Universidad de las Ciencias Informáticas

Facultad 3

Título: Redes de reglas de asociación

Trabajo de Diploma para optar por el título de
Ingeniero en Ciencias Informáticas

Autor: Alejandro Martínez Valdivia

Tutores: MSc. Julio César Díaz Vera

Ing. Guillermo Manuel Negrín Ortiz

DECLARACIÓN DE AUTORÍA

Declaro ser autor de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

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

Alejandro Martínez Valdivia

_____

Firma del Autor

MSc. Julio César Díaz Vera

_____

Firma del Tutor

Ing. Guillermo Manuel Negrín Ortiz

_____

Firma del Tutor

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 mi familia por su dedicación constante. A mi madre, a mi padre y a mis abuelos. A todos y cada uno de los que hicieron posible que este sueño se hiciera realidad.

Agradecimientos

Resumen

Las reglas de asociación son uno de las tareas de minería de datos más estudiadas y aplicadas en los últimos años. Una regla de asociación pretende encontrar implicaciones entre la ocurrencia de ítems en las transacciones de una base de datos. La cantidad de reglas de asociación que pueden generarse a partir de n ítems es igual a $2^n - 2^{n+1} + 1$ lo cual constituye un problema de alto coste computacional e intratable a la hora de interpretar por los seres humanos. Este trabajo pretende reducir el tamaño de los modelos utilizando una estrategia basada en redes que aproveche el grado de redundancia entre las reglas para reducir la cantidad de reglas que necesita explorar un especialista con fin de encontrar un modelo interpretable. Los resultados de experimentación alcanzados muestran que a partir de un conocimiento relativamente pequeño (alrededor del 1% delas reglas) se alcanzan tazas de reducción superiores de hasta el 90% de las reglas. A partir de estos resultados pueden construirse mecanismos de exploración de reglas de asociación que faciliten la tarea de utilización del conocimiento en la toma de decisiones empresariales.

Palabras claves:

Reglas de asociación, redes de reglas de asociación, redundancia en reglas de asociación.

Índice

Índice de Tablas

Índice de Figuras

Introducción

El análisis y procesamiento de datos es una de las tareas más complejas en la sociedad actual como consecuencia del desarrollo científico-tecnológico. En la sociedad mundial actual para el estudio de cada una de las ramas que componen su estructura se necesitan los grupos de datos sin procesar para garantizar la transformación de estos en información y de esta en conocimiento para la extracción de patrones, posibles tendencias, resultados, comportamiento de variables y relaciones entre los datos. El desarrollo de la computación y de las ciencias asociadas a esta han permitido facilitar el procesamiento y análisis de datos mediante el uso de varias técnicas vinculadas a la inteligencia artificial. Una de las técnicas eficientes para este estudio es la minería de datos.

La minería de datos es la encargada de proveer técnicas para extraer datos relevantes específicos en las bases de datos e incluye varios procesos para garantizar la extracción de información. (Pandey, y otros, 2017)

El proceso de extracción de información analizable se realiza mediante el uso de un grupo de técnicas de extracción, de las cuales constituyen las más utilizadas en las investigaciones referentes a la minería de datos:

Las **redes neuronales artificiales** se basan en el procesamiento automático y su funcionamiento está basado en las neuronas biológicas de los seres humanos. (González-Salcedo, y otros, 2016); la **regresión lineal** es usada para formar relaciones entre los datos, aunque no es eficiente en espacios multidimensionales, es decir, en colecciones de datos de más de dos variables (Roiger, 2017); **árboles de decisión** se emplean en el análisis predictivo para la construcción de reglas lógicas que sirven para categorizar una serie de condiciones sucesivas para dar solución a un problema (Franco Arcega, y otros, 2008)**;** los **modelos estadísticos** es una expresión simbólica en forma de igualdad o ecuación empleada en la regresión para indicar los diferentes factores que influyen en la variable de respuesta. (Sammut, y otros, 2017); el **agrupamiento o clustering** es un procedimiento de agrupación de una serie de vectores según varios criterios habitualmente de distancia. (Rajaraman, y otros, 2011); y las **reglas de asociación**.

La presente investigación se basa en la técnica de reglas de asociación para el estudio de las relaciones implícitas en los diferentes grupos de datos de estudio.

Las reglas de asociación es una de las técnicas más empleadas en la minería de datos. Su objetivo principal es encontrar correlaciones entre los diferentes elementos en una base de datos. Las reglas

condicionales son un aspecto básico para comprender los sistemas basados en reglas ya que son de fácil comprensión para los usuarios.

Las reglas de asociación son ampliamente empleadas para el descubrimiento de las correlaciones en los datos empresariales, en las ciencias médicas como la imagenología para determinar padecimientos o tendencias que pueda padecer un paciente como tumores, células cancerígenas y otros tipos de patologías. Además, se han aplicado técnicas de minado para la detección y prevención de fraudes mediante tarjetas de crédito.

En el marco de los proyectos de la Universidad de las Ciencias Informáticas el desarrollo de este tipo de investigaciones pudiera brindar mejoras para el procesamiento de datos que son almacenados en la universidad. El software encargado de la gestión de los proyectos universitarios (GESPRO). La función de este sistema es procesar, almacenar y clasificar todos los proyectos de la universidad. Esto sería facilitado con el empleo un modelo de reglas de asociación.

Los problemas asociados a los modelos de reglas de asociación están vinculados al número de reglas que son extraídas mediante la minería de datos. Una cantidad de reglas superior a 1000 en un modelo dificulta el estudio por parte de los usuarios y la toma de decisiones basados en las mismas.

En este sentido las investigaciones recientes referentes a la temática han desarrollado soluciones parciales. Por una parte, se han desarrollado estudios basados en las métricas de calidad de soporte y confianza (entre otras) para determinar el conjunto de reglas relevantes en un modelo de reglas de asociación. Las métricas de calidad para evaluar cuán importante o no es una regla en un modelo es un mecanismo eficiente dependiendo del tamaño de la base de datos de estudio y del modelo de reglas extraídos.

Otra línea de investigación está orientada a determinar las reglas más importantes de un modelo de reglas de asociación para determinar cuáles son los ítems más relevantes en el modelo extraído y eliminar las reglas redundantes del modelo. Aunque investigaciones más recientes enfocan la exploración de los modelos con el empleo de redes. Las conocidas como redes de reglas de asociación son estructuras que nos permiten facilitar el estudio de las reglas de asociación determinando cuales son los nodos de mayor significancia en cada red conformada.

El uso de la minería de datos y las redes de reglas de asociación contribuyen a facilitar la extracción del conocimiento y permiten obtener información de utilidad para realizar estudios y apoyar el proceso de toma de decisiones. Con el desarrollo de esta investigación, mediante el uso de redes de reglas de asociación, brindará la aplicación de poder estudiar con mayor facilidad una colección de datos de un

tamaño considerable facilitando el análisis del conocimiento extraído y el apoyo a la toma de decisiones. El propósito de la investigación presente es desarrollar un procedimiento que permita, con el empleo de redes y métricas de calidad, explorar los modelos de reglas de asociación con el objetivo de facilitar el análisis de los modelos de reglas de asociación, generalizando las reglas relevantes y eliminando las redundantes.

A partir de la problemática antes descrita se identifica como **problema a resolver:**

¿Cómo explorar modelos de reglas de asociación mediante el uso de redes?

Enmarcándose en el **objeto de estudio:** Reglas de asociación

Su **campo de acción** se enfoca en la representación de las reglas de asociación mediante el uso de redes.

Para darle solución al problema, se plantea como **objetivo general:** Obtener un procedimiento para explorar un modelo de reglas de asociación mediante redes.

Para dar cumplimiento al objetivo general se definen los **objetivos específicos:**

- Establecer el marco conceptual para el desarrollo de la investigación.
- Definir un procedimiento de construcción de redes de reglas de asociación utilizando el conocimiento como métrica de similitud.
- Validar la propuesta de solución.

Se plantea como **idea a defender** que si se desarrolla un procedimiento para la exploración de redes de reglas de asociación se puede optimizar el proceso de reducción de redundancia de lo extraído mediante el proceso de minería.

Como **posibles resultados** se espera obtener un procedimiento para explorar un modelo de reglas de asociación mediante redes.

Para asegurar el cumplimiento del objetivo general se definen como **tareas a cumplir:**

- Recopilar la bibliografía referente al tema, selección y análisis de la bibliografía.
- Estudiar las soluciones previas al modelo de reglas de asociación.
- Seleccionar los principales elementos asociados al estudio de una base de datos empleando un modelo de redes de reglas de asociación.
- Identificar los principales requisitos para la implementación del algoritmo.

- Implementar un algoritmo que permita obtener la representación de una red de reglas de asociación de fácil entendimiento al usuario.
- Aplicar pruebas al algoritmo implementado.
- Presentar los resultados.

Descripción de los capítulos

A continuación, se presenta la descripción de los capítulos asociados a la investigación presente:

Capítulo 1: “Fundamentación teórica”. Describe el método de revisión sistemática de la bibliografía para la realización de un estudio de las soluciones existentes y las investigaciones referidas al tema de la presente investigación. Posteriormente se selecciona un procedimiento para orientar el desarrollo de la investigación

Capítulo 2: “Desarrollo de la propuesta”. Se presenta el procedimiento para la exploración de modelos de reglas de asociación. Descripción de las herramientas, clases, ficheros y métodos necesarios para garantizar la viabilidad del resultado.

Capítulo 3: “Validación de la propuesta”. Se analizan los resultados de la implementación diseñada con el tratamiento de varios Dataset para verificar la calidad y el funcionamiento en cada una de las exploraciones realizadas.

Capítulo 1 Fundamentación Teórica

Introducción

Describe el método de revisión sistémica para la realización del estudio de soluciones homólogas al problema planteado y análisis bibliográfico para identificar las principales temáticas de investigaciones y las soluciones obtenidas en cada una de las mismas.

1.1. Método de revisión sistemática.

Se emplea el método de revisión sistemática para la construcción del estado de arte. Este permite realizar una evaluación rigorosa y confiable de la bibliografía, además de realizar síntesis de la información seleccionada.

Esta metodología ha sido ampliamente empleada. Por ello la investigación se basa en la recomendación de trabajos previos, que es dividir el proceso en tres fases: planeación, conducción y análisis de resultados. Las siguientes secciones destacan cada uno de estos escenarios.

1.1.1 Planeación de la investigación

La planeación está orientada a definir las preguntas de la investigación, identificación de bases de datos, definición de las palabras clave, estrategias de búsqueda, criterios de inclusión y exclusión y calidad de los artículos. Fueron definidas como preguntas de investigación las siguientes:

PI1- ¿Cuál es el enfoque de los investigadores?

