset. Además, se percibe que la mayoría de los artículos utilizan varios dataset, predominando los sintéticos con datos generales, seguido de los reales con datos médicos. Para una mayor comprensión en la tabla 5 se resumen cada uno de los tipos de dataset relacionado con los artículos seleccionados para la presente investigación.

Tabla 5: Tipos de dataset

Tipos de dataset	Artículos relacionados
Sintético con datos generales	• No. 2, No. 4, No. 6, No. 10, No. 12, No. 13, No. 14, No. 15
Reales con datos generales	• No. 8, No. 11
Sintético con datos de seguridad de red	• No. 9

Reales con datos médicos	• No. 1, No. 3, No. 5
Sintético con datos de administración de red	• No. 7

Fuente: Elaboración propia

1.5 Procedimiento para la comparación de los clasificadores asociativos (ClassPro)

Una vez realizado el estudio del arte llevado a cabo a través de la revisión sistemática, las respuestas a las preguntas de investigación y los principales resultados de los estudios se determina como elementos fundamentales dentro del procedimiento de la comparación de clasificadores asociativos, los siguientes:

Extracción de las reglas de asociación de clasificación: En esta etapa se centran los esfuerzos en preparar los datasets para la extracción de reglas de clasificación aplicando un algoritmo de extracción de reglas de asociación. Esta tarea se realiza descartando todas aquellas reglas que tengan más de un elemento a la derecha y que estos no pertenezcan a las clases a asignar.

Implementación de los clasificadores: En esta fase se procede a implementar el clasificador: CBA, HP method y High classify pruning method (HCP), mediante el lenguaje de programación Python.

Comparación de los resultados de los clasificadores utilizando diferentes dataset definidos: En esta fase se procede a realizar un estudio que permita definir las métricas que brinden información acerca de la calidad de la clasificación. En un segundo momento se procede a implementar estas métricas y finalmente se realizan experimentos con los algoritmos sobre los datasets y posteriormente se presentan los resultados. Las métricas seleccionadas fueron: sensibilidad, especificidad, precisión, valor de predicción negativa, error de clasificación, veracidad y prevalencia.

Conclusiones parciales

A partir del estado del arte se determina que la mayoría de las investigaciones se centran en las propuestas de nuevos métodos o algoritmos, se utilizan dataset sintéticos para hacer las validaciones. Las técnicas más utilizadas para el ordenamiento de las reglas utilizan las métricas de soporte, confianza y cardinalidad. Las estrategias más seguidas para establecer las clases usan la mejor clase.

CAPÍTULO 2: IMPLEMENTACIÓN DE LOS CLASIFICADORES ASOCIATIVOS

Introducción

En el presente capítulo se desarrollan los dos primeros pasos definidos en el capítulo anterior como procedimiento para la comparación de clasificadores asociativos. Para la extracción de las reglas se definen los mecanismos de selección de los dataset y el algoritmo de extracción. Para la fase de implementación de los clasificadores se define el lenguaje de programación y los clasificadores a implementar.

2.1 Extracción de las reglas de asociación:

Retomando el primer paso definido en el procedimiento en el capítulo anterior se decide utilizar un dataset para la ejemplificación de todos los elementos desarrollados en este capítulo.

Los dataset seleccionados fueron tomados del repositorio de Machine Learning de la Universidad de California Irvine (UCI, 2019). En la tabla 6 se presenta la descripción de cada uno de los dataset seleccionados. Para la selección solo se tuvieron en cuenta dataset que hayan sido previamente utilizados en tareas de clasificación.

En cada uno de los dataset se eliminaron las filas en las que aparecen atributos nulos y los atributos continuos fueron transformados a discretos siguiendo una aproximación de distribución de frecuencias iguales y 3 categorías. Cada dataset fue separado en 2 clases de equivalencia que contienen la primera el 80 % de las filas y es utilizada para el entrenamiento de los clasificadores y la segunda el 20 % de las filas y es utilizada para la evaluación del clasificador.

Tabla 6: Dataset seleccionados

Nombre	Columnas	Filas	Tipo de dato	Distribución de clases	Cantidad de clases
Breast-Cancer	10	545	Discreto	Balanceada	2
Car	7	1728	Discreto	Desbalanceada	4
Credit	16	653	Discreto	Balanceada	2
Habermant	4	284	Discreto	Desbalanceada	2
Nursery	9	12960	Discreto	Desbalanceada	5
Zoo	17	80	Discreto	Desbalanceada	7

Fuente: Elaboración propia

La generación de reglas de asociación por parte del algoritmo FP-Growth es volátil, por lo tanto, para evitar tener que ejecutar el algoritmo cada vez que se vaya a utilizar un clasificador es necesario contar con una estructura que permita a los algoritmos de clasificación leer de ella para evitar este problema de ejecución repetida del algoritmo.

Se determina realizar una única ejecución y almacenar las reglas resultantes en un archivo de texto. Este fichero se nombra a partir del dataset que lo origine y del soporte que haya sido seleccionado, por ejemplo, si se calcula los itemset frecuentes con un soporte mayor que 0.05 para el dataset brast-cancer, el nombre del fichero con los itemset sería nombrado breast-cancer_0.05.txt.

Cada itemset frecuente dentro del fichero tendrá la siguiente estructura [$ítem_1$, $ítem_2$, …, $ítem_n$] soporte donde ítem 1 ..n hace referencia a cada uno de los ítem individuales que componen el itemset y soporte es un valor real entre 0 y 1 que representa el soporte del itemset. Por ejemplo si tenemos un itemset con los valores 1, 2, 3 y un soporte de 0.5, su representación seria [1, 2, 3] 0.5.

En la generación de las reglas se utiliza como entrada el archivo generado anteriormente, posteriormente se generan reglas de asociación que satisfacen las métricas pasadas por parámetros. Se genera un archivo que tiene como nombre dataset_soporte_rules_conf_confianza.txt por ejemplo, si se pretenden encontrar las reglas de asociación para el dataset brast-cancer con un soporte de 0.05 y una confianza de 0.7, entonces el nombre del fichero seria *breast-cancer_0.05_rules_conf_0.7.txt.* La estructura de cada regla contenida en este fichero seria [antecedentes] -> [consecuente] supp: soporte conf: confianza donde antecedente y consecuente son itemsets mientras que soporte y confianza son valores reales que representan a estas métricas respectivamente. Por ejemplo, una regla en la que 81 implica a 91 con un soporte de 0.63 y una confianza de 0.76 se escribiría, [81]-> [91] supp: 0.63 conf: 0.76.

