Universidad de las Ciencias Informáticas

Facultad 2

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

"Extensión de índices de validación de grupo en la herramienta WEKA para la evaluación de algoritmos de agrupamiento"

Autores: Yohandra Echverria Castillo

Clara Haydee González Carbonell

Tutor: MsC. Héctor Raúl González Diez

Ciudad de La Habana, junio 2015

Si me preguntan cuál es la palabra más bella, diré "Patria", y si me preguntan por otra casi tan bella como "Patria", responderé "Amistad".

Declaración de autoría

Declaramos ser los legítimos autores del trabajo de diploma titulado: “Extensión de índices de validación de grupo en la herramienta WEKA para la evaluación de algoritmos de agrupamiento”, y reconocemos a la Universidad de las Ciencias Informáticas (UCI) los derechos patrimoniales del mismo, con carácter exclusivo.

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

_____ _____

Yohandra Echeverria Castillo Clara Haydeé González Carbonell

_____

MsC. Héctor Raúl González Diez

DEDICATORIA

Eres la razón por la que seguí adelante cuando muchos pensaban que se había acabado el camino. A ti dedico mi tesis, por ser la abuela más adorable del mundo, a mi abuelita Haydee por creer en mí.

Porque cuando me dices "Mami te quiero" mi mundo cambia de color. Le dedico me tesis a mi hijo, con amor de mami. Te quiero mi niño.

Clara Haydee González Carbonell

Cuando no tenía a quien acudir, sabía que podía contar contigo. Cuando todos las puertas se cerraban tu eres la única abierta. Cuando todo se ponía difícil ahí estabas tú a mi lado, diciéndome que todo saldría bien. Gracias Mamá por todo lo hiciste, todo te lo debo a ti.

Yohandra Echeverria Castillo

AGRADECIMIENTOS

Luego de 5 años, que requirieron esfuerzo y sacrificio, finalmente llegó el momento de agradecer a las personas que fueron importantes en esta etapa.

Uno de tus sueños siempre ha sido verme convertida en una profesional, hoy he alcanzado ese sueño para ti,

A mi mamita, por siempre estar ahí para mí, por darme la solución a cada problema que se me presento, por creer en mí cuando todo indicaba que las cosas no irían bien.

A mi papa por hacerme sentir orgullosa de ser su hija.

A mi abuela Haydee por cada palabra, por cada gesto de comprensión, amor y regano. Abue ya tu nieta es ingeniera.

Agradecer a la hermana que encontré en esta escuela, el destino nos unió y dudo que pueda separarnos. Sino fuéramos tan parecidas no nos lleváramos tan bien. Te agradezco por hacer de tu casa mi casa, por transitar junto a mi estos cinco largos años, que relacionan llantos y sobre todo alegría. Gracias por ser mi compañera de tesis y porque durante este proceso aguantaste mis arrebatos.

A mi titi, por tu amor, tu comprensión, por darme tú apoyo incondicional. Gracias por estar ahí cuando lo necesite y darme tu ayuda en estos últimos meses que han sido duros para ti y para mí. Te quiero mucho.

Clara Haydee González Carbonell

AGRADECIMIENTOS

Han sido 5 años llenos de esfuerzo y sacrificio, cerrada esta etapa me queda agradecer

A mi madre por ser madre y padre, por la devoción que tiene a sus hijos, por el apoyo limitado e incondicional que siempre me has dado, por tener la fortaleza de salir adelante sin importar los obstáculos, por haberme formado como una mujer de bien, y por ser la mujer que me dio la vida y me enseña a vivirla. No hay palabras en este mundo para agradecerte.

A mis hermanos Yarislán y Yoander, por su empeño, su constancia y su cariño incondicional.

A mis abuelos, porque han sido y serán ejemplo incuestionable de fortaleza, integridad y sabiduría.

A toda mi familia, por sus palabras de aliento, especialmente a mis tíos Michel y Alexis.

A mi novio quien se estresó y me hizo en cada momento junto a mí. Por ser capaz de sostenerme cundo todo iba mal, por ser mi amigo y mi novio. Gracias por amarme y hacerme fuerte.

A la persona que me acompaño en todas las batallas libradas, en la universidad. Te doy las gracias hoy por las pelas, por los regaños, por escucharme, por ser mi espía, por defenderme con unas y diente, porque más que una amiga, eres mi hermana, por ayudarme a levantarme cuando tropezaba. Le doy las gracias al gato por habernos acercado, en fin te doy las gracias por ser insoportablemente tú Clara Haydee.

Yohandra Echeverria Castillo

Les agradecemos a nuestro tutor Héctor, por ser profesor y amigo y guiarnos en esta difícil tarea.

Les agradecemos a todos aquellos que contribuyeron de una forma u otra a la terminación de esta tesis, especialmente a Exxon, Robert, Jova, Vichi, German, Frank, Gaby, Darío.

Les agradecemos a las chicas del Latex, por acompañarnos en este difícil camino, por brindarnos casa, comida y cama.

En especial a Robert y a Exxon por soportarnos y cargar con nosotras. Los queremos

Clara Haydee González Carbonell y Yohandra Echeverria Castillo

RESUMEN

En el ámbito del creciente desarrollo de los algoritmos de agrupamiento en innumerables áreas de la sociedad, y debido a la imprecisión de la herramienta WEKA para la validación de la calidad de resultados de estos algoritmos, surge el Trabajo de Diploma: “Extensión de índices de validación de grupo en la herramienta WEKA para la evaluación de algoritmos de agrupamiento”. Esta investigación tiene como objetivo medir la calidad de las particiones resultantes de los algoritmos de agrupamiento y así contribuir en el diseño de experimentos, para recopilar información sobre diferentes tipos de datos. Es por ello que se integra un conjunto de métricas o índices de validación externas e internas a la herramienta. La solución ofrecerá que la herramienta WEKA permita la validación de calidad para los algoritmos de agrupamiento, lo que permitirá llevar acabo comparación entre algoritmos.

Palabras clave: agrupamiento, validación, calidad, índice de validación de grupo, validación interna, validación externa, WEKA.

ABSTRACT

In the space of increasing development of the clustering's algorithms in innumerable areas of the society, and due to the imprecision of the tool WEKA for the validation of quality of aftermath of these algorithms, Diploma's Work appears: "Extending validity index WEKA cluster in the tool for evaluating clustering algorithms". This research aims to measure the quality of partitions resulting from the clustering algorithms and thus contribute to the design of experiments, to collect information about different types of data. That is why a set of metrics or indices of external and internal validation tool is integrated. The solution will allow the validation tool WEKA quality for clustering algorithms, which will carry out comparison between algorithms.

Keywords: *clustering, validation, quality, cluster validity index, internal validity, external validity, WEKA.*

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

El Aprendizaje Automático es la rama de la Inteligencia Artificial que se dedica al estudio de los programas que aprenden o evolucionan basados en su experiencia, para realizar una tarea determinada cada vez mejor. En las últimas décadas, el uso de técnicas de aprendizaje automático en diversos campos como la informática, la estadística, la robótica, la medicina, etc. se ha desarrollado de manera extraordinaria.

Además, el aprendizaje automático es una técnica que utiliza la minería de datos en su proceso de conversión de datos en conocimiento, Para agilizar el mismo se maneja la extracción de modelos, utilizando los árboles de clasificación como herramientas para eliminar los resultados innecesarios. Esto lo convierte en un motor de consultas que permite realizar ordenamientos y selección de datos. Entre los diferentes algoritmos de aprendizaje automático, se encuentran los de clasificación supervisada y los de clasificación no supervisada (Caparrini, 2013).

Entre las tareas más utilizadas del aprendizaje supervisado se encuentra la predicción y la clasificación, mientras que en el aprendizaje no supervisado están agrupamiento y asociación (Weiss, 1998). La clasificación supervisada es conocida también como clasificación con aprendizaje, se conoce a priori la clase a la que pertenecen cada uno de los objetos. La clasificación no supervisada es conocida como clasificación sin aprendizaje, no se toma en cuenta la información de las clases, debido a que aprende a partir de la naturaleza intrínseca de los datos.

Entre los algoritmos que utiliza la clasificación no supervisada se encuentran: Los algoritmos de agrupamiento. Estos algoritmos tienen como objetivo aglomerar un conjunto de objetos, en dependencia de la naturaleza de los rasgos que caracterizan dichos objetos, basándose en la similitud entre estos. Para medir la similitud entre objetos se utilizan diferentes funciones de distancia: distancia Euclídea, de Manhattan, de Mahalanobis, etc.