PI2- ¿Cuáles son los Dataset empleados en las investigaciones?

PI3- ¿Cuáles son los algoritmos aplicados?

PI4- ¿Cuáles son los resultados obtenidos?

Usualmente los criterios de inclusión y exclusión son determinados después de conformar las preguntas de la investigación. A continuación, se describen los criterios de selección de los artículos.

1.1.1.1 Criterios de calidad

Están orientados a aplicar filtros de información a la bibliografía recopilada para obtener los artículos que estén realmente vinculados a la problemática de la investigación.

Criterios de inclusión.

Están orientados a obtener los artículos que cumplan las pautas definidas a continuación.

- Estudios realizados en idioma inglés.
- Documentos que constituyan artículos de revistas o artículos de conferencias.

Criterios de exclusión.

Están orientados a obtener los artículos que cumplan las pautas definidas a continuación.

- Estudios fuera del contexto de trabajo.
- Estudios redactados en otro idioma que no sea inglés.
- Estudios publicados antes del 2014.

La definición de los criterios de inclusión y exclusión permiten emitir un criterio de selección para cada uno de los artículos asegurando la calidad de los mismos y el desarrollo de la investigación evaluando el título, las palabras clave, los resúmenes, los mecanismos de solución y las conclusiones en cada uno de los mismos. Para la realización de la selección de los artículos que cumplan con los criterios especificados se seleccionaron las bases electrónicas siguientes:

- ACM Digital Library: dl.acm.org
- Springer: www.springer.com
- IEEE Xplorer Digital Library: ieeexplore.ieee.org
- Semantic Scholar: www.semantic.scholar.org

Los motores o llaves de búsqueda son mecanismos para la recopilación de información referente a un tema específico con el uso de caracteres o palabras relacionadas al objetivo de la búsqueda. En la investigación se selecciona como llave de búsqueda la cadena (“association rules” or (network or graph)) para encontrar los artículos vinculados al objetivo de la investigación.

1.1.2 Plan de ejecución

El plan de ejecución se desarrolla para, una vez definidos los criterios del análisis bibliográfico, orientar el estudio de los artículos que sean de utilidad para guiar el curso de la investigación. El mismo consta de 5 etapas:

- Realización de la búsqueda en las bases de datos
- Comparación de resultados para excluir artículos de igual solución
- Aplicación de los criterios de inclusión, exclusión y calidad
- Evaluación de todos los estudios que aprobaron la revisión inicial
- Síntesis de datos.

Realizado el plan de ejecución se obtuvieron 1402 trabajos. Al aplicar el criterio de exclusión referente a los trabajos que fueron publicados antes del 2014 se redujo el número de trabajos en 370. Estos trabajos fueron sometidos al análisis de título, palabras clave y resumen, obteniéndose un total de 1019 trabajos rechazados. Finalizado el plan de ejecución se obtuvo como resultado un total de 13 artículos, los cuales fueron analizados a partir de las preguntas de investigación.

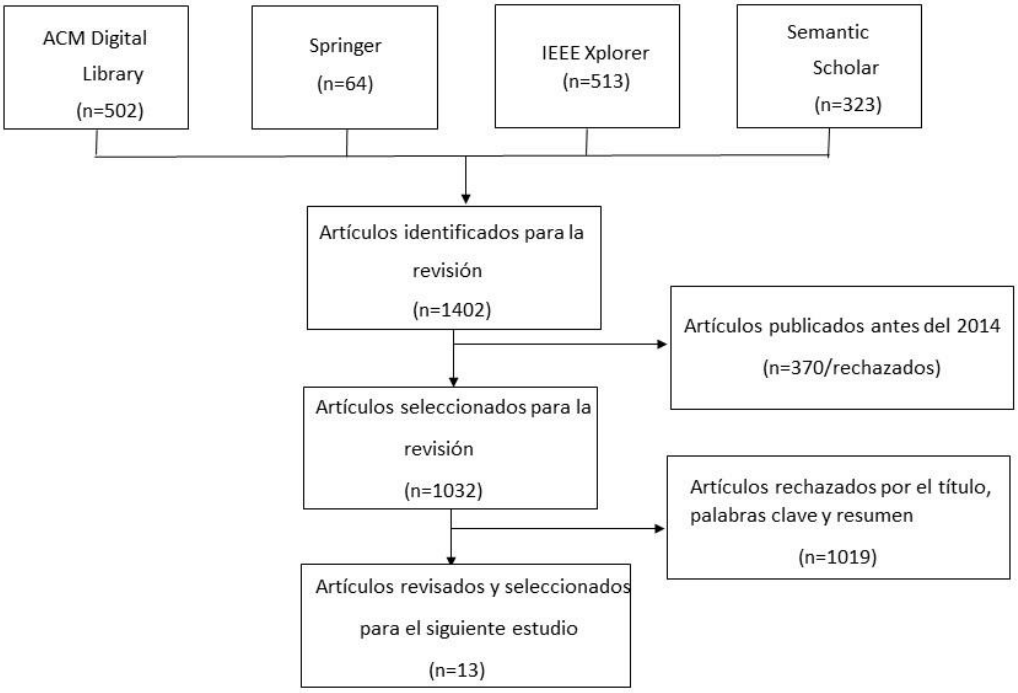

Figura 1. Flujo de la investigación

1.1.3 Resultados

Siguiendo el orden planteado en la fase de planeación de la investigación se darán respuestas a cada una de las preguntas de investigación.

¿Cuál es el enfoque de los investigadores?

Los trabajos seleccionados teniendo en cuenta el enfoque de los investigadores se encuentran listados en la Tabla 1. Haciendo un análisis de cada uno se adquiere un panorama actualizado de los enfoques de análisis de modelos de reglas de asociación. Los enfoques en cada una de las investigaciones pueden ser clasificados como los campos en los cuales serán aplicados los resultados de los artículos. Por ello se categorizaron los trabajos en siete clases: ciencia básica, análisis de conceptos formales, exploración de relaciones musicales, análisis de descomposición del abono, monitoreo de riesgos acoplados, toma de decisiones y clasificación de texto. (ver Tabla 2). En la Tabla se observa cómo los trabajos están orientados a desarrollar soluciones que sean aplicadas a varias áreas con la utilización de técnicas que propongan elementos de mejora en análisis de post-procesamiento, la clasificación de textos, extracción y exploración. En la mayoría de los artículos estudiados la extracción de información

es uno de los pasos más difíciles ya tiene asociado un gran costo y dificultad computacional. A continuación, se detallan cada una de las temáticas:

- **Ciencia básica:** orientada al desarrollo de soluciones para las redes de reglas de asociación que puedan ser aplicadas a cualquier área de trabajo, con el propósito de desarrollar soluciones adaptables al ambiente del problema.
- **Análisis de conceptos formales:** se trata de encontrar los conjuntos más frecuentes en un grupo de datos mediante el uso de algoritmos basados en los conceptos de enrejado y análisis de conceptos formales.
- **Exploración de relaciones musicales:** aborda un mecanismo para la recuperación de información musical con la construcción de redes de reglas de asociación basadas en patrones temporales.
- **Análisis de descomposición del abono:** se emplea para el análisis de parámetros del abono para la demostración de la utilidad de las técnicas computacionales para incrementar la productividad.
- **Monitoreo de acoplamiento de riesgos:** expone el empleo de las redes complejas para el estudio de los potenciales riesgos y de las relaciones acopladas con el uso de las reglas de asociación.
- **Toma de decisiones:** asociada al apoyo de la toma de decisiones en una empresa mediante el uso de un modelo bayesiano de características categorizadas y ponderadas con la incorporación de redes de reglas de asociación.
- **Clasificación de textos:** mejorar los resultados que se obtienen en la clasificación de textos con los algoritmos de clasificación transductiva utilizando redes heterogéneas bipartitas.

A continuación, se muestra el listado de artículos seleccionados:

Tabla 1: Artículos de reglas de asociación

ID	Artículos
1	Post-procesamiento de reglas de asociación usando redes y aprendizaje transductor
2	Post-procesamiento de reglas de asociación: aprovechamiento de una red basada en la propagación de etiquetas
3	Aprendizaje semi-supervisado para dar soporte a la exploración de reglas de asociación
4	Algoritmo para la propagación de etiquetas libre de parámetros usando redes heterogéneas bipartitas para la clasificación de texto

5	Optimización y propagación de etiquetas en redes heterogéneas bipartitas para mejorar la clasificación de los textos
6	Aprovechamiento de la construcción de grafos de conflicto basados en reglas de asociación para las redes ultra densas
7	Extracción de reglas de asociación no redundantes de conceptos de matrículas basado en un sistema IsoFCA
8	Exploración de los datos usando extensiones de reglas de asociación
9	Exploración relaciones musicales usando redes de reglas de asociación
10	Análisis de descomposición de parámetros de abono verde en el noreste brasileño con reglas de asociación
11	Medidas objetivas asimétricas aplicadas al filtrado de reglas de asociación
12	Minado de reglas de asociación combinado con redes complejas para monitorear riesgos acoplados
13	Incorporación de las redes de reglas en un modelo de Bayes ingenuo ponderado por categoría para soportar la toma de decisiones

La Tabla 2 especifica las relaciones entre las temáticas y los artículos seleccionados:

Tabla 2. Temáticas de investigación

temática	**Artículo**
ciencia básica	[1,2,3,6,8,11]
análisis de conceptos formales	[7]
exploración de relaciones musicales	[9]
análisis de descomposición del abono	[10]
monitoreo de acoplamiento de riesgos	[12]
toma de decisiones	[13]
clasificación de texto	[4,5]

Analizados los artículos teniendo en cuenta los enfoques de los investigadores La ciencia básica posee el mayor número de trabajos con un total de seis investigaciones asociadas, le sigue la clasificación de textos con dos trabajos, mientras que el resto presenta un trabajo por cada temática.

¿Cuáles son los Dataset empleados en las investigaciones?

La extracción y análisis de la información de los Dataset empleados depende de los valores cualitativos y cuantitativos de los datos analizados en cada uno de los algoritmos. La limitación más importante es encontrar una fuente segura y seleccionar atributos relevantes para solucionar el problema.

La Tabla 3 muestra los Dataset asociados a cada uno de los artículos analizados.

Tabla 3. Dataset

Artículo	Dataset
1	Clima nominal (clima), lente de contacto (medicina), grupo de datos de un supermercado de Sao Carlos (estudio de mercado)
2	No tiene Dataset
3	No tiene Dataset
4	Conjunto de documentos de texto
5	E-mail, documentos médicos, artículos nuevos, documentos científicos (colección de textos)
6	No tiene Dataset
7	No tiene Dataset
8	No tiene Dataset
9	Colección de pistas de música divididas en 10 géneros
10	Grafo asociado a los cultivos y el tiempo de vida promedio
11	Lentes de contacto (medicina), granos de soya (cultivo), estado del abono verde (plantación)
12	No tiene Dataset
13	Colección de los registros de los pacientes en un hospital en Taiwán

¿Cuáles son los algoritmos aplicados?