La extraccion de itemsets frecuentes se realiza con el algoritmo fp-growth implementado en el fichero fp-growth.py

Fragmento del código de *fp_growth.py*:

```
from collections import defaultdict, namedtuple

def find_frequent_itemsets(dataset, min_support,
include_support=False):
 from load_transactions import load_transactions
 transactions = load_transactions(dataset)
 if type(min_support) == float:
 min_support = int(len(transactions) * min_support)
 items = defaultdict(lambda: 0)

 for transaction in transactions:
 for item in transaction:
 items[item] += 1

 items = dict((item, support) for item, support in items.items() if
support >= min_support)
```

Se realiza la modificación para que se descarten aquellas reglas que tienen más de un elemento a la derecha o que no sean las clases predefinidas para este dataset en la fase de entrenamiento.

Fragmento del código:

```
def conditional_tree_from_paths(paths):

 tree = FPTree()
 condition_item = None
```

```
items = set()

for path in paths:
if condition_item is None:
condition_item = path[-1].item
```

Fragmento del código de *extraccion_reglas.py*:

```
from test_fp_growth import fq2file

fq2file('nursery.csv', 34, 'nursery_0.05.txt')
var=([28],[29],[30],[31],[32])

import generate_rules as gr

gr.read_itemsets('nursery_0.05.txt')

gr.write_rules('nursery_0.05_rules_conf_0.7.txt',0.7,'confidence',var)
```

2.3 Implementación de los clasificadores:

Los algoritmos de extracción de reglas de clasificación generan muchas reglas y eso afecta el desempeño de los clasificadores. Las diferencias más importantes entre los clasificadores están asociadas al mecanismo que se sigue para establecer un ranking y podar las reglas que no son esenciales para la clasificación.

2.3.1 Ordenamiento de la regla

El mecanismo de ordenamiento de reglas más utilizado en la biografía asociado a esta temática, combina las métricas de soporte y confianza, favoreciendo aquellas reglas con mayor soporte en el caso de empate entonces se favorecen las reglas con mayor confianza y en el caso de empate se chequea la cardinalidad de la regla. La cardinalidad de la regla está asociada a la cantidad de ítems que están presente en la regla.

La implementación del procedimiento para ordenar las reglas se realizó en el archivo el archivo rank.py el cual, va a darle mayor nivel de importancia a la regla según su soporte, si este no es punto de desempate entonces opta por medir entre sus confianzas, y sino por su cardinalidad, dando así las reglas ya organizadas bajo estos criterios.

Ejemplo del código en el *archivo rank.py*:

```
from turtledemo.chaos import line
from rule import Rule

def scc(rulea):
 return (rulea.support * 10000) + (rulea.confidence *1000) +
rulea.cardinality()

def rank(file):
 arule= []
 with open(file, 'r') as f:
 for line in f:
 rule = Rule.from_line(line)
 arule.append(rule)

 arule.sort(key=scc, reverse=True)
 return arule
```

2.3.2 Implantación de los clasificadores

Una vez obtenidas las reglas se le da paso a la implementación de los clasificadores, en este caso tres, HP method, High classify pruning method (HCP) y Full and partial match rule pruning (CBA).

El método **HP** permite que una regla sea insertada en el clasificador si su cuerpo parcialmente coincide con el caso de entrenamiento sin importar la similitud de clase entre la clase de la regla y aquella del caso de entrenamiento. Por tanto, una vez que las reglas se extraigan y se establezca un ranking, este método itera sobre las reglas empezando con la de más alto ranking y todos los casos de entrenamiento cubiertos por la regla seleccionada se descartan y la regla se inserta en el clasificador. Cualquier regla que no cubra un caso de entrenamiento se elimina. El ciclo termina cuando el conjunto de entrenamiento esté vacío o se hayan probado todas las reglas. La diferencia entre los métodos HP y HCP estriba en que el primero una regla se inserta en el clasificador si cubre parcialmente al menos un caso de entrenamiento, independientemente si lo clasifica correctamente o no. Por otro lado, en el segundo, una regla debe clasificar correctamente un caso de entrenamiento para ser considerada en el clasificador.

```
Input: Given a set of generated rules R, and training data set T
Output: classifier (Cl)
1 R ' = sort(R);
2 For each rule rj in R ' Do
3 Find all applicable training cases in T that partially match rj's
condition
4 Insert the rule at the end of Cl
5 Remove all training cases in T covered by rj
6 If rj cannot correctly cover any training case in T
7 Remove rj from R
8 end if
9 end for
```

Ejemplo del código en el *archivo hp.py*

```
import rank as rk

def read_trainig(training):
 with open('zoo-80.csv', 'r') as tr:
 for line in tr:
 a = line.rstrip('\n')
 a = a.split(',')
 a = [int(x) for x in a]
 size = len(a)
 element = frozenset(a[:size - 1])
 training.append(element)

def partial_coverage(lines, rule, hp):
 covered = []
 for index, b in enumerate(lines):
 for item in rule.premise:
 if frozenset([item]).issubset(b):
 if rule not in hp:
 hp.append(rule)
 covered.append(index)
 lines = [lines[x] for x in range(len(lines)) if x not in covered]

 return lines
```

HCP

Muchas de las reglas encontradas en el paso de entrenamiento no pueden ser usadas para predecir casos de prueba y por lo tanto algunas reglas descubiertas se eliminan. Este método de evaluación High classify pruning method (HCP) (Abumansour, 2010), itera sobre el conjunto completo de reglas después del ranking y aplica cada regla contra el dataset de entrenamiento. Si la regla cubre (coincide parcialmente) un caso de entrenamiento y tiene una clase común con aquella del conjunto de entrenamiento, será insertada en el clasificador y todos los casos de entrenamiento cubiertos por la regla serán eliminados. El método

repite el mismo proceso para cada regla que queda hasta que el conjunto de entrenamiento quede vacío y se consideren las reglas dentro del clasificador dentro del paso de predicción. La diferencia distintiva entre este método y el cubrimiento de la base de datos es que se agrega una regla al clasificador si cubre parcialmente al menos un caso de entrenamiento, sin importar si clasifica el caso correctamente o no. Por otra parte, en el método HCP, una regla debe clasificar un caso de entrenamiento correctamente para ser considerada dentro del clasificador.