Existen varios enfoques de algoritmos de agrupamiento, debido a que para un mismo conjunto de datos, aplicando diferentes algoritmos de agrupamiento se pueden obtener resultados diferentes. Es por esto que surge la necesidad de evaluar las particiones obtenidas y poder determinar la calidad de los resultados alcanzados. Existen medidas que permiten realizar una evaluación de la estructura resultante de cada algoritmo de agrupamiento, obteniendo de manera cuantitativa un valor de calidad de las mismas. Estas medidas de calidad se conocen bajo el nombre de índices de validación de grupo.

Un índice de validación de grupo es una función, que mide la bondad o calidad de una partición. Se define que a menor valor de la función mejor será la partición. Un índice de validación proporciona una medida objetivo del resultado de un agrupamiento, y su valor óptimo, se usa frecuentemente para indicar la mejor selección posible ya que permiten:

- Cuantificar a través de una medida la calidad del agrupamiento para una determinada base de datos.
- Determinar una configuración adecuada de los parámetros de entrada para cierto algoritmo de agrupamiento. (Ej. Número óptimo de grupos para un conjunto de datos)

La validación de grupo se clasifica en tres categorías: externa, interna y relativa. La investigación se centrará en las categorías externa e interna, debido a que la categoría relativa algunos autores la tratan como un caso particular de la validación interna.

En la validación externa se tiene un resultado ideal esperado para un conjunto de datos. Se realiza una comparación con el resultado obtenido que se deriva de un algoritmo de agrupamiento, y nos define si el algoritmo realizado es adecuado. En la validación externa existen diversos tipos de medidas de similitud, una de ellas es el recuento de pares (Goikoetxea, 2010), sobre la cual se han definido varios índices como es el caso de Rand (Rand, 1971) y Jaccard (Jaccard, 1908). Otra medida de similitud es la teoría de la información donde se encuentra la entropía y la pureza (Amigo, 2009).

A diferencia de la validación externa, los métodos de validación interna no utilizan la información de la clase para evaluar la calidad del agrupamiento. Su funcionamiento está relacionado con los conceptos de cohesión, la cual cuantifica la proximidad entre elementos de un mismo grupo. Mientras que el concepto de separación determina las relaciones entre los elementos de grupos diferentes. Este enfoque de validación se ha utilizado en problemas donde se hace necesario ajustar parámetros de entrada de diferentes algoritmos, como puede ser el número de grupos que mejor se ajusta para particionar un conjunto de datos (ej. cuando se desea determinar el mejor K). Existen un gran número de índices de validación interna (Bouguessa, 2006) (González, 2010) muchos de ellos definidos hace algunos años, sin embargo continúan empleándose en problemas reales.

Existen un conjunto de herramientas con técnicas y algoritmos de aprendizaje automático, que su uso se ha extendido por la comunidad internacional. Se destacan las herramientas Matlab y R, las cuales contienen índices de validación de agrupamiento. Otra de las herramientas más utilizada en el aprendizaje automático es WEKA. Desarrollada en el departamento de Ciencias

de la Computación de la Universidad de Waikato en Nueva Zelanda. Es una herramienta de código abierto, funciona en plataforma Windows y Linux, contiene una gran colección de algoritmos de aprendizaje automático, implementado en Java. Entre los principales tipos de problemas de aprendizaje que pueda hacer frente se encuentran la clasificación, regresión, agrupamiento y existe cierto apoyo a la minería de reglas de asociación (Jímenez, 2003). Esta herramienta crea un entorno favorable para el diseño de experimentos y la comparación de algoritmos para el desarrollo de investigaciones científicas.

En cuanto a los algoritmos de aprendizaje automático que soporta WEKA esencialmente cuenta con un conjunto de herramientas de pre-procesamiento de datos, alrededor de 76 algoritmos de clasificación/regresión para resolver problemas de clasificación supervisadas, 8 algoritmos de agrupamiento y 3 algoritmos de reglas de asociación para problemas no supervisados. Este balance entre los algoritmos supervisados y no supervisados en WEKA, denota que el mayor esfuerzo en el desarrollo de esta herramienta ha estado dirigido en los algoritmos de clasificación supervisados. Esta herramienta no tiene una implementación de los diferentes índices de validación de agrupamiento, que pueden ser utilizados para evaluar cuál de los algoritmos de agrupamiento funciona mejor sobre un conjunto de datos.

Por la situación problemática anteriormente descrita, se tiene como **problema a resolver**, ¿Cómo medir, haciendo uso de la herramienta WEKA, la calidad de las estructuras que se forman a partir de aplicar un algoritmo de agrupamiento?

Se define como **objeto de estudio** de la presente investigación el método de validación de agrupamiento.

De ahí que el **objetivo general** de esta investigación es desarrollar un paquete de métricas para la validación interna y externa de algoritmos de agrupamiento integrado en la herramienta WEKA.

Del objetivo general se derivan los siguientes **objetivos específicos**:

- Caracterizar el marco teórico-conceptual de los métodos de validación de agrupamiento interna y externa.
- Desarrollar un paquete de métricas de validación interna y externa para evaluar algoritmos de agrupamiento.
- Validar la solución implementada mediante casos de estudios y experimentación con varios métodos de agrupamiento existentes en WEKA.

El **campo de acción** viene dado por el análisis de los métodos de validación de agrupamiento interna y externa integrada a la herramienta WEKA.

Aporte práctico de la investigación.

Se desarrolla un paquete de métricas para la validación interna y externa de algoritmos de agrupamiento, que se integra a la herramienta WEKA, para la evaluación de la calidad de las particiones resultantes, apoyando así al desarrollo de experimentos e investigaciones.

Beneficio de la investigación.

Weka es una herramienta de código abierto, por lo que facilita su modificación y brinda facilidades para realizar experimentos e investigaciones en el ámbito del aprendizaje automático. Esta herramienta se emplea en la asignatura de Inteligencia Artificial 2 en la Universidad de Ciencias Informáticas, uno de los temas fundamentales que se trata en esta asignatura, son los algoritmos de agrupamiento. La herramienta WEKA no tiene implementadas, las métricas para evaluar el resultado del agrupamiento, por lo que debería incluirse como parte de las prácticas de laboratorio de la asignatura Inteligencia Artificial 2.

Durante la investigación se utilizarán **métodos científicos** para lograr caracterizar a fondo el objeto de estudio, así como garantizar el conocimiento del estado del arte, su evolución, relación con otros fenómenos y llevar a cabo la validación de los resultados de la investigación. Los métodos científicos a utilizar son:

Analítico-Sintético: A través de este método se analiza la bibliografía disponible para realizar una investigación lo más completa posible, y a su vez realizar una extracción de las características y elementos más importantes sobre los métodos de validación de grupo. Este método permite definir los principales conceptos, definiciones y otras soluciones ya existentes.

Histórico-Lógico: A través de este método se hace un estudio de la evolución de los métodos de validación de grupo, para así lograr un mejor entendimiento de estos enfoques y trabajar en su mejoramiento o actualización.

Métodos empíricos: Se utilizó como caso de estudio el algoritmo SimpleKMeans, realizando experimentos a diferentes bases de datos.

Organización de la tesis

Para exponer los resultados de la solución propuesta la tesis se estructuró de la siguiente manera:

- **Capítulo 1**. Fundamentos teóricos: Contiene la fundamentación teórica del tema a desarrollar. Se abordarán los conceptos asociados a los métodos de validación de grupo, así como una descripción de cada uno de estos enfoques.
- **Capítulo 2**. Análisis de la solución propuesta: Se hace referencia a las características que debe tener el sistema para su buen funcionamiento, haciendo constancia en las diferentes medidas de validación de grupo.
- **Capítulo 3**. Diseño de la solución propuesta: En este capítulo se realiza la propuesta de un Caso de Estudio, donde sobre un conjunto de datos, se evalúa el algoritmo de agrupamiento Simple-K-means contenido en la herramienta WEKA, y hacer uso de los métodos de validación de grupo tanto internos como externos, para evaluar dicho algoritmo de agrupamiento, para luego hacer una comparación de los resultados obtenidos por cada índice de validación de grupo.

CAPÍTULO 1.FUNDAMENTOS TEÓRICOS

Introducción

En el presente capítulo se brinda una visión general de los principales conceptos, asociados a los métodos de validación de grupo. Se describen los enfoques de validación externa e interna, utilizados para comprender y darle solución al trabajo de investigación. Se muestran además, algunos lenguajes y herramientas a utilizar, justificando en cada caso su utilización.