Analizados los artículos se seleccionaron de cada uno los algoritmos más relevantes, los cuales se exponen a continuación:

Algoritmo para el aprovechamiento del post-procesamiento transductor: elabora una nueva forma para extraer y emplear el conocimiento del usuario seleccionando las reglas más importantes en una red y propagando esta información hacia las reglas no clasificadas en la red con el uso del aprendizaje transductor para la clasificación de los elementos.

Algoritmo de redes de propagación basada en etiquetas: incorpora el concepto de red de propagación basada en etiquetas para realizar la exploración de las reglas minadas en el proceso de extracción.

Algoritmo de aprendizaje semi-supervisado para la exploración: utiliza la ponderación para la clasificación de las reglas de más significación para posteriormente propagar esta clasificación hacia los elementos sin clasificar.

Algoritmo para la propagación de etiquetas en redes heterogéneas bipartitas: clasifica las reglas sin explorar mediante la propagación de las clases ponderadas más relevantes mediante el uso de redes heterogéneas bipartitas con la ventaja de poder clasificar textos etiquetados, no etiquetados y términos.

Algoritmo de categorización transductiva basada en redes bipartitas heterogéneas: asegura la asignación de puntuaciones a los términos de relevancia mediante términos etiquetados, asignación de información de clases a documentos no etiquetados e induce la relevancia de términos usando puntajes mediante información asignada a los documentos sin etiquetar.

Algoritmo CGC: la construcción del grafo de conflicto se utiliza para la representación del modelo de reglas de asociación para el tratamiento de reglas ultra densas es el principio de funcionamiento del algoritmo.

Algoritmo ENAR: es capaz de generar reglas de asociación no redundantes establecidas directamente desde el concepto de enrejado y el número de reglas es mucho menor y más intuitivo.

Algoritmo ARN extendido: utiliza una red de reglas de asociación extendida con de principio de determinar las clases de más relevancia para propagar la clasificación hacia los elementos sin clasificar.

Algoritmo ExARN: es un algoritmo recursivo encargado de reunir las reglas directa o indirectamente relacionadas con el objetivo aplicando un conjunto de restricciones de redes de reglas de asociación de extracción.

Algoritmo B-Graph: después del proceso de extracción de reglas se encarga de transformar una red de reglas de asociación mediante un hipergrafo de propagación hacia atrás.

Algoritmo ARN filtrado: utiliza métricas asimétricas para explorar los modelos de reglas de asociación.

Algoritmo de clasificación para redes complejas: primero utiliza el A priori para el minado de reglas de asociación para posteriormente utilizar las redes complejas para su análisis.

Algoritmo de reglas categorizadas y ponderadas con clasificador bayesiano ingenuo: utiliza un clasificador bayesiano ingenuo para la clasificación de reglas de asociación.

La tabla 4 muestra los algoritmos asociados a cada artículo.

Tabla 4. Algoritmo

Articulo	Algoritmo
1	Algoritmo para el aprovechamiento del post-procesamiento transductor
2	Algoritmo de redes de propagación basada en etiquetas
3	Algoritmo de aprendizaje semi-supervisado para la exploración
4	Algoritmo propuesto para la propagación de etiquetas en redes heterogéneas bipartitas
5	Algoritmo de categorización transductiva basada en redes bipartitas heterogéneas
6	Algoritmo CGC
7	Algoritmo ENAR
8	Algoritmo ARN extendido
9	Algoritmo ExARN
10	Algoritmo B-Graph
11	Algoritmo ARN filtrado
12	Algoritmo de clasificación para redes complejas

13	Algoritmo de reglas categorizadas y ponderadas con clasificador bayesiano ingenuo

¿Cuáles son los resultados obtenidos?

La tabla 7 se reflejan los resultados obtenidos en cada una de las investigaciones planteadas.

Tabla 5. Resultados de las investigaciones

Artículo	Resultados
1	Se presentaron mejoras en la reducción del tiempo de exploración con el empleo del algoritmo con respecto al aprovechamiento de las medidas objetivas
2	Se logró mejorar el tiempo de procesamiento en la exploración del modelo de reglas de asociación usando la red de propagación de etiquetas
3	La ponderación no redujo el espacio del exploración de las reglas pero si redujo de manera mínima el tiempo de respuesta del algoritmo Apriori
4	El uso de redes bipartitas heterogéneas mejora el tiempo de clasificación de los objetos de las redes homogéneas de documentos
5	Se logró una mejora considerable en la clasificación de la versión inductiva del algoritmo, incluso destaca un mejor funcionamiento en documentos sin etiquetar en el proceso de clasificación.
6	El grafo de conflicto mejora el tiempo de exploración del algoritmo Apriori
7	Los algoritmos empleados lograron reducir de manera considerable el número de reglas a explorar que en las empleadas en el tradicional Apriori
8	Se mejora el espacio de exploración porque se seleccionan los elementos más significativos del modelo
9	Se lograron encontrar patrones significativos con el uso del algoritmo en los Dataset de estudio con nuevas restricciones como un umbral de soporte bajo en las reglas definidas como no redundantes
10	Se obtuvieron elementos asociados a la conformación del hipergrafo que reducen el tiempo de exploración de las reglas en forma de redes de reglas de asociación
11	Se logra una mejora en la clasificación de reglas con la utilización de las métricas

12	La investigación proporciona conocimiento a los gerentes el conocimiento necesario para monitorear y predecir el acoplamiento de riesgos mejorando redes complejas con reglas de asociación
13	Reduce el espacio de exploración al utilizar el clasificador bayesiano ingenuo

Después de analizar los enfoques, algoritmos, Dataset y resultados de cada uno de los artículos se define que la mayoría de las soluciones están orientadas a ciencia básica; que se necesitan un conjunto de Dataset de diferente procedencia para garantizar una mejor experimentación y que los algoritmos empleados en cada una de las investigaciones pueden tributar al desarrollo de esta investigación en la creación de un modelo de reglas de asociación

Luego del análisis de los principales resultados asociados a los trabajos estudiados se necesita garantizar un proceso de conformación del modelo de reglas de asociación; además se necesita una red para almacenar este modelo y auxiliar el proceso de análisis del mismo mediante la aplicación de métricas de calidad. Por ello se decide realizar la implementación de un procedimiento para la exploración de modelos de reglas de asociación mediante redes.

1.2 Procedimiento para la exploración de modelos de reglas de asociación mediante redes

Para dar solución al problema de la investigación se define un procedimiento que guie el proceso de implementación y garantizar la calidad en los resultados de la investigación.

El procedimiento se basa en la realización de 3 etapas:

- **Extracción de las reglas de asociación:** conformación del modelo de reglas de asociación.
- **Establecimiento de la red:** representación a partir de una red del modelo de reglas de asociación.
- **Análisis de la red conformada:** análisis de la red conformada para determinar los nodos representativos del conocimiento en la estructura planteada.

1.2.1 Extracción de las reglas de asociación

En esta fase se definen las métricas para examinar y extraer las reglas en una base de datos. Para su selección y definición es necesario definir restricciones y métricas que identifiquen reglas significativas. Asociado a la extracción de reglas de asociación es importante tener en cuenta las siguientes definiciones:

Transacción: sea *T* una colección de elementos que un cliente toma en realizar una transacción e *i (k=1, 2, …n)* los elementos que componen *T* entonces el conjunto de todos los elementos en *D* son *I= {i1, i2, …ik}.* (Zhou, y otros, 2019)

Reglas de asociación: son presentadas como una implicación $A \rightarrow B$ donde A es el antecedente de la regla y B es el consecuente. Semánticamente hablando son reglas condicionales y su significado es si A aparece en una transacción en una base de datos entonces B también debe aparecer.

Métricas de calidad:

Fundamental en la etapa de extracción de las reglas de asociación son las métricas para evaluar la relevancia de cada regla extraída. Una de ellas es *soporte* denotado $supp(A \rightarrow B)$ que representa la porción de las transacciones en una base de datos donde $A \cap B$ sea verdadero. Teniendo como fórmula:

$$(A \Rightarrow B) = \frac{|\{T: A \cup B \subseteq T, T \subseteq D\}|}{|D|}$$

Otra de las métricas es la *confianza,* definida como la medida para una porción en la base de datos en las que las transacciones que contienen Y también contienen a X. (Bhargava, 2016). Teniendo como fórmula:

$$(A \Rightarrow B) = \frac{|\{T: A \cup B \subseteq T, T \subseteq D\}|}{|\{T: A \subseteq T, T \subseteq D\}|} = \frac{Support\,(A \Rightarrow B)}{Support\,(A)}$$

El *soporte* y la *confianza* son las métricas, según la bibliografía, más utilizadas para evaluar el minado de reglas de asociación. Si bien establecer umbrales de *soporte* y *confianza* es un componente importante para evaluar la relevancia de la regla; esto no garantiza que aquellas que excedan los umbrales sean las más significativas. El marco de evaluación basado en el *soporte* y la *confianza* tiene serias limitaciones como por ejemplo no es capaz de identificar las dependencias negativas, no es capaz de determinar la independencia estadística ni incluye al consecuente dentro del cálculo de la métrica. Se da el caso que ítems con un alto soporte son predichos por casi cualquier combinación de elementos. (Yuan, 2017)

Con vistas a resolver este tipo de problemas se han propuesto otros marcos de evaluación a partir de otro tipo de métricas como son convicción, interés, y el factor de certeza.

En el caso de la convicción se define de la siguiente manera:

$$Conv(A \rightarrow B) = \frac{Supp(A) * Supp(\neg B)}{Supp(A \cup \neg B)}$$

(Zhang, y otros, 2011)

En el caso del interés se define:

$$Int(A \rightarrow B) = \frac{Supp(A \rightarrow B)}{Supp(A) * Supp(B)}$$

(Hahsler, y otros, 2016)

El factor de certeza resuelve varios de los problemas presentes en el marco de la confianza y el soporte y permite encontrar reglas interesantes. Se define de la siguiente manera:

$$CF(A \rightarrow B) = \frac{Conf(A \rightarrow B) - Supp(B)}{1 - Supp(B)}$$

si $Conf(A \rightarrow C) > Supp(C)$ y

$$CF(A \rightarrow B) = \frac{Conf(A \rightarrow B) - Supp(B)}{Supp(B)}$$

(Berzal, et al.)

Para esta investigación se emplearon las métricas de soporte, confianza y factor de certeza.