Input: Given a set of generated rules R, and training data set T
Output: classifier (Cl)
1 R ′ = sort(R);
2 For each rule r_j in R ′ Do
3 Find all applicable training cases in T that partially match r_j's condition
4 If r_j correctly classifies a training case in T
5 Insert the rule at the end of Cl
6 Remove all training cases in T covered by r_j
7 end if
8 If r_j cannot correctly cover any training case in T
9 Remove r_j from R
10 end if
11 end for

Ejemplo del código en el *archivo hcp.py*

```
import rank as rk
from collections import defaultdict


rules = rk.rank('zoo_0.05_rules_conf_0.7.txt')

def read_trainig():
 training= defaultdict()
 with open('zoo-80.csv', 'r') as tr:
 for line in tr:
 a = line.rstrip('\n')
 a = a.split(',')
 a = [int(x) for x in a]
 size = len(a)
 element = frozenset(a[:size - 1])
 cl = frozenset((a[size - 1],))
 training[element] = cl

 return training


def partial_coverage(lines,rule,hcp):
 covered = []
 for b in lines.keys():
```

```
for item in rule.premise:
if frozenset([item]).issubset(b):
if lines[b] == rule.conclusion:
if rule not in hcp:
hcp.append(rule)
covered.append(b)
return covered
```

CBA

El clasificador CBA se construye teniendo en cuenta las reglas que satisfacen de manera parcial o total los casos en el dataset de entrenamiento. Una regla satisface totalmente un caso en el dataset de entrenamiento si todos los atributos presentes en el cuerpo de la regla están presentes en la instancia del dataset de entrenamiento. Una regla satisface un caso en el dataset de entrenamiento si al menos uno de los atributos presentes en el cuerpo de la regla está presente en el caso de entrenamiento.

Se parte de la regla de más alto nivel y se escanea el dataset de entrenamiento buscando las instancias que complementan parcial o totalmente la regla, estas instancias son eliminadas del dataset y la regla es añadida al clasificador. Si alguna regla complementa al menos alguna instancia del datset de entrenamiento, se elimina la regla. El algoritmo termina cuando no quedan instancias en el dataset de entrenamiento o se han utilizado todas las reglas disponibles.

Implementación de *cba.py*

```
import rank as rk

rules = rk.rank('breast-cancer_0.05_rules_conf_0.7.txt')

fpm = []

def coverage(line, rules):
 a = line.rstrip('\n')
 a = a.split(',')
 a = [int(x) for x in a]
 size = len(a)
 element = frozenset(a[:size-1])
 cl = frozenset((a[size-1],))
 match = []
 for index, rule in enumerate(rules):
 if rule.conclusion == cl:
 if rule.premise.issubset(element):
 match.append(index)

 return match
```

```
def fpm_rules():
 fpm=[]
 with open('breast-cancer', 'r') as tr:
 for line in tr:
 temp = (coverage(line, rules))
 for x in temp:
 if x not in fpm:
 fpm.append(x)
 fpm = [rules[x] for x in fpm]


 return fpm
```

Conclusiones parciales

Se realizó una prueba de eficiencia a la implementación de ClassPro que permite concluir que es un marco de referencia valido para el cálculo de las métricas Sensibilidad, Especificidad, Precisión, Valor de Predicción Negativa, Error de Clasificación, Veracidad y Prevalencia en clasificadores asociativos.

CAPÍTULO 3: COMPARACIÓN DE LOS CLASIFICADORES ASOCIATIVOS

Introducción

En este capítulo se van a desarrollar un grupo de experimentos con vista a determinar cuál es el mejor clasificador asociativo dentro de los seleccionados. Para ello en la sección 3.1 y 3.2 se detallan las métricas utilizadas en la comparación, mientras en la sección 3.3 se muestran los resultados individuales de cada clasificador y se grafican los valores de cada métrica para todos los dataset.

3.1 Matriz de confusión

La matriz de confusión (2018, Koldo Pina) dará una mejor idea de cómo está clasificando el modelo, dando un conteo de los aciertos y errores de cada una de las clases por las que se está clasificando. Así se puede comprobar si el modelo está confundiendo clases, y en qué medida. Una matriz de confusión de dos clases, en este caso + y -, se puede representar de la siguiente forma:

		Clasificador	
		+	-
Valor real	+	TP	FN
	-	FP	TN

Cada columna de la matriz representará el número de predicciones para cada clase realizadas por el modelo, y cada fila los valores reales por cada clase. Con lo cual los conteos quedan divididos en 4 clases, TP, FN, FP y TN, que significan lo siguiente:

TP – Verdadero Positivo: Son el número verdaderos positivos, es decir, de predicciones correctas para la clase +.

FN – Falso Negativo: Son el número de falsos negativos, es decir, la predicción es negativa cuando realmente el valor tendría que ser positivo. A estos casos también se les denomina errores de tipo II.

FP – Falso Positivo: Son el número de falsos positivos, es decir, la predicción es positiva cuando realmente el valor tendría que ser negativo. A estos casos también se les denomina errores de tipo I.

TN – Verdadero Negativo: Son el número de verdaderos negativos, es decir, de predicciones correctas para la clase -.

3.2 Implementación de las métricas

A partir de estas 4 categorías se implementaron una serie de métricas para medir la efectividad del clasificador, tales como Sensibilidad, Especificidad, Precisión**,** Valor de predicción Negativa, Error de clasificación y Prevalencia que brindaran una medida exacta para definir la precisión de cada clasificador con respecto al dataset Breast-Cancer.

- **Sensibilidad:** también se la llama recall o tasa de verdaderos positivos. Da la probabilidad de que, dada una observación realmente positiva, el modelo la clasifique así. Se calcula de la siguiente manera: $\text{Sensibilidad} = \frac{\text{tp}}{(\text{tp} + \text{fn})}$
- **Especificidad:** también llamado ratio de verdaderos negativos. Da la probabilidad de que, dada una observación realmente negativa, el modelo la clasifique así. Se calcula de la siguiente manera: $\text{Especificidad} = \frac{\text{tn}}{(\text{tn} + \text{fp})}$
- **Precisión:** también llamado valor de predicción positiva. Da la probabilidad de que, dada una predicción positiva, la realidad sea positiva también. Se calcula de la siguiente manera: $\text{Presición} = \frac{\text{tp}}{(\text{t} \ + \text{fp})}$
- **Valor de predicción Negativa:** da la probabilidad de que, dada una predicción negativa, la realidad sea también negativa. Se calcula de la siguiente manera: $\text{Valor de predicción negativa} = \frac{\text{tn}}{(\text{tn} + \text{fn})}$

- **Error de clasificación:** Porcentaje de errores del modelo. Se calcula de la siguiente manera:$\text{Valor de predicción negativa} = \frac{(fp+}{(\text{tp} + \text{tn}+\text{fp}+\text{fn})}$
- **Veracidad:** Porcentaje total de los aciertos en el modelo. Se calcula de la siguiente manera: $\text{Veracidad} = \frac{(tp+t}{(\text{tp} + \text{tn}+\text{fp}+\text{fn})}$
- **Prevalencia:** La probabilidad de un positivo en el total de la muestra. Se calcula de la siguiente manera: $\text{Prevalencia} = \frac{(tp+fn)}{(\text{tp} + \text{tn}+\text{fp}+\text{fn})}$

Solamente con la sensibilidad y la especificidad se puede definir qué tan bien clasifica el modelo.