1.1 Definiciones y notaciones

Antes de entrar en una descripción más detallada de algunos elementos, se definen varios conceptos y notaciones que se emplean a lo largo del documento.

Definición 1: Una base de datos, $X = \{x_1, x_2, x_3, \dots, x_n\} \subseteq X \in \Re^J$ es un conjunto de N objetos o puntos representados como vectores de características en un espacio $J - dimensional$. Denotamos el valor del atributo correspondiente a la dimensión J del vector X_i como x_{ij} (Piper, 2003)

Definición 2: Datos continuos: Son los datos que pueden tomar cualquier valor, estos datos pueden tener valores decimales. Al estar muy cerca unos de otros, no se pueden estudiar de uno en uno y se agrupan en intervalos. Los valores que asume son números reales (Buuren, 2007).

Los algoritmos de agrupamiento reciben un conjunto de datos como entrada y mediante un proceso no supervisado, son capaces de descubrir a partir de la naturaleza de los datos, grupos de objetos que son similares entre sí.

Definición 3: Algoritmos de agrupamiento: Su objetivo es agrupar un conjunto de objetos, atendiendo a la naturaleza de los rasgos que caracterizan dichos objetos. Se emplea un modelo que permite agrupar los objetos representados en un conjunto de datos de modo que, aquellos que se encuentren bien cercanos entre sí, pertenecerán a un mismo grupo mientras que los objetos que estén bien separados deben quedar en grupos diferentes. De este modo se obtiene, como resultado del agrupamiento, una partición $C = \{C_1, C_2, \dots C_k\} \subseteq X,$ que cumple con las siguientes propiedades: $U_{C_k \in C}\, C_k = X,\ C_k \cap C_l = \emptyset\ \forall\ k \neq l.$ Esta definición es válida cuando nos referimos a agrupamiento duro.

Para realizar el agrupamiento de los objetos, es necesario determinar cuándo dos objetos del espacio son parecidos y cuándo no. Con este fin, se definen las funciones de similaridad o de disimilaridad, entre estas últimas se encuentran las métricas o distancias.

Definición 2: Un espacio métrico es un par (X, d) donde X es un conjunto no vacío y d es una función real definida en $X \times X$, llamada distancia o métrica, y que satisface los siguientes axiomas (Métricos, 2011):

$$d(x,y) \geq 0 \; \forall_{x,y} \in X, \;\; y \;\; d(x,y) = 0 \Leftrightarrow x = y. \qquad (1)$$

$$d(x,y) = d(y,x) \;\; \forall_{x,y} \in X \qquad (2)$$

$$d(x,z) \leq d(x,y) + d(y,z) \;\; \forall_{x,y,z} \in X \qquad (3)$$

En aplicaciones prácticas de algoritmos de agrupamiento, existen cuestiones que deben ser resueltas, entre ellas, la determinación del número de grupo y la evaluación de la calidad de un grupo. Existen algunos índices de validación de grupo que proporcionan una medida objetivo del resultado de un algoritmo de agrupamiento.

Definición 3: Un índice de validación de grupo es una función $F(P)$, que mide la bondad o calidad de una partición. Se define que a menor valor de la función F mejor será la partición.

1.2 Distancias o Métricas

Las medidas de distancias más tradicionales son aquellas que se aplican sobre dos instancias o ejemplos, tales que todos los atributos sean continuos. La función de distancia más utilizada es la euclidiana:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (4)$$

Otras funciones que podemos mencionar son:

Distancia de Mahalanobis:

$$d(x,y) = \; (\sum_{i=1}^{I} |x_i - y_i|^r)^{\frac{1}{r}}, r \; \geq 1 \qquad (5)$$

Distancia de Manhattan:

$$d_1(x,y) = \; \sum_{i=1}^{I} |x_i - y_i| \qquad (6)$$

La distancia de Mahalanobis, a diferencia de la distancia Euclídea, tiene en cuenta las correlaciones del conjunto de datos y no depende de la escala de las mediciones. En la literatura, se han propuesto diversas funciones para calcular la distancia entre objetos con atributos no numéricos, por ejemplo en (Stanfill, 1986), (Wilson, 2000) y (Olvera, 2005).

1.3 Algoritmos de Agrupamiento

El agrupamiento es la técnica principal de la clasificación no supervisada. Aunque existen diversas definiciones de agrupamiento, podríamos decir que el objetivo del agrupamiento es permitir la identificación de grupos. Donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros grupos, a partir de las características definidas para estos objetos (WUNSCH, 2009).

1.3.1 Aplicaciones del agrupamiento

El agrupamiento de datos se ha utilizado principalmente para tres propósitos fundamentales:

- Detectar la estructura subyacente en los datos: conocer mejor los datos, generar hipótesis, detectar anomalías e identificar las características principales.
- Clasificación natural: identificar el nivel de similitud de diferentes organismos.
- Compresión: método para organizar los datos y resumirlos a través de prototipos o representantes de grupo (Goikoetxea, 2010).

Siguiendo cualquiera de estos objetivos el agrupamiento se ha utilizado en numerosos campos científicos como: la Psicología (Estrada, 2013), Seguridad Informática (Gurrutxaga, 2008) (I. Perona, 2009), Telecomunicaciones (Oñate, 2012), Medicina (Almeida, 2014), Neurociencia (Sánchez, 2014). Una búsqueda del término agrupamiento en Google Schoolar arroja un récord de 2, 700,000 entradas en el mes de marzo del año 2015, este dato da una pista acerca de la gran extensión del agrupamiento.

1.3.2 Características de los algoritmos de agrupamiento

Las características deseables de la mayoría de los algoritmos de agrupamiento son las siguientes (Orallo, 2001):

Escalabilidad. La mayoría de los algoritmos de agrupamiento trabajan de manera apropiada con un número pequeño de observaciones (hasta 200 aproximadamente), mientras que se

necesita una gran escalabilidad para realizar agrupamiento en bases de datos con millones de observaciones.

Habilidad para trabajar con distintos tipos de atributos. Muchos algoritmos se han diseñado para trabajar sólo con datos numéricos, mientras que en ocasiones, es necesario trabajar con atributos asociados a tipos numéricos, binarios, discretos y alfanuméricos.

Descubrimiento de grupos con formas arbitrarias. La mayoría de los algoritmos de agrupamiento se basan en la distancia Euclidiana, lo que tiende a encontrar grupos, todos con forma (circular) y densidad similares. Es importante diseñar algoritmos que puedan establecer grupos de formas arbitrarias.

Habilidad para tratar con datos ruidosos. La mayoría de las base de datos (BD) contienen datos con comportamiento extraño, datos faltantes, desconocidos o erróneos. Algunos algoritmos de agrupamiento son sensibles a tales datos y pueden derivarlos a grupos de baja calidad.

Alta dimensionalidad. Una BD o DW (DataWarehouse) puede contener varias dimensiones o atributos, por lo que es bueno que un algoritmo de agrupamiento pueda trabajar de manera eficiente y correcta, no sólo en repositorios con pocos atributos, sino también en repositorios con un alto espacio dimensional, o gran cantidad de atributos.

Interpretación y uso. Los usuarios esperan que los resultados del agrupamiento sean comprensibles, fáciles de interpretar y de utilizar.

1.3.3 Tipos de algoritmos de agrupamiento

En la literatura existe una gran cantidad de técnicas de agrupamiento que varían de acuerdo a la arquitectura que utilizan. Los algoritmos de agrupamiento se clasifican en diferentes tipos (Jain, 1999):

Agrupamiento particional: Intenta descomponer directamente el conjunto de datos en un conjunto de grupos disjuntos. Específicamente, trata de determinar un número entero de particiones que optimicen una función de criterio determinada. La función de criterio puede enfatizar el local o la estructura global de los datos y su optimización es un procedimiento iterativo.

Agrupamiento jerárquico: En los métodos de agrupamiento jerárquicos se crea un árbol o dendrograma. El árbol no es un conjunto de grupos, sino más bien una jerarquía de varios

niveles, donde los grupos en un nivel se unen como grupos en el siguiente nivel. Esto le permite decidir el nivel o escala de agrupación que es el más apropiado para cada caso.

Agrupamiento basado en densidad: Este tipo de algoritmos se basa en la idea de que un grupo es una región del espacio densa y aislada. Por lo tanto, detectan los grupos como puntos en el espacio conectados por zonas de una densidad mínima establecida como umbral.

A continuación se describen detalladamente los tipos de algoritmos de agrupamiento.