Técnica empleada para la extracción del modelo de reglas de asociación

El algoritmo FP Growth (Lin, y otros, 2015) es uno de los algoritmos más empleados para la extracción de reglas de asociación. El mismo se divide en 2 pasos fundamentales:

1-Encontrar los grupos de elementos frecuentes en una base de datos de estudio asignando un umbral de soporte para cada uno de los elementos extraídos.

2-Almacenar los grupos de itemsets frecuentes expresados en reglas de asociación en estructuras FPTree y FPNode para posteriormente evaluar la importancia de cada una de las reglas seleccionando los nodos críticos en la red conformada

1.2.2 Establecimiento de la red

Una red puede ser definida como una matriz donde n es el total de nodos presentes en una red. Si existe una arista entre los nodos *Vi* y *Vj*, entonces *Aij=1*, de otra forma *Aij=0*. Las redes sin ciclos son

denotadas como *Aii=0*. En las redes ponderadas las aristas pueden ser asociadas con pesos que definan cuan fuertes, intensas o capaces son las reglas. (Zhou et al. 2019).

Se definen los siguientes pasos para la creación de la red:

- **Creación de los nodos:** cada nodo es una regla de asociación con un soporte y una confianza asociados.
- **Creación de las aristas:** cada arista es creada entre dos reglas de asociación. Esta representación permite establecer una secuencia para estructurar los caminos que son formados por las reglas extraídas, lo cual es de mucha utilidad para definir los nodos de mayor importancia.

Tipos de redes.

Se definen dos tipos de redes para el almacenamiento de las reglas extraídas con el propósito de evaluar los resultados en dependencia del tipo de red empleada, buscando un método eficiente para garantizar el éxito del procedimiento.

Para la investigación se implementaron las redes: Kear y red de clasificación:

- Kear: La red Kear (reglas de asociación para la extracción del conocimiento) se define con la creación de los nodos con cada una las reglas posicionadas en ellos y las aristas de acuerdo al orden del resultado obtenido con el algoritmo de extracción.
- Red de clasificación: Se define un elemento desde el cual debe partir la creación de la red. Posteriormente se almacenan las reglas de las cuales el elemento es un consecuente para posteriormente almacenar las reglas de las cuales sus antecedentes son consecuentes hasta almacenar cada una de las reglas extraídas. (Pandey, y otros, 2009).

Métricas para el establecimiento de la red Kear

Las métricas para apoyar el análisis de la red de reglas de asociación son establecidas con el objetivo de seleccionar las más relevantes y crear un ranking de reglas que facilite el estudio. Para la construcción de aristas se utiliza la métrica de distancia de redundancia.

Distancia de redundancia:

Evalúa la probabilidad de que la regla B sea redundante con respecto a la regla A. Una regla tendrá mayor probabilidad de ser redundante en la medida que contenga mayor proporción de elementos que pueden ser redundantes con respecto al total de elementos. (Boudane, y otros, 2017).

Existen 3 condiciones fundamentales por las que un elemento puede ser redundante y corresponden a las condiciones de redundancia siguientes:

$$Sa_1 = \begin{cases} |Y \cap Y_1| & if\ X \subseteq X_1 \\ 0 \end{cases}$$

$$Sa_3 = \begin{cases} |Y \cap X_1| & if\ X \subseteq X_1 \\ 0 \end{cases}$$

$$Sa_4 = \begin{cases} |Y \cap Y_1| & if\ X \subseteq Y_1 \\ 0 \end{cases}$$

1.2.3 Análisis de la red conformada

Para el análisis de la red es importante:

- Identificar las formas eficientes para la exploración de la estructura.
- Proveer un método sistemático para identificar los nodos representativos del conocimiento y sus interdependencias en la red.

Métricas para el análisis de la red Kear

Centralidad local alcanzada

La centralidad local alcanzada es una forma para la identificación de los nodos representativos del conocimiento en el grafo. Es la proporción de todos los nodos en la red que pueden ser alcanzados partiendo desde un nodo i, es denotada como Cr(i). La regla con el valor más alto de la centralidad local es la más interesante en consideración a reducir la redundancia en el modelo de regla de asociación. (Suchacka, y otros, 2016)

Comprensibilidad

Sea $R: A \rightarrow B$ una regla de asociación, la comprensibilidad $C(R)$ de la misma se define como (Bastide, y otros, 2000):

$$C(R) = \frac{1}{|A \cup B|}$$

Creación de ranking de reglas

Para la creación de un ranking de reglas se considera la métrica de centralidad local alcanzada. Sea G una red $kear$ con N reglas y $\alpha \in [0,1]$ el ranking de reglas Ri, con $s0 < i \leq N$, sobre G se denota como $Rg(Ri)$ y se define como:

$$Rg(Ri) = \alpha \; x \; Cr(Ri) + (1 - \alpha) \; x \; C(Ri)$$

Donde Cr es la centralidad local alcanzada por un nodo y C la comprensibilidad. (Agrawal, y otros, 1993)

1.3 Entorno de desarrollo

Lenguaje de programación Python 3.7

Se selecciona el lenguaje de programación Python debido a que ofrece un número considerable de librerías de gran utilidad para los estudios asociados al tratamiento de datos basados en técnicas de inteligencia artificial.

Lenguaje de programación: Python versión 3.7.

Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, usa tipado dinámico y es multiplataforma.

Biblioteca de Python NetworkX

Es un paquete del lenguaje de programación Python para la creación, manipulación, y estudio de estructuras dinámicas y funciones de redes complejas. Manipula estructuras de datos como los grafos, dígrafos y multígrafos y muchos algoritmos basados en los mismos. Ofrece estructuras de redes y medidas de análisis.

En este caso es necesario el empleo de una librería del lenguaje de programación seleccionado que permita el tratamiento de estructuras con representación en forma de redes. Por lo que se selecciona la biblioteca Networkx para realizar esta fase del procedimiento. (Brito Calcada, y otros)

IDE de programación Pycharm

Conclusiones parciales

A partir del análisis bibliográfico se concluye que:

- El enfoque de los investigadores se asocia a proponer soluciones que son clasificadas como ciencia básica por lo que las soluciones son adaptables al contexto de diferentes problemas
- Se emplean diferentes Dataset para la realización de pruebas y experimentos en algunas de las investigaciones

- Se implementan algoritmos que proporcionan una mejora al uso de las exploraciones de modelos de reglas de asociación tradicionales.
- Los resultados resuelven parcialmente la problemática planteada facilitando la exploración del modelo de reglas de asociación
- A partir del análisis de resultados en el método de revisión sistémica de la bibliografía se logra proponer un procedimiento con tres etapas para la construcción de redes de reglas de asociación.

Capítulo 2: Desarrollo de la propuesta de solución

Introducción

En este capítulo se definen las bases para el procedimiento definido en la implementación de la solución, así como los componentes necesarios incluyendo herramientas, técnicas, clases y elementos vinculados a la programación del algoritmo. El epígrafe 2.1 detalla el procedimiento para la exploración de modelos de reglas de asociación mediante redes. Este se divide en tres subepígrafes 2.1.1 que argumenta la etapa de extracción del modelo de reglas de asociación, el 2.1.2 que trata el establecimiento de la red y la 2.1.3 que aborda el análisis de la red conformada.

2.1 Procedimiento para la exploración de modelos de reglas de asociación mediante redes

Para demostrar la validez del procedimiento de extracción de reglas, construcción de la red y análisis de las reglas más relevantes es necesario partir de un Dataset de estudio que será utilizado para ejecutar una prueba de conceptos. El Dataset que se escogió para este fin se denomina breast_cancer el cual contiene los datos de pacientes con padecimientos de cáncer de mama, no para validar los resultados en materia de experimentación de la solución implementada sino para demostrar que el código de codificación del proyecto obtiene resultados basados en el objetivo propuesto.

2.1.1 Extracción de reglas de asociación

Se presentan los ficheros y clases necesarios para garantizar la obtención de todas las reglas que pueden ser generadas de una base de datos de entrada. Se definen para cada una de las mismas los umbrales de soporte y confianza, métricas analizadas en la relevancia de cada una, como se refleja en el capítulo anterior.

La extracción de reglas de asociación se realiza haciendo uso del algoritmo FP_growth. Se utiliza un Dataset de entrada para extraer los conjuntos de ítems presentes en cada una de las transacciones que componen los datos. Seguidamente se determina el soporte de cada uno de los elementos de la base de datos, así como las relaciones entre cada uno de estos ítems en forma de implicación lógica. Finalmente se definen cuáles los valores de las métricas de soporte y confianza para cada uno de las reglas de asociación extraídas.

Clases y ficheros

La etapa de extracción de las reglas de asociación define dos ficheros fundamentales en el tratamiento de diferentes tipos de Dataset los cuales son **generate_rules.py** y **fp_growth.py**, además de otros necesarios para apoyar el proceso. El primer fichero del que se hace uso es **rule.py** el cual define los atributos y métricas de una regla de asociación.

Fichero rule.py

Contiene los atributos, las métricas y los métodos para definir una regla de asociación. La estructura de cada una de las reglas de asociación está definida por un antecedente y un consecuente, los cuales son elementos de la base de datos de estudio que representan implicaciones lógicas para representar tendencias en los datos de estudio.

Para cada una de las reglas de asociación se definen las métricas de calidad: soporte y la confianza. La confianza es el porcentaje de las transacciones de una base de datos donde aparezca el Y aparezca el X. En el caso del soporte expresa el porcentaje en el que aparecen antecedente y consecuente entre el total de las reglas.

Fichero fp_growth.py

El fichero **fp_growth.py** implementa el algoritmo FP_growth para la extracción de reglas de asociación y la definición de soporte y confianza para cada una de ellas. El funcionamiento de este algoritmo está basado en encontrar los grupos de datos frecuentes en un Dataset que satisfagan un soporte mínimo en cada una de las transacciones de la base de datos de estudio y construye una estructura **FPTree**, definida en dicho fichero **fp_growth.py** para almacenar los itemsets frecuentes y determinar las rutas y nodos presentes en el Dataset de análisis.