Se determinó implementar por separado las métricas para cada clasificador.

Ejemplo del código de **métricas** para HP en el archivo *metricas_hp.py*

```
from rule import Rule
from hp import hp_rules

tp,fn,fp,tn = (0,0,0,0)
hp_rules = hp_rules()


file='zoo-20.csv'
with open('zoo-20.csv', 'r') as tr:
 for line in tr:
 a = line.rstrip('\n')
 a = a.split(',')
 a = [int(x) for x in a]
 size = len(a)
 element = frozenset(a[:size - 1])
 cl = frozenset((a[size - 1],))
 for var in hp_rules:
 if var.premise.issubset(element):
 if cl == var.conclusion:
 tp+=1
 tn+=1
 else:
 fp+=1
 fn+=1

def sensibilidad(tp, fn):
 a = tp / (tp + fn)
 return a
```

```
def especificidad(tn, fp):
 b = tn / (tn + fp)
 return b


def presicion(tp, fp):
 c = tp / (tp + fp)
 return c


def valor_predc_negativa(tn, fn):
 d = tn / (tn + fn)
 return d


def error_clasificacion(ep, fn, tp, tn):
 e = (ep + fn) / (tp + tn + ep + fn)
 return e


def prevalencia(tp, tn, ep, fn):
 g = (tp + fn) / (tp + tn + ep + fn)
 return g
```

3.3 Resultados

Aplicando cada una de las métricas a cada uno de los dataset seleccionados se obtuvieron por cada uno de los clasificadores los siguientes resultados:

3.3.1 Dataset Breast-Cancer

HCP tn: 113 tp: 113 fn: 25 fp: 25

Sensibilidad: 0.8188405797101449

Especificad: 0.8188405797101449

Precisión: 0.8188405797101449

Valor de predicción Negativa: 0.8188405797101449

Error de clasificación: 0.18115942028985507

Veracidad: 0.8188405797101449

Prevalencia: 0.5

HP tn: 112 tp: 112 fn: 26 fp: 26

Sensibilidad: 0.8115942028985508

Especificidad: 0.8115942028985508

Precisión: 0.8115942028985508

Valor de predicción Negativa: 0.8115942028985508

Error de clasificación: 0.18840579710144928

Veracidad: 0.8115942028985508

Prevalencia: 0.5

CBA tn: 111 tp: 111 fn: 27 fp: 27

Sensibilidad: 0.8043478260869565

Especificad: 0.8043478260869565

Precisión: 0.8043478260869565

Valor de predicción Negativa 0.8043478260869565

Error de clasificación 0.1956521739130435

Veracidad 0.8043478260869565

Prevalencia: 0.5

3.3.2 Dataset Car

HCP tn: 791 tp: 197 fn: 100 fp: 100

Sensibilidad: 0.8877665544332211

Especificad: 0.8877665544332211

Precisión: 0.6632996632996633

Valor de predicción Negativa: 0.8877665544332211

Error de clasificación: 0.16835016835016836

Veracidad: 0.8316498316498316

Prevalencia: 0.25

HP tn: 791 tp: 197 fn: 100 fp: 100

Sensibilidad: 0.8877665544332211

Especificidad: 0.8877665544332211

Precisión: 0.6632996632996633

Valor de predicción Negativa: 0.8877665544332211

Error de clasificación: 0.16835016835016836

Veracidad: 0.8316498316498316

Prevalencia:0.25

CBA tn: 791 tp: 197 fn: 100 fp: 100

Sensibilidad: 0.8877665544332211

Especificidad: 0.8877665544332211

Precisión: 0.6632996632996633

Valor de predicción Negativa: 0.8877665544332211

Error de clasificación: 0.16835016835016836

Veracidad: 0.8316498316498316

Prevalencia:0.25

3.3.3 Dataset Credit

HCP tn: 108 tp: 108 fn: 17 fp: 17

Sensibilidad: 0.864

Especificad: 0.864

Precisión: 0.864

Valor de predicción Negativa: 0.864

Error de clasificación: 0.136

Veracidad: 0.864

Prevalencia: 0.5

HP tn: 108 tp: 108 fn: 17 fp: 17

Sensibilidad: 0.864

Especificad: 0.864

Precisión: 0.864

Valor de predicción Negativa: 0.864

Error de clasificación: 0.136

Veracidad: 0.864

Prevalencia: 0.5

CBA tn: 84 tp: 84 fn: 46 fp: 46

Sensibilidad: 0.6461538461538462

Especificad: 0.6461538461538462

Precisión: 0.6461538461538462

Valor de predicción Negativa 0.6461538461538462

Error de clasificación 0.6461538461538462

Veracidad 0.6461538461538462

Prevalencia:0.5

3.3.4 Dataset Haberman

HCP tn: 53 tp: 53 fn: 19 fp: 19

Sensibilidad: 0.7361111111111112

Especificad: 0.7361111111111112

Precisión: 0.7361111111111112

Valor de predicción Negativa: 0.7361111111111112

Error de clasificación: 0.2638888888888889

Veracidad: 0.7361111111111112

Prevalencia: 0.5

HP tn: 53 tp: 53 fn: 19 fp: 19

Sensibilidad: 0.7361111111111112

Especificad: 0.7361111111111112

Precisión: 0.7361111111111112

Valor de predicción Negativa: 0.7361111111111112

Error de clasificación: 0.2638888888888889

Veracidad: 0.7361111111111112

Prevalencia: 0.5

CBA tn: 56 tp: 56 fn: 22 fp: 22

Sensibilidad: 0.717948717948718

Especificad: 0.717948717948718

Precisión: 0.717948717948718

Valor de predicción Negativa 0.717948717948718

Error de clasificación 0.28205128205128205

Veracidad 0.717948717948718

Prevalencia:0.2

3.3.5 Dataset Nursery

HCP tn: 6903 tp: 1719 fn: 9 fp: 9

Sensibilidad: 0.9986979166666666

Especificad: 0.9986979166666666

Precisión: 0.9947916666666666

Valor de predicción Negativa: 0.9986979166666666

Error de clasificación: 0.0020833333333333333

Veracidad: 0.9979166666666667

Prevalencia: 0.2

HP tn: 3456 tp: 864 fn: 0 fp: 0

Sensibilidad: 1.0

Especificidad: 1.0

Precisión: 1.0

Valor de predicción Negativa: 1.0

Error de clasificación: 0.0

Veracidad: 1.0

Prevalencia:0.2

CBA tn: 10113 tp: 2337 fn: 255 fp: 255

Sensibilidad: 0.9754050925925926

Especificad: 0.9754050925925926

Precisión: 0.9016203703703703

Valor de predicción Negativa 0.9754050925925926

Error de clasificación 0.03935185185185185

Veracidad 0.9606481481481481

Prevalencia:0.2

3.3.5 Dataset Zoo

HCP tn: 60 tp: 10 fn: 0 fp: 0

Sensibilidad: 1.0

Especificad: 1.0

Precisión: 1.0

Valor de predicción Negativa: 1.0

Error de clasificación: 0.0

Veracidad: 1.0

Prevalencia: 0.14285714285714285

HP HCP tn: 60 tp: 10 fn: 0 fp: 0

Sensibilidad: 1.0

Especificad: 1.0

Precisión: 1.0

Valor de predicción Negativa: 1.0

Error de clasificación: 0.0

Veracidad: 1.0

Prevalencia: 0.14285714285714285

CBA tn: 80 tp: 10 fn: 4 fp: 4

Sensibilidad: 0.9523809523809523

Especificad: 0.9523809523809523

Precisión: 0.7142857142857143

Valor de predicción Negativa 0.9523809523809523

Error de clasificación 0.08163265306122448

Veracidad 0.9183673469387755

Prevalencia: 0.14285714285714285

3.4 Gráficas

A continuación, se muestra una representación en forma de tablas de barras los valores de las métricas con respecto a todos los datasets con cada uno de los clasificadores:

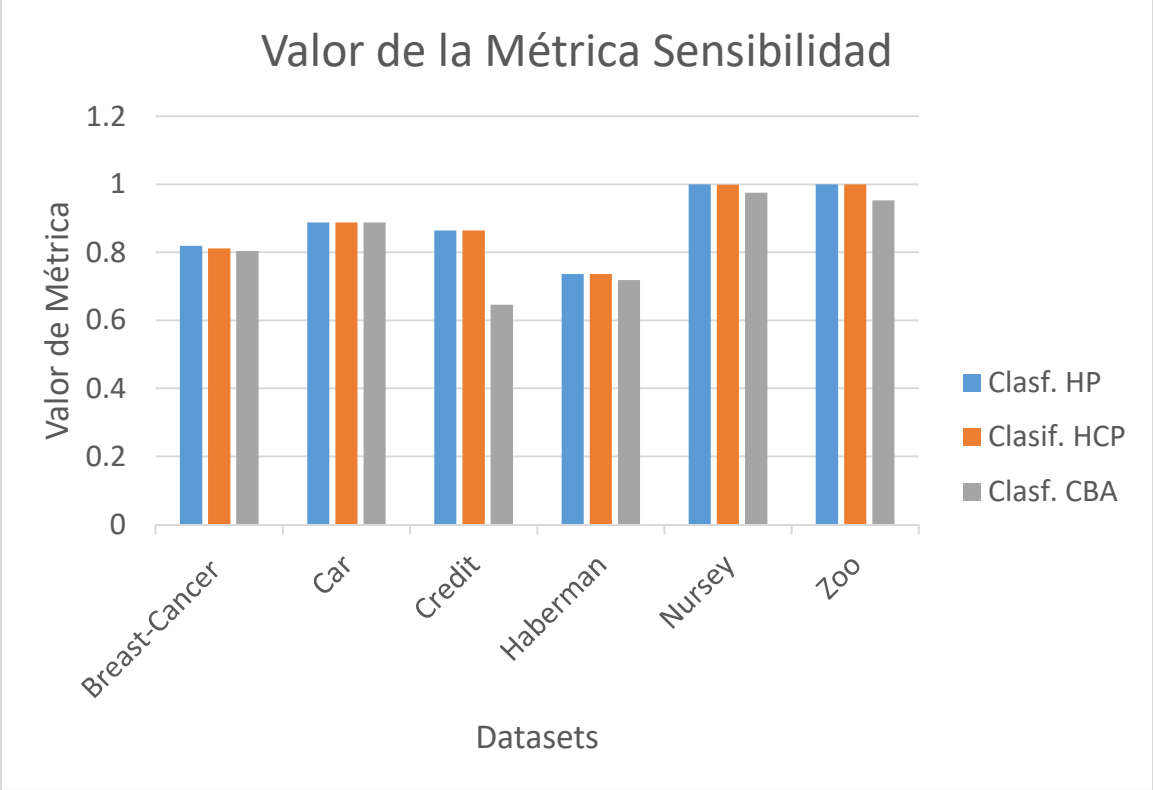


Figura 2: Valor de la Métrica Sensibilidad

Fuente: Elaboración propia

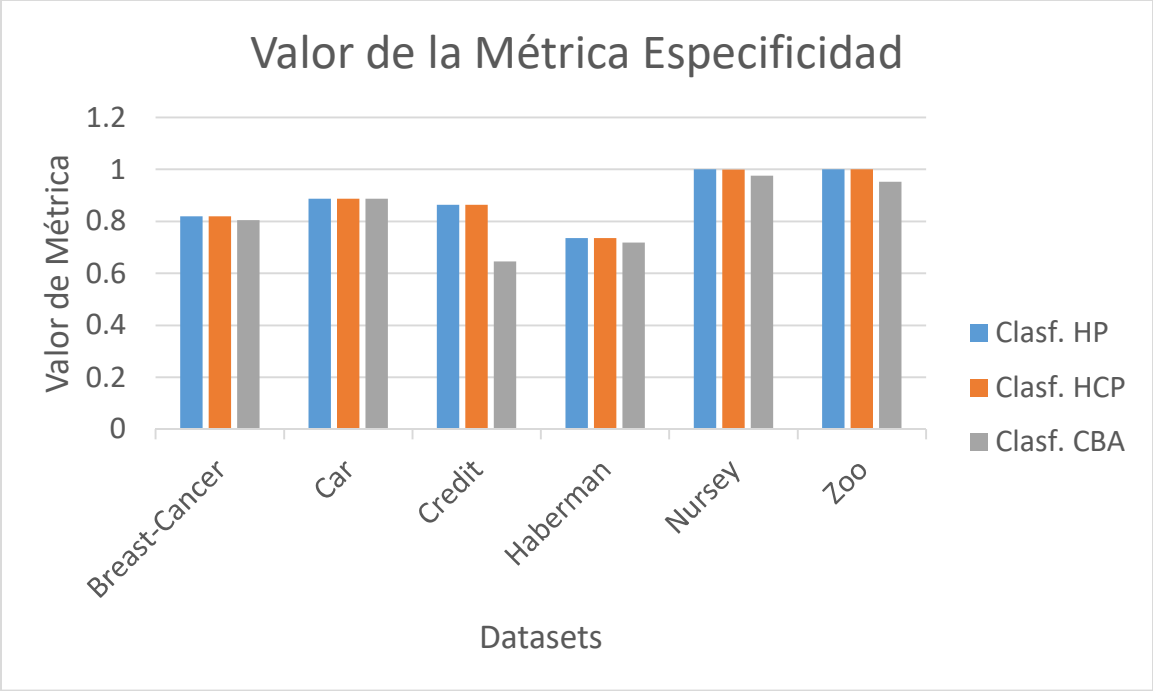


Figura 3: Valor de la métrica Especificidad

Fuente: Elaboración propia

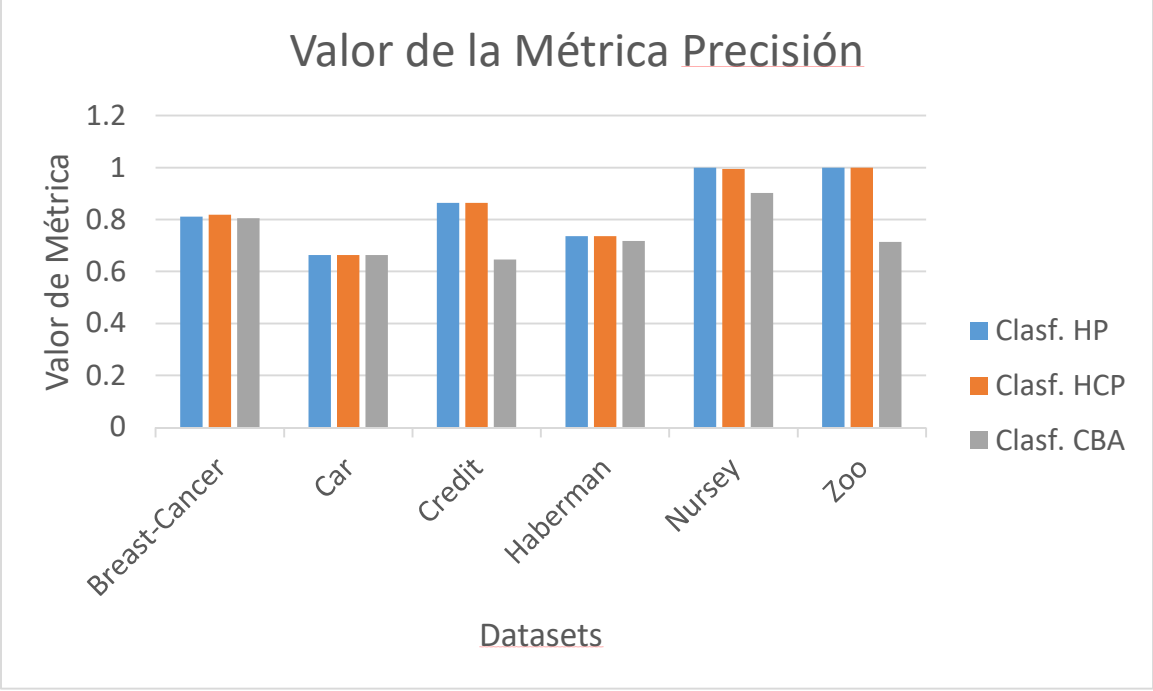


Figura 4: Valor de la Métrica Precisión

Fuente: Elaboración propia

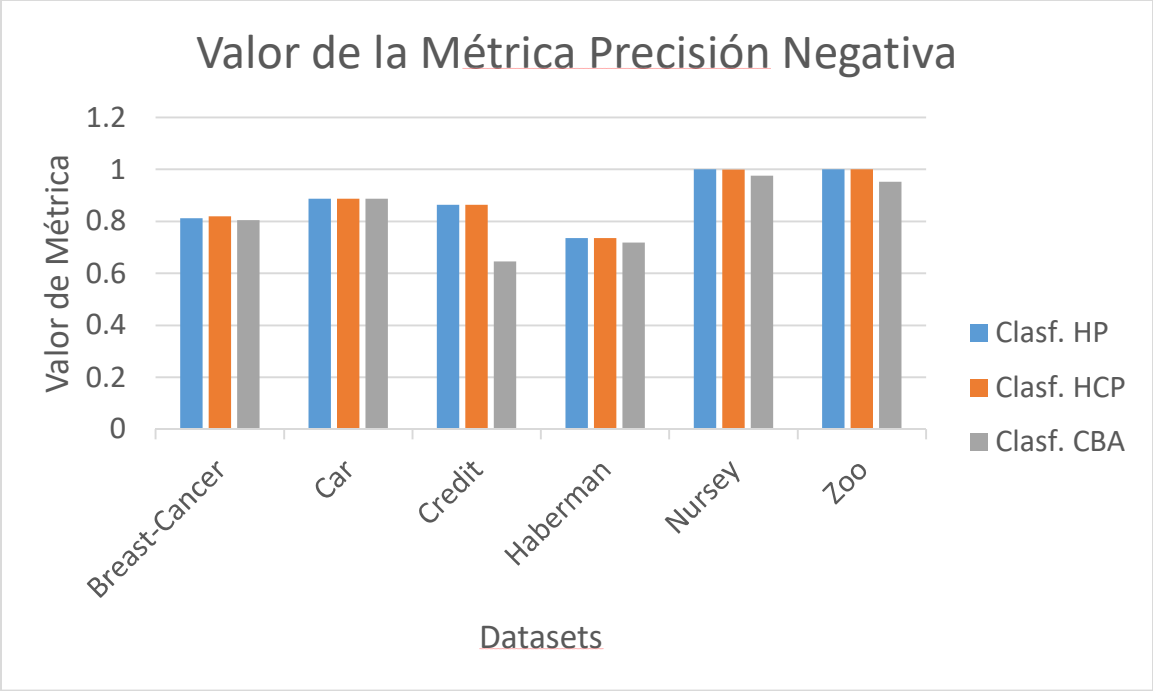


Figura 5: Valor de la Métrica Precisión Negativa

Fuente: Elaboración propia

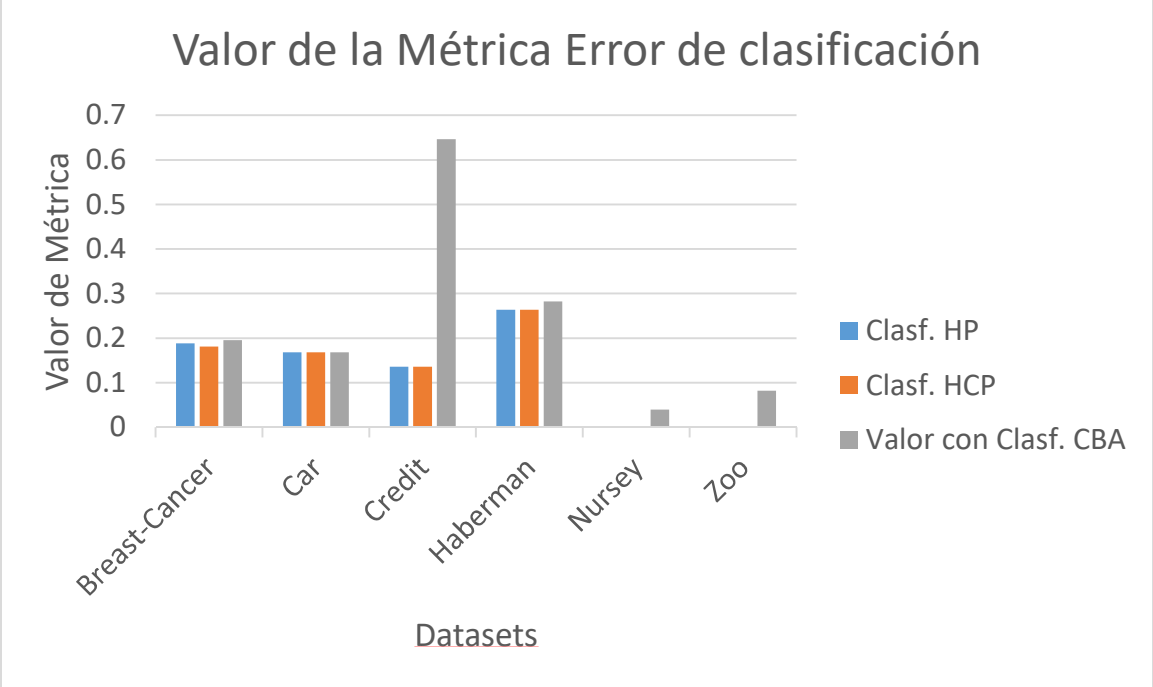