Método Jerárquico

Los algoritmos jerárquicos utilizan la matriz de distancia como criterio de agrupamiento, estos algoritmos no necesitan saber el número de grupos que deben usar, pero si una condición de término.

En los métodos de agrupamiento jerárquico se crea un árbol o dendrograma. Los nodos hoja, representan la primera partición de un proceso aglomerativo (o la última de uno divisivo), mientras que los nodos internos representan la unión de varios grupos. El árbol es una jerarquía de varios niveles, donde los grupos en un nivel se unen como grupos en el siguiente nivel. Esto le permite decidir el nivel o escala de agrupación que es el más apropiado para cada caso (WUNSCH, 2009).

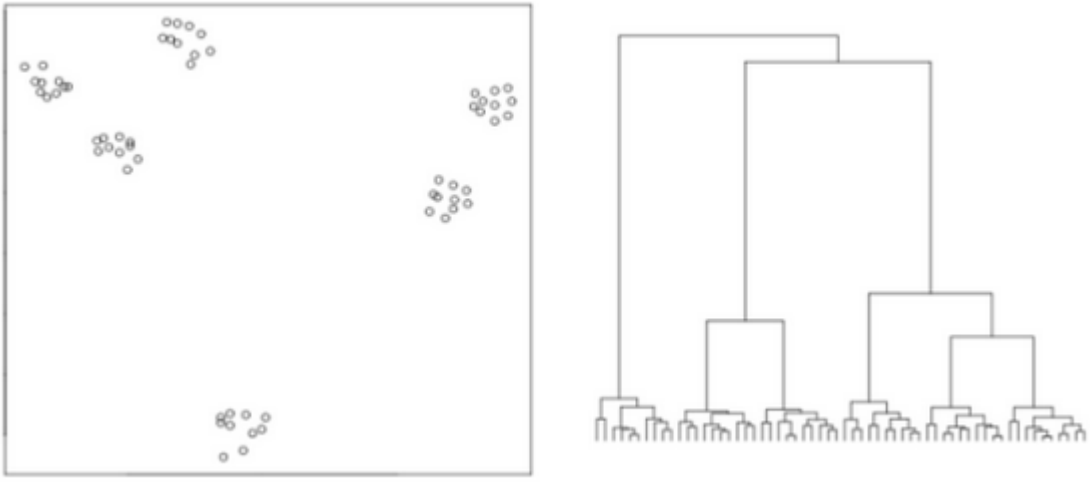

Figura 1 Una base de datos con estructura jerárquica y un posible dendrograma.

Existen dos tipos de algoritmos jerárquicos: aglomerativo y divisivo. El algoritmo aglomerativo funciona de abajo hacia arriba y mezcla grupos iterativamente. Este algoritmo comienza situando cada patrón u objeto por separado, y va mezclando estos grupos en grupos cada vez

mayores, hasta que los objetos están en un único grupo. Los grupos se van agrupando en función de la similitud existentes entre ellos, para ello se evalúa la distancia entre los distintos grupos y aquellos dos más cercanos se fusionan formando uno sólo, hasta que estén todos agrupados (HALKIDI, 2001).

Las estrategias jerárquicas aglomerativas más conocidas basadas en distancias son Single Link (SL) (Sibson, 1973), Average Link (AL) y Complete Link (Defays, 1977). En cada nivel de la jerarquía, se unen los dos grupos más cercanos. La siguiente figura ilustra estas estrategias.

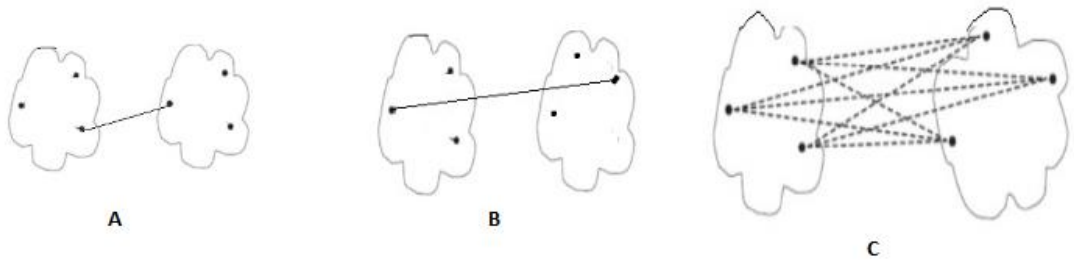

Figura 2 A- Single Link, B- Complete Link, C-Average Link

Algoritmos de densidad

Los algoritmos basados en densidad localizan zonas de alta densidad separadas por regiones de baja densidad. Entre los algoritmos que utilizan este enfoque se encuentran:

DBSCAN [1] es uno de los primeros algoritmos que utiliza este enfoque (Ester, 1996). DBSCAN presenta como ventajas que: descubre grupos de formas arbitrarias, trata el ruido y genera automáticamente el número de grupos. No es bueno para datos de alta dimensionalidad, con grupos de diferentes densidades y muy solapados.

Inicia seleccionando un punto t arbitrario, si t es un punto central, se empieza a construir un grupo alrededor de él, tratando de descubrir componentes denso-conectadas; si no, se visita otro objeto del conjunto de datos. Puntos centrales son aquellos tales que en su vecindad de radio, hay una cantidad de puntos mayor o igual que un umbral especificado. Este algoritmo busca grupos comprobando la vecindad de cada punto de la base de datos y va añadiendo puntos que son denso-alcanzables desde un punto central.

[1] Density Based Spatial Clustering of Aplications with Noise.

DENCLUE[2] este algoritmo logra buenos agrupamientos en bases de datos con puntos ruidosos (Hinneburg, 1998). Entre sus limitaciones esta la selección, de σ, que no es más que la determinación de la influencia de un punto a su vecindad, y la de ξ , el cual es el umbral de densidad.

OPTICS[3] se basa en la necesidad de introducir parámetros de entrada en casi todos los algoritmos de agrupamiento existentes, que en la mayoría de los casos son difíciles de determinar (OPTICS, 1999). En conjuntos de datos reales, no existe una manera de determinar estos parámetros globales, por lo que trata de resolver este problema basándose en el esquema del algoritmo DBSCAN, creando un ordenamiento de la base de datos para representar la estructura del agrupamiento basada en densidad.

Método Particional.

El agrupamiento particional es una división de objetos, en conjuntos no superpuestos, donde cada objeto es exactamente un subconjunto. El algoritmo particional genera una única partición de los datos. Generalmente este tipo de algoritmo es menos costoso.

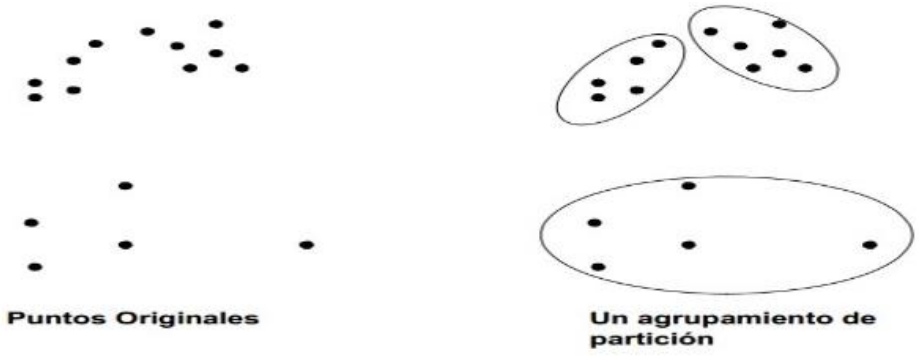

Figura 3 Ejemplo de agrupamiento particional.

Los algoritmos de agrupamiento varían entre sí por las reglas heurísticas que utilizan y el tipo de aplicación para el cual fueron diseñados. La mayoría de ellos se basa en el empleo sistemático de distancias entre vectores (objetos a agrupar), así como entre grupos que se van formando a lo largo del proceso de agrupamiento.

Entre los algoritmos más representativos que engloba el método particional se encuentra el K-means (WUNSCH, 2009). Pese a que el algoritmo K-means es conocido desde hace más de 5

[2] Density-based Clustering.
[3] Ordering Points to Identify the Clustering Structured.

décadas sigue siendo uno de los algoritmos de aprendizaje automático más utilizados y mejor valorados.

Dado un conjunto de datos y un valor de K, el algoritmo genera una partición de los datos con K grupos. El objetivo del algoritmo es minimizar el error cuadrático medio de todos los grupos. El error cuadrático de un grupo se define, como la media de los cuadrados de las diferencias de los objetos de un grupo, al centroide del grupo. El centroide de un grupo es el punto medio de todos los objetos del grupo. Este algoritmo también se conoce como algoritmo de las medias móviles porque en cada iteración se recalculan los centros de los agrupamientos, a esta función hay que darle como parámetro el número de agrupamientos (K) que debe encontrar (Raymond, 2013).