Las funcionalidades más importantes para la implementación del algoritmo FP_growth son **find_frequent_itemsets,** y las clases **FPTree y FPNode.**

En el caso de la función **find_frequent_itemsets** utiliza el fichero **load_transactions.py** para leer un fichero de entrada de tipo csv y utilizarlo como base de datos para la extracción de transacciones. Posteriormente define un listado de ítems al cual se les asigna un umbral de soporte y son almacenados en una estructura de tipo diccionario para después ser introducidos en una estructura **FPTree**.

```
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)

    def clean_transaction(transaction):
        transaction = list(filter(lambda v: v in items, transaction))
        transaction.sort(key=lambda v: items[v], reverse=True)
        return transaction
    master = FPTree()
    for transaction in map(clean_transaction, transactions):
        master.add(transaction)
```

Figura 2:Obtención de elementos frecuentes con el algoritmo FP_growth

Las clases **FPTree** Y **FPNode** son estructuras auxiliares para garantizar el funcionamiento del algoritmo. Los resultados de la función **find_frequent_itemsets** son almacenados en un **FPTree** y este contiene los elementos en los **FPNode**.

```
class FPTree(object):
    Route = namedtuple('Route', 'head tail')

    def __init__(self):
        self._root = FPNode(self, None, None)
        self._routes = {}

    @property
    def root(self):
        return self._root

    def add(self, transaction):
        point = self._root

        for item in transaction:
            next_point = point.search(item)
            if next_point:
                next_point.increment()
            else:
                next_point = FPNode(self, item)
                point.add(next_point)
                self._update_route(next_point)
            point = next_point

    def _update_route(self, point):
        assert self is point.tree
        try:
            route = self._routes[point.item]
            route[1].neighbor = point
            self._routes[point.item] = self.Route(route[0], point)
        except KeyError:
            self._routes[point.item] = self.Route(point, point)
```

Figura 3:Clase FP_Tree en el algoritmo FP_growth

```
class FPNode(object):
    def __init__(self, tree, item, count=1):
        self._tree = tree
        self._item = item
        self._count = count
        self._parent = None
        self._children = {}
        self._neighbor = None

    def add(self, child):
        if not isinstance(child, FPNode):
            raise TypeError("Can only add other FPNodes as children")
        if not child.item in self._children:
            self._children[child.item] = child
            child.parent = self

    def search(self, item):
        try:
            return self._children[item]
        except KeyError:
            return None
```

Figura 4:Clase FP_Node en el algoritmo FP_growth

Fichero generate_rules.py

Contiene los métodos necesarios para la definición de la estructura de las reglas de asociación extraídas del Dataset de estudio.

Los métodos más importantes de este fichero son el **generate_rules** para la construcción del conjunto de reglas que satisfagan las métricas y los umbrales definidos por parámetro definiendo antecedente, consecuente, soporte y confianza de cada regla, así como la función **write_rules** que escribe las reglas con la estructura especificada.

```
def generate_rules(min_metric=0.5, metric=CONFIDENCE):
    for itemset, support in frequent_itemsets.items():
        index = len(itemset)
        if index > 1:
            for conclusion in get_sub_sets(itemset):
                premise = itemset - frozenset(conclusion)
                if metric == CONFIDENCE:
                    rule_confidence = confidence(premise, itemset)
                    if rule_confidence >= min_metric:
                        yield Rule(premise, conclusion, support, rule_confidence)
                elif metric == CF:
                    cfactor = cf(premise, itemset)
                    if cfactor >= min_metric:
                        yield Rule(premise, conclusion, support, cfactor)
                elif metric == CONVICTION:
                    pass
                elif metric == JACCARD:
                    pass
                elif metric == INTEREST:
                    pass
                elif metric == LEAF:
                    pass
```

Figura 5:Función generate_rules en el fichero generate_rules.py

```
def write_rules(rule_file, value=0.5, metric='confidence'):
    rules_count = 0
    with open(rule_file, 'w') as sfile:
        for rule in generate_rules(value, metric):
            rules_count += 1
            sfile.write(str(list(rule.premise)) + '->' + str(list(rule.conclusion)) + ' supp: ' + str(rule.support)
                        + ' conf: ' + str(rule.confidence) + '\n')
    print(rules_count)
    return rules_count
```

Figura 6:Función write_rules en el fichero generate_rules.py

2.1.2. Establecimiento de la red

Se establecen los tipos de red que pueden implementarse de acuerdo a lo definido en el capítulo anterior para almacenar el modelo de reglas de asociación para posteriormente analizar la red conformada.

Kear

La red Kear se define como un grafo dirigido y en el cual las reglas de asociación están conectadas a otras reglas de asociación en el conjunto de reglas de asociación de acuerdo a algunas métricas de

similaridad. Se propone como métrica de similaridad: similaridad del conocimiento, definidas en el capítulo anterior en las condiciones de redundancia por las que una regla puede ser o no relevante.

El objetivo es obtener conocimiento nuevo que aparece en una regla con respecto a la otra. Las reglas de asociación representan conocimiento mediante las asociaciones entre los ítems de una base de datos, por tanto, es posible usar las asociaciones no novedosas para definir la similaridad del conocimiento.

Clases y ficheros

Para la etapa de establecimiento de la red se definieron dos ficheros fundamentales para el tratamiento de diferentes tipos de Dataset las cuales son **kear.py** y **class_network.py**

El fichero **kear.py** construye una red del tipo **Kear** de reglas de asociación mediante el uso de varias reglas, definidas previamente en el capítulo anterior.

Los métodos que incluye este fichero son el **from_file** que toma un fichero de entrada que almacena el conjunto de reglas extraídas en la fase anterior y las almacena en los nodos de un grafo dirigido que tiene su implementación en el paquete de Python Networkx. Y llama a la función **edge_from_rules**.

```
    @classmethod
    def from_file(cls, file):
        net = nx.DiGraph()

        with open(file, 'r') as rfile:
            for index, line in enumerate(rfile):
                rule = Rule.from_line(line)
                net.add_node(index, line=line, comprehensibility=rule.comprehensibility())
        KearNet.edge_from_rules(net, file)
        network = cls(net)
        return network
```

Figura 7:Función from_file en el fichero kear.py

En el caso del método **edge_from_rules** se crean las aristas en la red entre los diferentes nodos (que contienen las reglas de asociación) hasta definir todas las conexiones posibles evitando los ciclos en el grafo.

```
@staticmethod
def edge_from_rules(network, file):
    for node in network.nodes():
        frule = Rule.from_line(network.node[node]['line'])
        for node2 in network.nodes():
            if node != node2:
                trule = Rule.from_line(network.node[node2]['line'])
                dist = KearNet.redundancy_distance(frule, trule)
                if dist > 0:
                    network.add_edge(node, node2, weight=dist)
```

Figura 8:Función edge_from_rules en el fichero kear.py

Red de clasificación

En el caso específico de este tipo de red se definió una estructura auxiliar para la construcción del grafo denominada camino (path, en inglés). Los caminos son diccionarios que almacenan los elementos que van siendo recorridos a partir de la definición de un consecuente hasta llegar a recorrer el total de nodos que son necesarios para llegar a este objetivo, además de los nodos que fueron recorridos en esa secuencia.

Clases y ficheros

El tipo de red clasificación se encuentra implementado en el fichero **class_network.py** siguiendo el principio de construcción de la red a partir de un consecuente.

Es importante en la implementación de la red la utilización de los métodos vinculados a la obtención de los caminos en el grafo de acuerdo al consecuente definido. El método **get_paths** obtiene todos los caminos presentes en una red de entrada. Esta red de entrada solo contiene los nodos (reglas de asociación) pues la creación de aristas es dependiente a los caminos resultantes.

```
def get_paths(network,class_item):
    notstop = True
    temp = defaultdict(lambda :{'node':[],'items':[], 'end': 0})
    paths = defaultdict(lambda :{'node':[],'items':[], 'end': 0})
    paths[0]['node'].append(len(network.nodes))
    paths[0]['items'].append(class_item)
    while notstop:
        notstop = False
        for key in paths.keys():
            if paths[key]['end'] == 0:
                rules = ClassNet.get_rules(network,frozenset([paths[key]['items'][len(paths[key]['items'])-1]]),paths[key]['items'])
                if len(rules) > 0:
                    paths[key]['end'] = 1
                    notstop = True
                    for node in rules:
                        rule = Rule.from_line(network.nodes[node]['line'])
                        antecedents = ClassNet.get_antecedents(rule)
                        for item in antecedents:
                            i = len(temp)
                            temp[i]['node'] = deepcopy(paths[key]['node'])
                            temp[i]['node'].append(node)
                            temp[i]['items'] = deepcopy(paths[key]['items'])
                            temp[i]['items'].append(item)

                else:
                    paths[key]['end'] = 2
        for key in temp.keys():
            i = len(paths)
            paths[i]['node'] = deepcopy(temp[key]['node'])
            paths[i]['items'] = deepcopy(temp[key]['items'])
        temp.clear()
    return paths
```

Figura 9:Función get_paths en el fichero class_network.py

En el caso de la creación de aristas se implementa el método **edge_from_path** que crea los enlaces entre los nodos de acuerdo a los caminos obtenidos conformando así una red que nos ofrezca todas las vías de llegada para arribar al nodo objetivo.

```
@staticmethod
def edge_from_path(network,class_item):
    paths = ClassNet.get_paths(network,class_item)
    for path in paths.keys():
        if paths[path]['end'] == 2:
            edges = ClassNet.get_path_edges(paths[path]['node'])
            network.add_edges_from(edges)
```

Figura 10: Función edge_from_path en el fichero class_network.py

2.1.3 Análisis de la red conformada

Se aplican métricas de calidad para la obtención de reglas que representen conocimiento en el grafo. Estas reglas son importantes porque brindan información general del modelo de reglas de asociación permitiendo la reducción de redundancia y la disminución de la complejidad del modelo.

Kear

Para el análisis de la red Kear se seleccionaron la métrica de centralidad local alcanzada y la definición de un ranking de reglas, definidos en el capítulo anterior con el propósito de encontrar las reglas y los ítems que representan conocimiento en el modelo de reglas de asociación extraído. El objetivo es presentar al usuario un conjunto de reglas reducido para facilitar el estudio del modelo. Es importante también la definición de la complejidad global del modelo que se define con la siguiente fórmula $(\text{Chadha, y otros, } 2000)$:

$$Cm = 1 - e^{(Nr/(Ne*10)}$$

Siendo Nr el número de reglas de asociación del modelo y Ne el número de elementos. Esta fórmula solo es válida cuando el número de reglas no excede a 1000. De otra manera se calcula:

$$Cm = (1 - \alpha) * Cp$$

Siendo $0 < \alpha < 1$ y Cp la comprensibilidad del modelo.

Conclusiones parciales

La utilización del método científico de implementación y la aplicación de estándares asociados al uso de las bibliotecas de clases Networkx permitió la construcción de Kear una red de reglas de asociación que utiliza la redundancia como métrica de conexión entre los nodos.

Capítulo 3: Validación de la propuesta de solución

Introducción

Se presentan los elementos necesarios para la validación de la propuesta. Se realizaron dos experimentos: el primero para comprobar la independencia del método de reducción de redundancia con respecto a la métrica de calidad selecciona para la extracción de reglas de asociación. El segundo para comprobar la reducción de complejidad en el modelo de reglas de asociación obtenido.