Figura 6: Valor de la Métrica Error de clasificación

Fuente: Elaboración propia

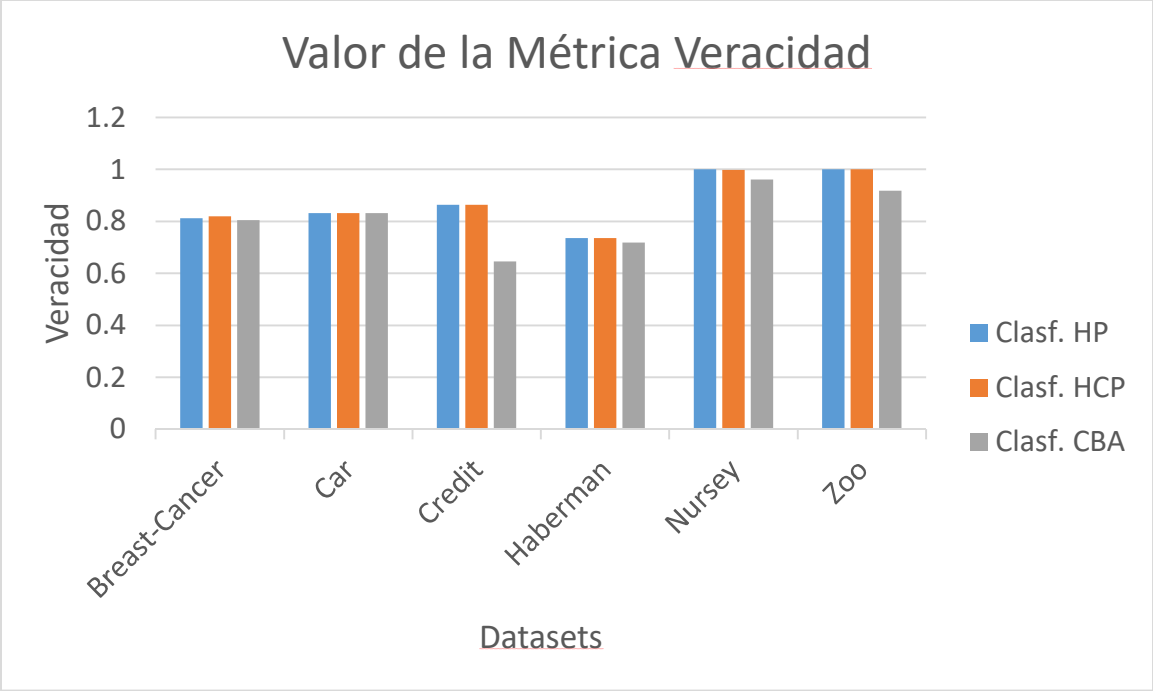


Figura 7: Valor de la Métrica Veracidad

Fuente: Elaboración propia

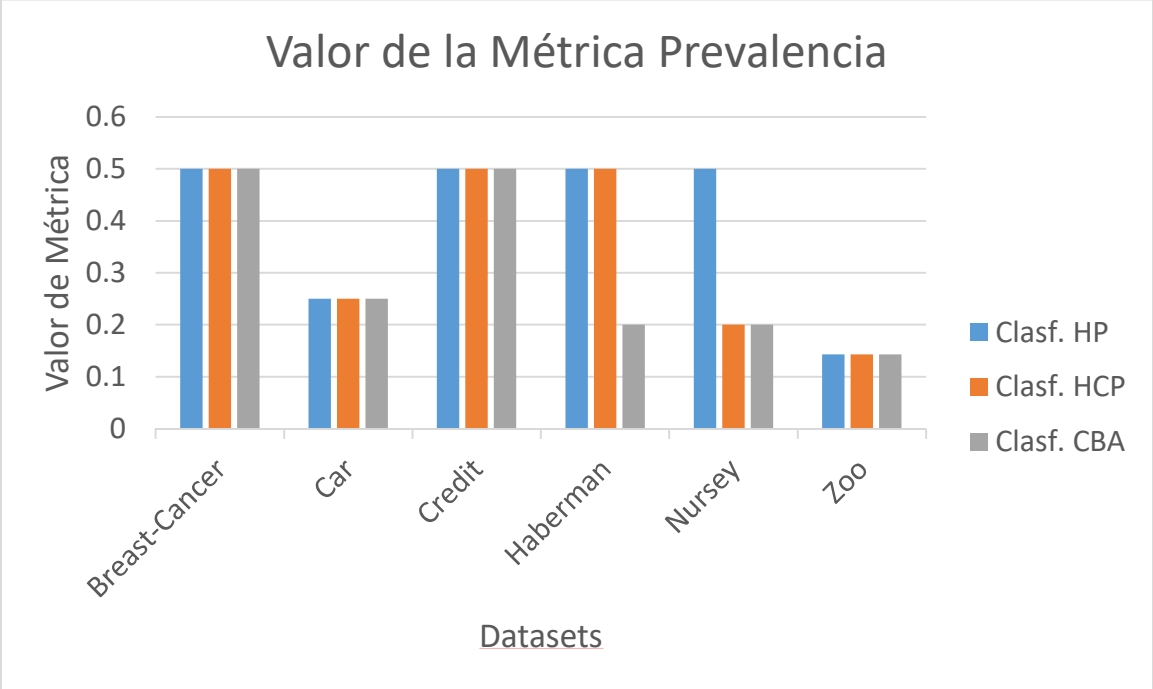


Figura 8: Valor de la Métrica Prevalencia

Fuente: Elaboración propia

Análisis de los resultados

A partir de los resultados alcanzados en los experimentos se puede observar que los clasificadores HP y HCP obtuvieron mejores resultados, para todas las métricas, con respecto al CBA.