Características del K-means

- Cada grupo está asociado con un centroide (punto central).
- Cada punto es asignado al grupo más cercano al centroide.
- El número de grupos “K” debe ser especificado.
- El algoritmo básico es muy simple.

El proceso de funcionamiento del algoritmo es el siguiente:

Algoritmo 1: K-means
Seleccionar K casos como centroides iniciales repeat Asignar cada caso al grupo con centroide más cercano Recalcular los centroides de cada grupo until los centroides no se han modificado

Tabla 1 Algoritmo K-means

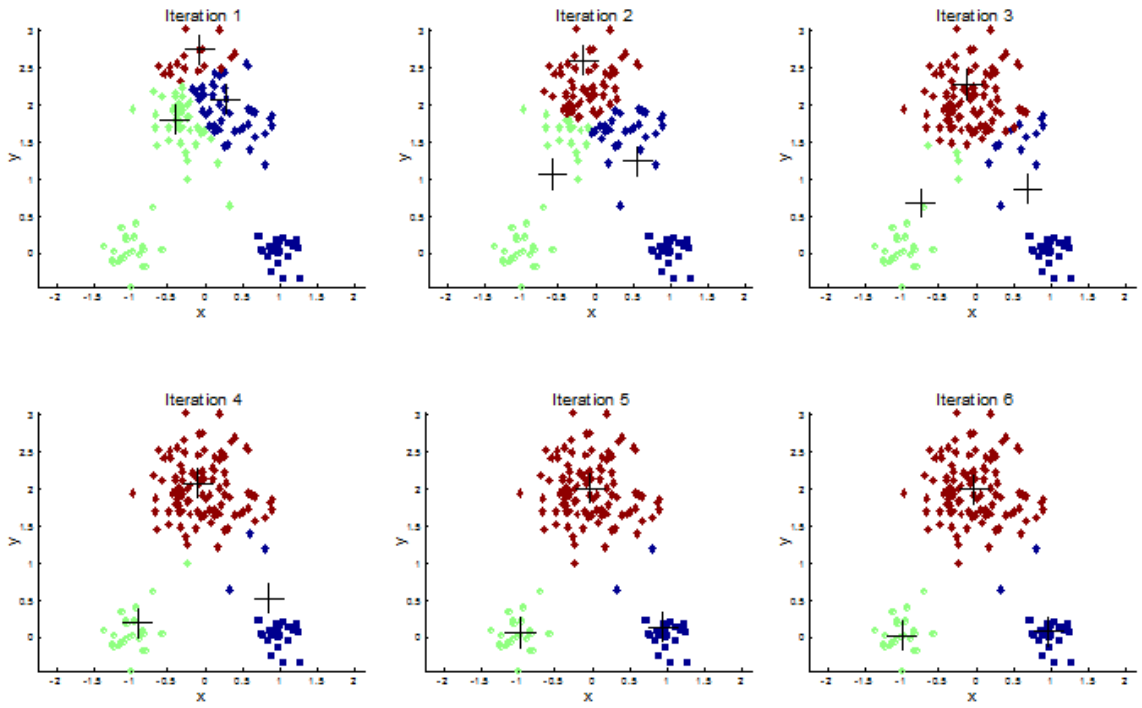

Figura 4 Ejemplo de aplicar el algoritmo K-Means.

Este algoritmo tiene como ventajas que: Converge eficientemente a un óptimo local. Además de ser adecuado en grupos compactos y bien separados. Sin embargo es inadecuado para descubrir grupos no convexos, de tamaño y densidad diferente.

1.4 Validación de agrupamiento

El objetivo de la validación de agrupamiento, es medir de forma objetiva la calidad del resultado obtenido por un algoritmo de agrupamiento. Como se ha descrito anteriormente, este resultado puede ser una partición o una jerarquía de particiones. Existen tres categorías de índices de validación: índices externos, internos y relativos, estos a su vez no dependen del tipo del algoritmo de agrupamiento, ya que se espera que sean objetivos y no tengan ninguna preferencia sobre algún algoritmo en particular. Dado que la práctica generalizada, es la de analizar las particiones de manera comparativa, y algunos autores han situado a la validación relativa como un caso particular de la validación interna, parece más adecuado limitar la validación a dos categorías: externa e interna.

1.4.1 Validación Externa

La validación externa asume que es conocida de antemano la partición ideal a las que se ajustan los datos, este enfoque se basa en determinar cuán cercanas se encuentra la partición obtenida producto de un algoritmo de agrupamiento y la partición ideal. Convirtiendo a los

índices de validación externa en medidas de similitud, las cuales dependen de la información que contienen. Entre estas medidas se encuentran el recuento de pares y teoría de la información.

El recuento de pares verifica que cada par de objetos pertenezca o no a un mismo grupo en ambas particiones, existen diversas medidas de similitud basadas en el recuento de pares como: Rand (Rand, 1971), Jaccard (Ramakrishna, 2005), Fowlkes-Mallows (Fowlkes y Mallows, 1983) y Adjusted Rand (W. J Matthijs, 2008a). Mientras que la teoría de la información se ocupa de la medición de la información y de la representación de la misma, entre estas medidas encontramos la entropía conjunta (Goikoetxea, 2010), la entropía y la pureza (E. Rendón, 2011).

1.4.2 Validación Interna

Existe un gran número de índices para la validación interna. La mayoría se basan en los conceptos de cohesión y separación. En el caso de Cohesión, el miembro de cada clúster debe ser lo más cercano posible a los otros miembros del mismo clúster. Mientras que en separación, los clúster deben estar ampliamente separados entre ellos. Existen varios enfoques para medir esta distancia entre clúster: distancia entre el miembro más cercano, distancia entre los miembros más distantes o la distancia entre los centroides. Existen formas de medir estos conceptos: sin usar centroides y usando centroides (Berzal, 2012).

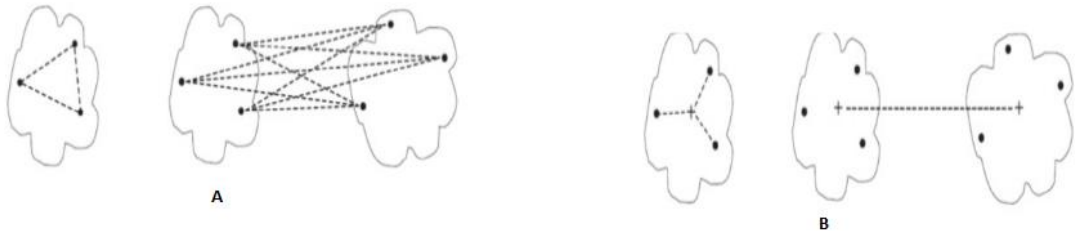

Figura 5 A-Sin usar centroide, B-Usando centroide

Kim y Ramakrishna (Ramakrishna, 2005) definen estas formas como índices basados en radio e índices basados en suma. A continuación se presentan varios índices de validación interna, citados por varios autores:

Índice de validación interna	Tipo
Calinski-Harabasz	Radio

C-Index	Suma
DUNN	Suma
Xie-Beni	Radio
Davies Bouldin	Radio

Tabla 2 Índices de validación interna con el grupo al que pertenecen

1.5 Análisis del agrupamiento

El análisis del agrupamiento es un proceso importante para interpretar datos. Las técnicas de agrupamiento son la herramienta fundamental empleada para la exploración de datos, específicamente cuando se necesita saber la estructura interna de los datos, sin tener información a priori disponible, pues se espera que los algoritmos de agrupamiento produzcan particiones que reflejen la estructura interna de los datos e identifiquen las clases naturales y jerarquías presentes en los mismos. En diversas técnicas de agrupamiento, el número K de grupos a construir es un parámetro de entrada del algoritmo de agrupamiento, lo que supone un conocimiento a priori de ese dato. Sin embargo, en la práctica, podemos desconocerlo. Muchos métodos e indicadores bajo el nombre de validación de grupo permiten evaluar los resultados del agrupamiento de esta manera (González, 2010).

La noción de validación de grupo se refiere a conceptos o métodos para la evaluación cuantitativa y objetiva del resultado de un algoritmo de agrupamiento. En la literatura consultada con relación a esta temática, dichos conceptos y métodos aparecen publicados en (Rand, 1971), (Rousseeuw, 1987), (Jain, 1986), (Levine, 2001), (Halkidi, 2002a), (Halkidi, 2002b), (N. Bolshakova, 2003), (Bouguessa, 2006) entre otros.