3.1 Experimento 1: Independencia de las métricas de calidad

Para la validación de la implementación se desarrolló un experimento para cada uno de los Dataset seleccionados desde el repositorio del UCI Machine Learning Repository. El objetivo del primer experimento es garantizar la mayor reducción de redundancia en cada una de las transformaciones de los Dataset para presentar un modelo analizable al usuario, mediante un conjunto de reglas representativas del conocimiento extraído. La Tabla 6 muestra los Dataset que van a ser utilizados en los experimentos.

Tabla 6. Dataset de estudio

Dataset	No.Filas	Atributos	Atributos numéricos
adult	48842	14	5
haberman	306	3	3
Car	1728	6	0

En primer lugar, se evalúa el efecto de las métricas de calidad sobre la reducción de las reglas para demostrar que la métrica de redundancia utilizada es independiente de la métrica de calidad. Para ello se computó el valor de reducción para varios Dataset utilizando la confianza y el factor de certeza. Cada experimento consiste en definir varios valores para el soporte para obtener el conjunto de ítems frecuentes en cada uno de los Dataset. Después se definieron varios valores para las métricas de calidad confianza y factor de certeza para obtener el conjunto de reglas de asociación que satisfacen el umbral definido. Finalmente se define el porciento de poda que fue obtenido con la aplicación de la métrica de calidad con respecto al total de reglas. Se realizaron un total de 54 corridas del experimento.

La Tabla 7 presenta las características generales del experimento:

- La columna uno contiene el id de las iteraciones.
- La columna dos contiene el nombre del Dataset.

- La columna tres contiene el umbral se soporte usado para extraer los ítems frecuentes.
- La columna cuatro contiene la métrica de calidad utilizada.
- La columna cinco contiene el total de reglas minadas.
- La columna seis contiene el radio entre de la cantidad de reglas podadas con respecto al total de reglas.

Tabla 7. Comparación entre las métricas de calidad

ID	Datos	Soporte	Métrica	Reglas	Radio de poda
1.	Adult	0.25	Conf:0.5	3136	62
2.	Adult	0.25	Fc:0.5	704	66.9
3.	Adult	0.25	Conf:0.6	2469	63.4
4.	Adult	0.25	Fc:0.6	579	68.2
5.	Adult	0.25	Conf:0.7	1999	65.2
6.	Adult	0.25	Fc:0.7	422	68.4
7.	Adult	0.3	Conf:0.5	1804	58.9
8.	Adult	0.3	Fc:0.5	369	57.4
9.	Adult	0.3	Conf:0.6	1426	59.3
10.	Adult	0.3	Fc:0.6	325	58.7
11.	Adult	0.3	Conf:0.7	1148	59.4
12.	Adult	0.3	Fc:0.7	422	61.1
13.	Adult	0.35	Conf:0.5	849	49.11
14.	Adult	0.35	Fc:0.5	79	31.6
15.	Adult	0.35	Conf:0.6	683	51.9
16.	Adult	0.35	Fc:0.6	77	32.4
17.	Adult	0.35	Conf:0.7	531	50.2
18.	Adult	0.35	Fc:0.7	64	35.9
19.	Haberman	0.25	Conf:0.5	5939	56
20.	Haberman	0.25	Fc:0.5	3978	62.4
21.	Haberman	0.25	Conf:0.6	5498	57.9
22.	Haberman	0.25	Fc:0.6	2626	65.4
23.	Haberman	0.25	Conf:0.7	4297	61.1
24.	Haberman	0.25	Fc:0.7	1493	64.6
25.	Haberman	0.3	Conf:0.5	5207	53.1
26.	Haberman	0.3	Fc:0.5	3606	59.9
27.	Haberman	0.3	Conf:0.6	4879	54.9

28.	Haberman	0.3	Fc:0.6	3909	63.2
29.	Haberman	0.3	Conf:0.7	1402	58.5
30.	Haberman	0.3	Fc:0.7	3070	62.6
31.	Haberman	0.35	Conf:0.5	2293	46.4
32.	Haberman	0.35	Fc:0.5	2961	51.6
33.	Haberman	0.35	Conf:0.6	1636	48.3
34.	Haberman	0.35	Fc:0.6	2514	56.7
35.	Haberman	0.35	Conf:0.7	993	50.2
36.	Haberman	0.35	Fc:0.7	92089	57.2
37.	Car	0.25	Conf:0.5	51159	49.1
38.	Car	0.25	Fc:0.5	76706	45.4
39.	Car	0.25	Conf:0.6	42066	47.7
40.	Car	0.25	Fc:0.6	61963	43.2
41.	Car	0.25	Conf:0.7	31320	45.6
42.	Car	0.25	Fc:0.7	78998	38.6
43.	Car	0.3	Conf:0.5	48390	50.3
44.	Car	0.3	Fc:0.5	68759	48.1
45.	Car	0.3	Conf:0.6	40776	49.2
46.	Car	0.3	Fc:0.6	57634	47.1
47.	Car	0.3	Conf:0.7	30351	47.7
48.	Car	0.3	Fc:0.7	13908	44.5
49.	Car	0.35	Conf:0.5	6369	47.0
50.	Car	0.35	Fc:0.5	12031	42.7
51.	Car	0.35	Conf:0.6	5397	45.6
52.	Car	0.35	Fc:0.6	9822	40.6
53.	Car	0.35	Conf:0.7	4380	43.2
54.	Car	0.35	Fc:0.7	5431	37.4

De esta tabla se pueden extraer dos hechos importantes. El primer hecho es que el índice porcentual de poda alrededor del 50 por ciento es obtenido con un conjunto pequeño de conocimiento, y la segunda es que los radios de poda más elevados se obtienen con los valores más bajos en la métrica. Posteriormente se determinó a partir del análisis ANOVA de una variable la independencia de la métrica con respecto a la reducción.

Tabla 8:adult_0.5

Origen de las variaciones	***Suma de cuadrados***	***Grados de libertad***	***Promedio de los cuadrados***	***F***	***Probabilidad***	***Valor crítico para F***
Entre grupos	33.1820167	1	39.015	0.2064832	0.6731043	7.70864742
Dentro de los grupos	755.8	4	188.95			
Total	794.815	5				

Tabla 9:adult_0.6

Origen de las variaciones	***Suma de cuadrados***	***Grados de libertad***	***Promedio de los cuadrados***	***F***	***Probabilidad***	***Valor crítico para F***
Entre grupos	39.015	1	33.1820167	0.17513486	0.69708067	7.70864742
Dentro de los grupos	757.862067	4	189.465517			
Total	791.044083	5				

Tabla 10:adult_0.7

Origen de las variaciones	***Suma de cuadrados***	***Grados de libertad***	***Promedio de los cuadrados***	***F***	***Probabilidad***	***Valor crítico para F***
Entre grupos	39.015	1	33.1820167	0.17513486	0.69708067	7.70864742
Dentro de los grupos	757.862067	4	189.465517			
Total	791.044083	5				

Tabla 11: haberman_0.5

Origen de las variaciones	***Suma de cuadrados***	***Grados de libertad***	***Promedio de los cuadrados***	***F***	***Probabilidad***	***Valor crítico para F***
Entre grupos	56.4266667	1	56.4266667	2.00782825	0.22944833	7.70864742
Dentro de los grupos	112.413333	4	28.1033333			
Total	168.84	5				

Tabla 12: haberman_0.6

Origen de las variaciones	***Suma de cuadrados***	***Grados de libertad***	***Promedio de los cuadrados***	***F***	***Probabilidad***	***Valor crítico para F***
Entre grupos	97.6066667	1	97.6066667	4.37861682	0.10453327	7.70864742
Dentro de los grupos	89.1666667	4	22.2916667			
Total	186.773333	5				

Tabla 13: haberman_0.7

Origen de las variaciones	***Suma de cuadrados***	***Grados de libertad***	***Promedio de los cuadrados***	***F***	***Probabilidad***	***Valor crítico para F***
Entre grupos	35.5266667	1	35.5266667	1.50973865	0.28652426	7.70864742
Dentro de los grupos	94.1266667	4	23.5316667			
Total	129.653333	5				

Tabla 14: car_0.5

Origen de las variaciones	***Suma de cuadrados***	***Grados de libertad***	***Promedio de los cuadrados***	***F***	***Probabilidad***	***Valor crítico para F***
Entre grupos	17.34	1	17.34	3.44047619	0.137216	7.70864742
Dentro de los grupos	20.16	4	5.04			
Total	37.5	5				

Tabla 15: car_0.6

Origen de las variaciones	***Suma de cuadrados***	***Grados de libertad***	***Promedio de los cuadrados***	***F***	***Probabilidad***	***Valor crítico para F***
Entre grupos	22.4266667	1	22.4266667	3.20992366	0.14767077	7.70864742
Dentro de los grupos	27.9466667	4	6.98666667			
Total	50.3733333	5				

Tabla 16: car_0.7

Origen de las variaciones	***Suma de cuadrados***	***Grados de libertad***	***Promedio de los cuadrados***	***F***	***Probabilidad***	***Valor crítico para F***
Entre grupos	42.6666667	1	42.6666667	1.50973865	0.28652426	7.70864742
Dentro de los grupos	39.0266667	4	9.75666667			
Total	81.6933333	5				

Los resultados obtenidos para cada Dataset (adult,haberman,car) para distintos valores en las métricas de calidad (0.5,0.6,0.7) se ven reflejados en las Tablas 8,9 y 10 en el caso del Dataset adult; las Tablas 11,12 y 13 en el caso del Dataset haberman; y las Tabla 14,15 y 16 en el caso del Dataset car. En todos los resultados obtenidos el valor de F nunca superó el valor de F crítica lo que apoya la hipótesis de que no existen diferencias significativas en la utilización de las métricas de calidad.

3.2 Experimento 2: Reducción de redundancia

Un segundo experimento está orientado a verificar los beneficios desempeño real usando conocimiento real en un Dataset sintético. Se generan algunos datos con un conjunto de reglas representativas del conocimiento y el proceso Kear es aplicado dos veces para con un conjunto de reglas seleccionadas aleatoriamente como conocimiento para ordenar las reglas conocidas de acuerdo a su prioridad.