Los resultados alcanzados entre el HCP y HP son comparables en todas las métricas. En la métrica Sensibilidad el HP se comportó ligeramente de mejor manera, obteniendo mayor valor dentro del dataset Breast-Cancer, ya que en el resto de dataset tuvieron valores iguales. Dentro de la métrica Especificidad no se logra definir un mejor clasificador entre el HP y HCP ya que obtienen buenos resultados pero iguales. Al analizar la métrica Valor de Predicción Negativa el comportamiento es parejo teniendo entre el clasificador HCP y HP mejor resultado uno del otro en los datasets Breast-C. y Haberman respectivamente, en el resto iguales. Al valorar la métrica Error de Clasificación, el clasificador HCP queda en mejor posición con un menor valor de errores en el dataset Breast-C e igual valores en el resto de los dataset. Con la métrica Veracidad el HCP destaca por muy poco en el dataset Breast-C e iguales valores en el resto, pero donde es determinante el clasificador HP es el en dataset Nursery dentro de la métrica Prevalencia que la diferencia es notable.

Debido a la mayor diferencia alcanzada en la métrica prevalencia se decidió que el mejor clasificador es el HP aun cuando en la mayoría de las métricas no hay diferencias significativas entre HP y HCP.

Conclusiones parciales

Las comparaciones realizadas entre los resultados de las métricas Sensibilidad, Especificidad, Precisión, Valor de Predicción Negativa, Error de Clasificación, Veracidad y Prevalencia a los datasets Breast-Cancer, Car, Credit, Haberman, Nursery y Zoo permiten determinar que el mejor clasificador es HP.

CONCLUSIONES GENERALES

A partir del estado del arte se determina que la mayoría de las investigaciones se centran en las propuestas de nuevos métodos o algoritmos, se utilizan dataset sintéticos para hacer las validaciones. Las técnicas más utilizadas para el ordenamiento de las reglas utilizan las métricas de soporte, confianza y cardinalidad. Las estrategias más seguidas para establecer las clases usan la mejor clase.

Se realizó una prueba de eficiencia a la implementación de ClassPro que permite concluir que es un marco de referencia valido para el cálculo de las métricas Sensibilidad, Especificidad, Precisión, Valor de Predicción Negativa, Error de Clasificación, Veracidad y Prevalencia en clasificadores asociativos.

Las comparaciones realizadas entre los resultados de las métricas Sensibilidad, Especificidad, Precisión, Valor de Predicción Negativa, Error de Clasificación, Veracidad y Prevalencia a los datasets Breast-Cancer, Car, Credit, Haberman, Nursery y Zoo permiten determinar que el mejor clasificador es HP.

RECOMENDACIONES

- Adicionar nuevas estrategias para el ordenamiento de las reglas.
- Adicionar nuevas estrategias para la asignación de clases.
- Comparar con otros clasificadores.

REFERENCIAS BIBLIOGRÁFICAS

- **(ICONTEC), Instituto Colombiano de Normas Técnicas y Certificación. 2011.** *Gestion del riesgo.Principios y directrices.* Colombia : s.n., 2011.
- **Aldape Pérez, Mario. 2017.** *Clasificadores de Patrones: Asociativos y k-NN.* s.l. : AldapeCorp, 2017.
- **Brereton, P., y otros. 2007.** *Lessons from applying the systematic literature review process within the software engineering domain.* s.l. : The Journal of Systems and Software, 2007. págs. 571–583. Vol. 80.
- **CGTI. 2019.** *Utilidad vs. Relevancia de la información.* Guadalajara. México : Red Universitaria de Jalisco, 2019.
- **Fabbr, S., y otros. 2014.** *Managing literature reviews information through visualization.* s.l. : ICEIS, 2014. págs. 36-45.
- **García Floriano, Andrés y Camacho Nieto, Oscar. 2015.** *Clasificador de Heaviside.* México : Universidad De La Salle Bajío, 2015. ISSN: 2007-0705.
- **Gredos. 2018.** www.gredos.com. [En línea] 2018.
- **Itzamá López, Yáñez y Yáñez Márquez, Cornelio. 2014.** *Predicción de la concentración de contaminantes atmosféricos basada en un clasificador asociativo de patrones.* 2014. ISSN 2357-53280.
- **Karimi, S., y otros. 2010.** *Boolean versus ranked querying for biomedical systematic reviews.* s.l. : BMC Medical Informatics and Decision Making, 2010. pág. 58. Vol. 10.
- **Liu, B, W Hsu and Y Ma. 1998.** *Integrating classification and association rule mining.* New York : Conference-KDD, 1998.
- **Menezes, S. L., Freitas, R. S. y Parpinelli, R. S. 2016.** *Mining of massive data bases using hadoop mapreduce and bio-inspired algorithms: A*

systematic review. s.l. : Revista de Informática Teórica y Aplicada, 2016. págs. 69–101. Vol. 23.

- **RAE. 2019.** Real Academia Española. [En línea] 2019. www.rae.com.
- **Research Gate. 2017.** www.researchgate.com. [En línea] 2017.
- **Scielo. 2016.** www.scielo.org.mx. [En línea] 2016.
- **Slimani, Thabets. 2015.** *Class Association Rules Mining based Rough Set Method.* Taif, Saudia Arabia : s.n., 2015.
- **Smith, V., y otros. 2011.** *Methodology in conducting a systematic review of systematic reviews of healthcare interventions.* s.l. : BMC Medical Research Methodology, 2011. pág. 15. Vol. 11.
- **Springer. 2018.** www.springer.com. [En línea] 2018.
- **Turley, Frank. 2016.** *The Prince2 Foundation Training Manual.* 2016.
- **UCI. 2019.** UCI Machine Learning Repository. [En línea] 2019. archive.ics.uci.edu/ml/index.php.