Un índice de validación proporciona una medida objetivo del resultado de un agrupamiento, y su valor óptimo se usa frecuentemente para indicar la mejor selección posible de los valores de los parámetros en el algoritmo de agrupamiento, por ejemplo, el número de grupos. Numerosos trabajos sugieren índices directos o indirectos para evaluar: agrupamientos duros (Dubes, 1988), agrupamientos probabilísticos (Duda, 1973) y agrupamientos difusos (Bouguessa, 2006). Algunos índices de validación de grupo evalúan los resultados de un algoritmo de agrupamiento

empleando dos criterios de medida: compacidad y separación. La compacidad se basa en la suposición de que los elementos de un mismo grupo deben estar tan cerca como sea posible, mientras que el criterio de separación considera que los grupos deben estar bien separados. La mayoría de los índices de validación son específicos del tipo de validación. Típicamente la validación externa se reduce a la comparación de dos particiones: la evaluada y la correcta. Por ello, los índices de validación externa suelen ser medidas de similitud de particiones. Los índices de validación interna, en cambio, suelen estimar la cohesión y la separación de los grupos.

1.6 Sistemas para el análisis de datos

Existen disímiles herramientas que permiten realizar experimentos evaluando algoritmos de agrupamiento. Entre ellas se encuentran:

Matlab:

Matlab es una herramienta de software matemático con lenguaje de programación propio. Está disponible para las plataformas Unix, Windows y Mac OS X.

Entre sus prestaciones básicas se hallan: la manipulación de matrices, la representación de datos y funciones, la implementación de algoritmos, la creación de interfaces de usuario y la comunicación con programas en otros lenguajes y con otros dispositivos hardware.

La función de la validez proporciona las medidas de la validez del agrupamiento para cada partición. Es útil cuando el número de grupos es a priori desconocido. La partición óptima se puede determinar por el punto de los extremos de los índices de la validación, en la dependencia del número de grupos. Los índices calculados son: Repartir el coeficiente (PC), la entropía de la clasificación (CE), el índice de la partición (SC), el índice de separación (S), el índice de Xie-Beni (XB), el índice de Dunn (DI) y el índice de Dunn de la alternativa (DII) (Programming, 2010).

R:

R es un lenguaje interpretado (scripting) y un conjunto de librerías creado por John Chambers en los Laboratorios Bell. Provee al usuario gran variedad de técnicas estadísticas (modelos lineales y no lineales, test estadísticos, análisis de series de tiempo, grupo, etc.) y gráficas. Puede ser ejecutado y compilado en plataformas como UNIX, Windows y MacOS y se distribuye sin costo y bajo licencia GPL. Permite la construcción de paquetes que favorecen la reutilización y combinación de los componentes creados por otros usuarios, y los predefinidos en el kernel

base-R. R puede integrarse con distintas bases de datos y existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Python.

WEKA:

WEKA (Waikato Environment for Knowledge Analysis) es un entorno de trabajo desarrollado por la universidad de Waikato en Australia, para el análisis de datos. Se distribuye como software libre desarrollado en Java y está constituido por una serie de paquetes de código abierto que permiten aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático.

El resultado de aplicar el algoritmo de clasificación se efectúa comparando la clase predicha con la clase real de las instancias. Esta evaluación puede realizarse de diferentes modos (Bouckaert, 2010):

Conjunto de entrenamiento: Con esta opción se entrena el método con todos los datos disponibles y posteriormente realiza la evaluación sobre los mismos.

Adicionar un conjunto de datos de prueba: Esta opción permite cargar un conjunto nuevo de datos. Sobre cada dato se realizará una predicción de clase para contar los errores.

Validación cruzada: Se realiza la evaluación mediante la técnica de validación cruzada, consiste en: dado un número n se divide los datos en n partes y, por cada parte, se construye el clasificador con las $n-1$ partes restantes y se prueba con esa. Así por cada una de las n particiones.

Dividir los datos según el porcentaje: esta opción divide los datos en dos grupos, de acuerdo con el porcentaje indicado (%). El valor indicado es el porcentaje de instancias para construir el nuevo clasificador, que a continuación es evaluado sobre los datos restantes. Cuando el número de instancias es suficientemente elevado, esta opción es suficiente para estimar con precisión las prestaciones del clasificador en el dominio.

1.7 Herramientas y tecnologías utilizadas

El desarrollo y el avance alcanzado por la informática han brindado la posibilidad de impulsar el perfeccionamiento de las metodologías, herramientas, lenguajes y tecnologías para la construcción y desarrollo de aplicaciones, que posibilitan un mejor manejo y procesamiento de la información.

1.7.1 Herramientas

Actualmente se consideran a las Herramientas de Desarrollo de Software (HDS), como herramientas basadas en computadoras que asisten el proceso de ciclo de vida de una aplicación, consolidadas en la literatura en la forma de Ingeniería de software asistida por computadora (CASE, por sus siglas en inglés) . Entre las HDS que automatizan metodologías de software se escogió a Visual Paradigm.

1.7.1.1 Herramienta de modelado

Visual Paradigm for UML (CE) 8.0

Esta herramienta CASE[4] es empleada para el modelado visual; simplifica la elaboración de artefactos mediante el uso del lenguaje de modelado UML. Soporta una amplia gestión de casos de uso y diseño de bases de datos relacionales. Esta herramienta proporciona eficaces medidas de análisis y diseño de sistemas. Es posible modelar cualquier tipo de diagrama de forma rápida ya que posee una serie de módulos y productos que permiten la modelación (VisualParadigm).

1.7.2 Entorno de desarrollo integrado

Un entorno de desarrollo integrado, llamado también IDE[5] (por sus siglas en inglés), es un programa informático formado por un conjunto de herramientas de programación, se basa en un editor de código, un compilador, un depurador y un constructor de interfaz gráfica.

NetBeans

EL entorno de desarrollo integrado (IDE) NetBeans en su versión 7.4, es de código abierto, permite desarrollar aplicaciones de escritorio, móviles y aplicaciones web sobre el lenguaje Java, y es compatible con la herramienta WEKA. Es soportado por diferentes aplicaciones en varios idiomas, incluyendo HTML5, PHP y C ++. El IDE se ejecuta en Windows, Linux, Mac OS X y otros sistemas basados en UNIX. Su código está bien organizado, lo que lo hace más entendible a la vista de los nuevos desarrolladores (Netbeans).

[4] Computer Aided Software Engineering, Ingeniería de Software Asistida por Computadoras.
[5] IDE: Integrated Development Environment

1.7.3 Lenguajes

Un lenguaje en el contexto informático es un idioma artificial, diseñado para expresar procesos que pueden ser llevados a cabo por las máquinas, lo que proporciona un vocabulario para lograr una comunicación entre el equipo de desarrollo y los ordenadores.

1.7.3.1 Lenguaje de programación

Un lenguaje de programación es un lenguaje diseñado para detallar el conjunto de acciones sucesivas que un equipo debe realizar. Son herramientas que nos permiten crear software.

Java

Java es un lenguaje multiplataforma, que incorpora en el propio lenguaje, muchos aspectos que en cualquier otro lenguaje son extensiones de propiedad de empresas de software o fabricantes de ordenadores (ejecución remota, componentes, seguridad, acceso a bases de datos, etc.). La herramienta WEKA está desarrollada sobre este lenguaje. Muchos expertos opinan que Java es el lenguaje ideal para aprender la informática moderna, porque incorpora todos estos conceptos de un modo estándar, mucho más sencillo y claro que con las citadas extensiones de otros lenguajes (Otero, 2007).

1.7.3.2 Lenguaje UML

El lenguaje para modelamiento unificado (UML[6]), es un lenguaje para la especificación, visualización, construcción y documentación de los artefactos de un proceso de sistema intensivo. Puede ser aplicado a cualquier elemento de modelado, incluyendo clases, paquetes, relaciones de herencia, etc (Rodriguez, 1999).

1.7.4 Metodología estadística de múltiples clasificadores

La evaluación estadística de resultados experimentales, es una parte esencial para la validación de los métodos de aprendizaje automático. Sin embargo, las pruebas usadas son ingenuas y sin verificar. Mientras que para comparar dos clasificadores sobre una colección de datos existen procedimientos desde hace más de una década, para la comparación de varios clasificadores sobre múltiples colecciones de datos se usan soluciones parciales e insatisfactorias (Demsar, 2006).