Para la obtención de reglas se usa el siguiente procedimiento. Los datos sintéticos contienen algunas variables independientes y seleccionadas aleatoriamente $X1, X2, X3$ y una variable booleana Z. La regresión logística del modelo es usada para predecir la probabilidad de que Z sea verdadera. La ecuación muestra la expresión para calcular la contribución de cada predictor individual, el cual es llamado logit.

$$y = \beta 0 + \beta 1 X1 + \beta 2 X2 + B3 X3 + \epsilon$$

El termino ϵ en la ecuación anterior representa el error que no puede ser aceptada por las variables predictoras. Se asume que para tener una distribución normal con significado 0 y 0.1 de varianza. El parámetro $\beta 0$ representa el efecto inicial del sistema, en este caso un valor de 1.999. Los parámetros $\beta 1, \beta 2, \beta 3$ son coeficientes de regresión de las variables $X1, X2, X3$ y son de valor -1. Para estos valores la ecuación es:

$$y = 1.999 - X1 - X2 - X3 + \epsilon$$

Dado el logit la solución logística está definida como:

$$f(y) = \frac{e^y}{e^y + 1}$$

Por ello el estado de Z puede ser determinado por el valor de salida de la función logística:

$$Z = \begin{cases} 1 \; if \; f(x) \leq 0.5 \\ 0 \; if \; f(x) > 0.5 \end{cases}$$

En la generación de los datos primero son asignados valores al azar a $X1, X2, X3$. Entonces el valor de logit es calculado. La función logística es evaluada. El estado de Z es calculado de la ecuación anterior. Los valores esperados de la función logística para todas las combinaciones son listados en la Tabla 18.

Tabla 17: Función logística para todas las combinaciones de predictores

$X1$	$X2$	$X3$	y	$f(y)$	$f'(y)$
0	0	0	1.999	0.88	0.12
1	0	0	0.999	0.73	0.27
0	1	0	0.999	0.73	0.27
0	0	1	0.999	0.73	0.27
1	1	0	-0.001	0.49	0.5
1	0	1	-0.001	0.49	0.5
0	1	1	-0.001	0.49	0.5
1	1	1	-1.001	0.27	0.73

Las regularidades listadas en la tabla 18 generan 16 reglas de asociación, si estas cumplen con los umbrales de soporte y confianza definidos. Existen cuatro reglas de asociación que satisfacen la confianza de 0.5 y valor de $Z = 1$ en el consecuente. Estas cuatro reglas son consideradas como las reglas importantes (listadas en la tabla 19). Las reglas de asociación sin Z en el consecuente son consideradas no interesantes.

Tabla 18: Reglas de asociación principales

Reglas de asociación
$\{x1 = 1, x2 = 1, x3 = 0\} \rightarrow \{z = 1\} confidence \geq 0.5$
$\{x1 = 1, x2 = 0, x3 = 1\} \rightarrow \{z = 1\} confidence \geq 0.5$
$\{x0 = 1, x2 = 1, x3 = 1\} \rightarrow \{z = 1\} confidence \geq 0.5$
$\{x1 = 1, x2 = 1, x3 = 1\} \rightarrow \{z = 1\} confidence \geq 0.5$

Se adicionan dos variables $X4, X5$. Estas invariables no tienen repercusión en el estado de Z. $x4$ depende de los predictores, cuando al menos dos predictores son activos $X4 = 1$ y $X5$ recibe un valor al azar.

Los resultados son mostrados en la tabla. La primera columna contiene el número de transacciones del Dataset sintético. La segunda columna contiene el número de reglas generadas. La cuarta columna presenta la reducción del espacio de exploración (para un conocimiento aleatorio en la mejor de 30 ejecuciones) mientras la quinta columna presenta la complejidad final del modelo.

Tabla 19: Reducción del espacio de exploración para los Dataset sintéticos

Transacción	No. Reglas	Conocimiento	Redundancia	Complejidad
1000	1906	random	64.14	0.80
1000	1906	know	98.06	0.28
5000	1826	random	64.14	0.80
5000	1826	know	97.21	0.28
10000	1883	random	64.14	0.80
10000	1883	know	98.15	0.28
50000	1845	random	59.95	0.81
50000	1845	know	90.79	0.28

En la Tabla 19 se puede observar en la columna el nombre de los Dataset de estudio. En la segunda el número de reglas del modelo extraído en cada una. En la tercera los mejores resultados en las iteraciones en la reducción de redundancia. En la cuarta columna los peores resultados en la reducción de redundancia, y en la quinta columna los mejores resultados en la reducción de la complejidad del modelo.

Tabla 20. Distribución de conocimiento, redundancia y complejidad de los Dataset

Nombre del Dataset	No. Reglas	Mejor reducción	Peor reducción	Complejidad del modelo
adult	677	671	285	0.07001646039942878
haberman	1180	1129	443	0.2745682801533705
Car	1036	985	331	0.27851027361742275

La variación existente entre el mejor y el peor resultado de reducción de redundancia se debe a la selección de las reglas que representan el conocimiento. La selección de reglas presentadas al usuario influye de manera significativa en la reducción del espacio de exploración y la complejidad final del modelo.

Realizando la experimentación con los Dataset *adult, haberman y car* se obtuvieron una serie de resultados reflejados en las tablas 21, 22 y 23 respectivamente donde la primera columna nos muestra el número de la iteración del experimento; la segunda nos muestra la longitud del conocimiento (contiene las reglas resultantes del proceso de ranking); la tercera columna nos muestra la complejidad global del modelo de reglas de asociación y la cuarta la redundancia del modelo. También se muestra una representación de esto en las ilustraciones 11,12 y 13. Esta distribución puede observarse más claramente en la distribución de las iteraciones en cada uno de los Dataset representada por los gráficos de barras:

Tabla 21. Iteraciones del experimento en el Dataset adult

No.Iteración	Conocimiento	Complejidad global	Redundancia
1	10	0.7954933387623639	285
2	19	0.5747927252078363	538
3	28	0.5747927252078363	538
4	38	0.3124919548279839	620
5	47	0.07001646039942878	671

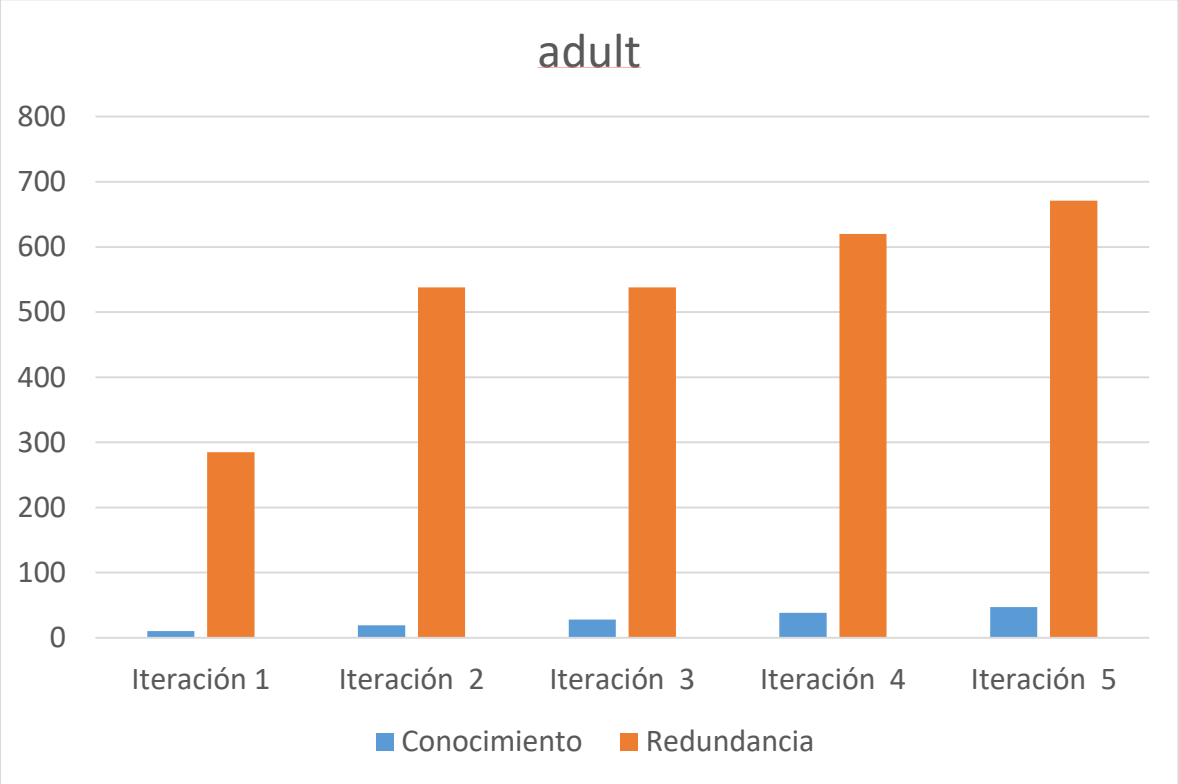

Figura 11. Reducción de redundancia en el Dataset Adult

Tabla 22. Iteraciones del experimento en el Dataset haberman

No.Iteración	Conocimiento	Complejidad global	Redundancia
1	8	0.799081523248422	443
2	15	0.787223221722477	690
3	25	0.45428221750806297	1079
4	35	0.2745682801533705	1129

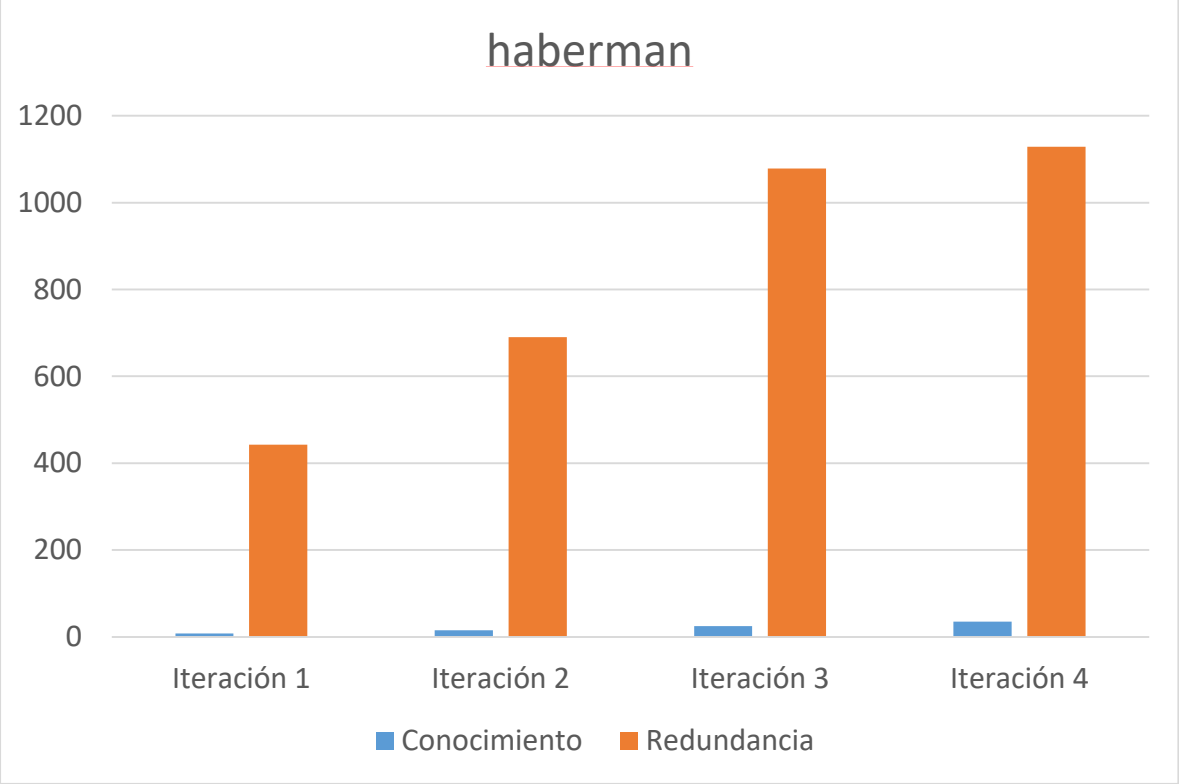

Figura 12. Reducción de redundancia en el Dataset Haberman

Tabla 23. Iteraciones del experimento en el Dataset car

No.Iteración	Conocimiento	Complejidad global	Redundancia
1	10	0.8060674944438415	331
2	19	0.800996223121452	472
3	29	0.7879391016623349	610
4	39	0.6054399898418553	872
5	49	0.5066530620961117	917
6	59	0.5066530620961117	917
7	68	0.5066530620961117	917
8	78	0.27851027361742275	985