[6] UML: Unifed Modeling Languaje

Dietterich (Dietterich, 1998) examina cinco pruebas estadísticas y recomienda el uso del 5x2cv t-test que supera el problema de la variación subestimada y el error de tipo I. Un t-test o prueba t de Student es donde la medida cuantitativa utilizada tiene una distribución t de Student si se acepta la hipótesis nula. Cuando el algoritmo se aplica múltiples veces y no es apropiado utilizarlo, Dietterich propone un procedimiento de McNemar sobre una matriz de clasificación errónea tan potente como el 5x2cv y advierte en contra de los t-test luego de muestreos aleatorios repetitivos y validación cruzada. Alpaydin (Alpaydin, 1999) mejoró el 5x2cv t-test al construir un 5x2cv f-test más robusto con un menor error de tipo I y mayor potencia. Un f-test o prueba f (nombrada f en honor a Ronald Fisher) es donde la medida cuantitativa utilizada sigue una distribución F si no puede ser rechazada la hipótesis nula. Según Bouckaert (Bouckaert, 2003) los grados teóricos de libertad son incorrectos ya que existen dependencias entre los experimentos. También plantea que deberían usarse en su lugar los valores hallados empíricamente.

Cuando se dispone de conjuntos de datos independientes es válido aplicar un t-test. Si en la práctica no se disponen de k conjuntos independientes, entonces se extraen repetidamente de la misma colección de datos conjuntos de entrenamiento y prueba, a esto se le llama t-test remuestreado (A. González). Este t-test re muestreado corregido propuesto por (Nadeau, 2000) ajusta la varianza basada en la superposición entre subconjuntos de ejemplos. Otros autores que escribieron sobre el tema fueron (Bouckaert, 2004) y (Bengio, 2004). La prueba ANOVA como una de las posibles soluciones es mencionada por (Salzberg, 1997), para luego describir el test binomial apoyándose en la corrección de Bonferroni para múltiples comparaciones. Esto tiene como desventaja que el test binomial no posee la potencia que brindan las mejores pruebas no paramétricas, además de ser excesivamente radical la corrección de Bonferroni.

Para la comparación entre múltiples modelos en solo una colección de datos, (Vazquez, 2001) (Pizarro, 2002) utilizan el ANOVA y el test de Friedman.
El primero en utilizar un test no paramétrico para la comparación de clasificadores en las tareas de recuperación de información y asesoramiento sobre su relevancia fue Hull (Hull, 1994), (Bradzil, 2000) al no priorizar la selección del algoritmo más óptimo utilizó rangos promedios para compararlos, lo cual no prueba estadísticamente la significancia de las diferencias entre ellos. Muchas publicaciones incluyen, en aras de determinar el método más óptimo o comparar su desempeño, algún procedimiento estadístico.

Existe una oposición entre las pruebas usadas para evaluar la diferencia entre dos clasificadores en una colección de datos y sobre múltiples clasificadores sobre varias. Cuando se está probando en una colección de datos, usualmente se computa el rendimiento promedio y

su varianza sobre datos de entrenamiento y de prueba repetitivamente sobre muestras aleatorias. Ya que estas muestras estén probablemente relacionadas, es necesario diseñar procedimientos y pruebas estadísticas con cautela para evitar complicaciones con estimaciones de la varianza parcial. El problema de comparar clasificadores en una colección de datos haciendo uso de pruebas estadísticas correctas no está relacionado a la comparación en múltiples colecciones de datos, ya que se debe resolver el primero para poder atacar el segundo. Al ejecutar los algoritmos en múltiples colecciones de datos genera una muestra de medidas independientes, las cuales son más simples que las comparaciones en una solo colección de datos (Demsar, 2006).

1.8 Conclusiones del capítulo

En este capítulo se identificaron dos enfoques: índices de validación externa e interna. En el caso de la validación externa se identifican cuatro índices fundamentales, como son entropía, pureza, Rand y Jaccard coefficient, a partir de las métricas analizadas. Por otra parte, en la validación interna se utiliza la distancia euclidiana para medir la similitud entre objetos, teniendo en cuenta los siguientes índices: Calinski-Harabasz, C-Index, Dunn, Davies Bouldin y Xie-Beni para la implementación de los índices internos. Se propone un paquete de métricas, utilizando como medidas de validación los índices internos y externos expresados anteriormente. Este paquete se integrará a la herramienta WEKA, debido a que es una herramienta de experimentación y de fácil integración con aplicaciones empresariales. Para determinar si existen diferencias entre los algoritmos propuestos se utilizará el test de Friedman.

CAPÍTULO 2. DESCRIPCIÓN DE LA SOLUCIÓN PROPUESTA

Introducción

En el presente capítulo se propone un paquete de métricas, que contribuyan a validar las particiones resultantes de aplicar los algoritmos de agrupamiento, a través de la herramienta WEKA, describiendo sus índices y diagrama de clases.

2.1 Propuesta de solución

La solución propuesta tiene en cuenta los métodos de validación externa e interna para evaluar la partición obtenida como resultado de un algoritmo de agrupamiento. Estos métodos conformarán un paquete de métricas con el objetivo de validar a través de la herramienta WEKA los algoritmos de agrupamiento. Esta herramienta servirá de apoyo para poder determinar si el algoritmo de agrupamiento realizado a un conjunto de datos es apropiado.

2.2 Índices de validación externa

La validación externa se basa en determinar cuán cercanas se encuentra la partición obtenida producto de un algoritmo de agrupamiento y la partición ideal. Como parte de la solución para dar respuesta a un problema de agrupamiento, se proponen los índices basados en teoría de la información y los índices basados en el recuento de pares, para la validación de grupo en la implementación de este enfoque.

2.3 Índices basados en teoría de la información

Según la teoría de la información, el nivel de información de una fuente se puede medir según la entropía y la pureza de la misma. La entropía de un grupo refleja cómo los miembros de un conjunto de datos X, se distribuyen por cada uno de los grupos. El aumento de la entropía depende exclusivamente del número total de elementos del grupo (Amigo, 2009). Esta se calcula como el promedio pesado de todos los grupos:

$$-\sum_j \frac{n_j}{n} \sum_i P(i,j) \times log_2 P(i,j) \qquad (7)$$

Dónde:

$P(i,j)$: La probabilidad de encontrar un elemento de la categoría i en el grupo j.

n_j: El número de elementos en el grupo j.

n: El número total de elementos de la distribución.

Para la implementación del índice basado en la Entropía se siguen los pasos descritos en el siguiente algoritmo:

Algoritmo 2: Entropía
Entrada: partición Salida: Índice de validación de grupo.
For k= 1 to Cantidad de clases do $E = -\sum_j \frac{n_j}{n} \sum_i P(i,j) \times log_2 P(i,j)$ Calcular la Entropía general

Tabla 3 Algoritmo de la Entropía

2.4 Pureza(P)

El índice de pureza es muy similar a la entropía, su valor global de la agrupación resultante, se obtiene como una suma individual de la pureza de los grupos y está dado por:

$$P = \sum_{j=1}^{M} \frac{n_j}{n} P_j \qquad (8)$$

Dónde:

$$P_{j=\frac{1}{n_j}} max_i \left(n_j^i\right) \qquad (9)$$

Siendo:

n_j: El tamaño del grupo j.

n_j^i: El número de objetos en la clase i que se encuentran en el grupo j.

M: El número de grupo.

n: Total de objetos.

Algoritmo 3: Pureza
Entrada: partición Salida: Índice de validación de grupo
For k= 1 to Numero de grupos do $P_{j=\frac{1}{n_j}} max_i(n_j^i)$ Calcular la Pureza general

Tabla 4 Algoritmo de la Pureza

Para lograr una implementación que ocupe el espacio de memoria adecuado de este índice, fue necesario un arreglo bidimensional, que forma una matriz llamada Contingency_Matix de dimensión NxM, siendo N la cantidad de grupos y M la cantidad de clases, donde se almacena la relación entre los elementos de un grupo y su asignación en dependencia de la información de la clase.

2.5 Índices basados en recuento de pares

Otra medida, para definir métricas de evaluación para la agrupación es el recuento de pares, la cual se basa en la comparación de dos particiones, de las cuales se recuentan las distintas situaciones, en las que se encuentran los posibles pares pertenecientes a cada una de las particiones (Amigo, 2009).

Estas medidas verifican que cada par de objetos X_i, X_j pertenezcan o no a un mismo grupo en ambas particiones. Finalmente las medidas de validación externas se modelan a partir de:

a: la cantidad de pares de objetos que se encuentran en el mismo grupo en ambas particiones.
b y c: en el mismo grupo en una partición y grupos diferentes en la otra.
d: en grupos diferentes en ambas particiones.