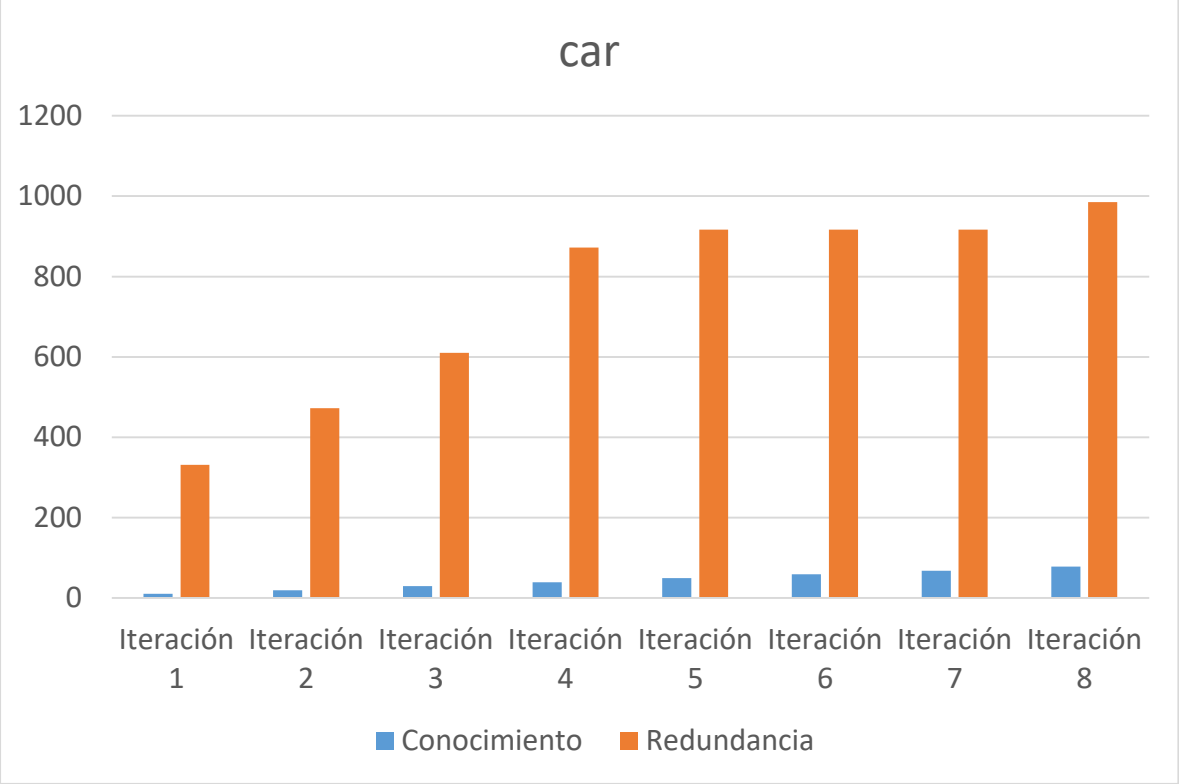

Figura 13. Reducción de redundancia en el Dataset Car

Los resultados son una mejora de las soluciones convencionales para la exploración de modelos de reglas de asociación. Los trabajos previos trataban de eliminar redundancia de un modelo con la utilización de reglas de inferencias. La utilización del conocimiento para la reducción de reglas redundantes en un modelo de reglas de asociación que no aportan conocimiento novedoso para facilitar la toma de decisiones.

Conclusiones parciales

Los experimentos realizados permitieron asegurar que:

- La reducción de redundancia es independiente de la métrica de calidad utilizada.
- Con conocimiento previo relativamente pequeño (1%) se obtienen tasas de reducción de hasta el 90%.

Conclusiones Generales

Finalizada la investigación se pudo concluir que:

- El proceso de revisión sistémica de la bibliografía permitió determinar que el enfoque de los investigadores se asocia a proponer soluciones que son clasificadas como ciencia básica por lo que las soluciones son adaptables al contexto de diferentes problemas. Se emplean diferentes Dataset para la realización de pruebas y experimentos en algunas de las investigaciones. Se implementan algoritmos que proporcionan una mejora al uso de las exploraciones de modelos de reglas de asociación tradicionales. Los resultados resuelven parcialmente la problemática planteada facilitando la exploración del modelo de reglas de asociación. A partir del análisis de resultados en el método de revisión sistémica de la bibliografía se logra proponer un procedimiento con tres etapas para la construcción de redes de reglas de asociación.
- La utilización del método científico de implementación y la aplicación de estándares asociados al uso de las bibliotecas de clases Networkx permitió la construcción de Kear una red de reglas de asociación que utiliza la redundancia como métrica de conexión entre los nodos.
- La reducción de redundancia es independiente de la métrica de calidad utilizada. Con conocimiento previo relativamente pequeño (1%) se obtienen tasas de reducción de hasta el 90%.

Recomendaciones

Extender la red para que sea capaz de utilizar otros tipos de reglas de asociación como por ejemplo las reglas de asociación difusas y multinivel.

Construir una interfaz para visualizar los resultados del análisis de la red.

Bibliografía

Boudane, Abdelhamid, et al. 2017. *Enumerating Non-redundant Association Rules Using Satisfiability.* 2017.

Hahsler, Michael and Karpienko, Radoslaw . 2016. *Visualizing association rules in hierarchical groups.* 2016.

Agrawal, Rakesh, Imiélinski, Tomasz and Swami, Arun. 1993. *Mining association rules between sets of items in large databases.* s.l. : ACM, 1993.

Bastide, Yves, et al. 2000. *Mining minimal non-redundant association rules using frequent closed itemsets.* s.l. : Computational LogicCL, 2000.

Berzal, Fernando, et al. *Measuring the accuracy and interest of association rules: A new framework.*

Bhargava, Niket. 2016. *Survey of Interestingness Measures for Association Rules Mining : Data Mining , Data Science for Business Perspective.* s.l. : IRACST -International Journal of Computer Science and Information Technology & Security, 2016.

Brito Calcada, Dario, de Padua, Renan and Oliveira Rezende, Solange. *Asymmetric objective measures applied to filter Association Rules Networks.* São Carlos : s.n.

Calcada, Dario, Oliveira Rezende, Solange and Teodoro, M.S. *Analysis of decomposition parameters of green manure in the Brazilian Northeast with Association Rules Network.*

Chadha, Atul , et al. 2000. *Dimension reduction using association rules for data mining application.* 2000.

de Padua, Renan, et al. *EXPLORING MUSICAL RELATIONS USING ASSOCIATION RULE NETWORKS.*

de Padua, Renan, et al. *Exploring the Data Using Extended Association Rule Network.*

de Padua, Renan, Oliveira Rezende, Solange and Oliverira de Carvalho, Veronica. 2014. *Post-Processing Association Rules using Networks and Transductive Learning.* s.l. : International Conference on Machine Learning and Applications, 2014.

Franco Arcega, Anilu, et al. 2008. *Árboles de decisión para grandes conjuntos de datos.* Puebla : s.n., 2008.

Gao, Yuanyuan , et al. 2016. *Incorporating association rule networks in feature category-weighted naive Bayes model to support weaning decision making.* s.l. : Decision Support Systems, 2016.

González-Salcedo, Luis Octavio, et al. 2016. *Redes neuronales artificiales para estimar propiedades en estado fresco y endurecido, para hormigones reforzados con fibras metálicas.* Tucuman : Cuaderno Activa, 2016.

Juan, Wang, Suqiu, Li and Xiaoliang, Feng. *Extraction of Non-redundant Association Rules from Concept Lattices based on IsoFCA.*

Lin, Dennis KJ and Changpetch, Pannapa. 2015. *Interestingness Measures for Classification Rule Mining: Model Selection Ability.* s.l. : International Journal of Computer Science Issues (IJCSI), 2015.

Pandey, Avinash, Sharma, Anuradha and Agrawal, Krishna Kant. 2017. *Developing Efficient Data Mining Algorithms.* s.l. : IEEE Xplore Compliant, 2017.

Pandey, Gaurav, et al. 2009. *Association Rules Network: Definition and Applications.* 2009.

Rajaraman, Anand and Ullman, Jeffrey David . 2011. *Mining of Massive Datasets.* 2011.

Roiger, R.J. 2017. *Data mining: a tutorial-based primer.* 2017.

Rossi, Rafael, Lopes, Alneu and Rezende, Solange. *A Parameter-Free Label Propagation Algorithm Using Bipartite Heterogeneous Networks for Text Classification.*

Sammut, Claude and Webb, Geoffrey I. . 2017. *Encyclopedia of Machine Learning and Data Mining.* 2017.

Suchacka, Grażyna and Chodak, Grzegorz . 2016. *Using association rules to assess purchase probability in online stores.* 2016.

Yong, Liu and Xueqing, Li. 2017. *Application of Formal Concept Analysis in Association Rule Mining.* Changsha : IEEE, 2017.

Yuan, Xiuli. 2017. *An improved Apriori algorithm for mining association rules.* s.l. : AIP Conference Proceedings, 2017.

Zhang, Shichao and Wu, Zindong. 2011. *Fundamentals of association rules in data mining and knowledge discovery.* 2011.

Zhou, Ying, et al. 2019. *Combining association rules mining with complex networks to monitor coupled risks.* s.l. : Reliability Engineering & System Safety, 2019.