Para la programación de este índice fue necesario implementar las medidas de Jaccard y Rand (Rand, 1971). Además se ha separado el cálculo de los parámetros a, b, c, d, en el método ComputedExternalParameter, de modo que la incorporación de un nuevo índice basado en el enfoque del recuento de pares reutiliza estos parámetros. Los índices implementados se determinan de la siguiente manera:

Índices	Fórmula
Rand índex	$R = \frac{a+d}{M}$
Jaccard coefficient	$J = \frac{a}{a+b+c}$

Tabla 5 Índices Rand y Jaccard con sus respectivas fórmulas

Para la implementación del índice basado en el recuento de pares se siguen los pasos descritos en el siguiente algoritmo.

Algoritmo 4: Basado en recuento de pares
Entrada: partición Salida: Índice de validación de grupo
Determinar los valores a, b, c, d a través del método (ComputedExternalParameter) Calcular Rand índex Calcular Jaccard coefficient

Tabla 6 Algoritmo basado en recuento de pares

2.6 Índices de validación interna

Los índices de validación internos no utilizan la información de la clase para evaluar la calidad del agrupamiento, dado que se concentran en el trabajo sobre una sola estructura de agrupamiento. Estos se dividen en dos grandes grupos los basados en suma y los basados en radio (centroide) como se mencionaba anteriormente. Se propone para la implementación de este enfoque a los índices: Davies-Bouldin, Dunn, Calinski-Harabasz, Xie-Beni y C-Index (Calinski, 1974), (Bouldin, 1979), (Rousseeuw, 1987), (Goikoetxea, 2010). La anexión de nuevos índices de validación de grupos resulta muy sencilla, pues existen funcionalidades en la implementación que se han definido, que pueden ser reutilizadas para la construcción de los mismos. A continuación se mencionan estas funcionalidades:

Funcionalidades	Expresión
Mínima distancia entre dos objetos de un mismo grupo.	$d = min_{x_i, x_{j \in C_k}} d(x_i, x_j)$

Mínima distancia entre dos objetos de grupos diferentes	$d = min_{x_i \in C_i, x_j \in C_k} d(x_i, x_j)$
Máxima distancia entre dos objetos de un mismo grupo	$d = max_{x_i, x_{j \in C_k}} d(x_i, x_j)$
Máxima distancia entre dos objetos de grupos diferentes	$d = max_{x_i \in C_i, x_j \in C_k} d(x_i, x_j)$
Mínima distancia entre dos centroides.	$d = min\, d(\bar{C}_i, \bar{C}_j)$
Máxima distancia entre dos centroides.	$d = max\, d(\bar{C}_i, \bar{C}_j)$
Distancia de un objeto a su centroide.	$d = d_{x_i \in C_i}(x_i, \bar{C})$

Tabla 7 Funciones de distancia

2.6.1 Índice Davies-Bouldin(DB)

Este índice se emplea para deducir la idoneidad de las particiones de datos. (Bouldin, 1979) utiliza como medida de compacidad de un grupo, la media de las distancias de sus puntos a su centroide, mientras que como medida de separabilidad, utiliza la distancia euclidiana, entre los grupos y su centroide. Es importante definir que el valor del índice obtenido no depende del número de grupos analizados. Se calcula a través de la expresión matemática:

$$DB(P) = \frac{1}{P}\sum_{C_k \in P, C_i \in P, C_k \neq C_i} max\left\{\frac{S(C_k)+S(C_i)}{d(\bar{C_k}, \bar{C_i})^2}\right\} \qquad (10)$$

Dónde:

$$S(C) = \sum_{x \in C} d(x, \bar{C})^2 \qquad (11)$$

Este índice es pequeño cuando los grupos son compactos y están lejos el uno del otro, al minimizar este índice, se logra una mejor partición. Para la implementación de este índice se siguen los pasos descritos en el siguiente algoritmo.

Algoritmo 5: Davies-Bouldin
Entrada: partición Salida: Índice de validación de grupo
Determinar la distancia que existe de cada objeto a su centroide Determinar la distancia que existe entre los centroides de la partición For k= 1 to Número de instancias do $\mathrm{S(C)} = \sum_{\mathrm{x \in C}} \mathrm{d(x, \bar{C})^2}$ Calcular Davies Bouldin

Tabla 8 Algoritmo Davies-Bouldin

2.6.2 Índice Dunn(D)

El índice Dunn corresponde al radio de la distancia más pequeña, entre las observaciones de diferentes grupos y la distancia inter-grupo más grande. Una de las desventajas del uso de este índice es el costo computacional, además del número de grupos y la dimensionalidad de los datos de aumento. Se calcula de la siguiente manera:

$$Dunn(P) = \frac{inter_{Dunn}}{intra_{Dunn}} \qquad (12)$$

Dónde:

$$inter_{Dunn}(P) = min_{C_K \in P}\left\{min_{C_i \in P, K \neq 1}\{\delta(C_K, C_i)\}\right\} \qquad (13)$$

$$\delta(C_K, C_i) = min_{X_i \in C_K}\left\{min_{X_j \in C_i}\left\{d\left(X_i, X_j\right)^2\right\}\right\} \qquad (14)$$

$$intra_{Dunn} = max_{C \in P}\left\{max_{X_i, X_j \in C}\left\{d\left(X_i, X_j\right)^2\right\}\right\} \qquad (15)$$

Para la implementación del índice Dunn se siguen los pasos descritos en el siguiente algoritmo:

Algoritmo 6: Dunn
Entrada: partición Salida: Índice de validación de grupo
Determinar la mínima distancia entre objetos de grupos diferentes Determinar la máxima distancia entre objetos que pertenezcan a un mismo grupo For k= 1 to Número de instancias do $inter_{Dunn}(P) = min_{C_K \in P}\left\{min_{C_i \in P, K \neq 1}\{\delta(C_K, C_i)\}\right\}$ $intra_{Dunn} = max_{C \in P}\left\{max_{X_i, X_j \in C}\left\{d\left(X_i, X_j\right)^2\right\}\right\}$ Calcular Dunn

Tabla 9 Algoritmo Dunn

2.6.3 Índice Calinski-Harabasz(CH)

Este índice está determinado por la razón entre la dispersión interior de los grupos y la dispersión entre los grupos. La función CH está definida de la siguiente forma:

$$CH(P) = \frac{SSB/(M-1)}{SSW/(N-M)} \qquad (16)$$

Dónde:

$$SSW = \frac{1}{N}\sum_{i=1}^{N} d(x_i, \bar{C}_i)^2 \qquad (17)$$

$$SSB = \frac{2}{M(M-1)}\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M} d\left(C_i, C_j\right)^2 \qquad (18)$$

Siendo N el número de objetos de la base de datos y M el número de grupos

Algoritmo 7: Calinski-Harabasz
Entrada: partición Salida: Índice de validación de grupo
Determinar la distancia entre los objetos y el centroide que le corresponde. Determinar la distancia de un centroide a otro centroide. For k= 1 to Número de instancias do $\mathrm{SSW} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{d}(\mathrm{x_i}, \bar{\mathrm{C}}_\mathrm{i})^2$ $\mathrm{SSB} = \frac{2}{\mathrm{M(M-1)}}\sum_{\mathrm{i=1}}^{\mathrm{M}}\sum_{\mathrm{j=1,j\neq i}}^{\mathrm{M}} \mathrm{d}\left(\mathrm{C_i}, \mathrm{C_j}\right)^2$ Calcular Calinski-Harabasz

Tabla 10 Algoritmo Calinski-Harabasz

2.6.4 Índice C-Index

Para el cálculo de este índice es necesario conocer que la suma de las distancias entre todos los pares de objetos en el mismo grupo es:

$$S(P) = \sum_{C \in P} \sum_{X_j, X_j \in C} d(X_i, X_j)^2 \qquad (19)$$

Sabemos que (n_w) es el número de pares de objetos. Por tanto:

(S_{min}): Es la suma de los (n_w) de menor distancia sobre la base de datos.

(S_{max}): Es la suma de los (n_w) de mayor distancia sobre la base de datos.

Por lo que queda definido de esta manera:

$$CI = \frac{S(P) - S_{min}(P)}{S_{max}(P) - S_{min}(P)} \qquad (20)$$

Algoritmo 8: C-Index
Entrada: partición Salida: Índice de validación de grupo
Determinar la suma de las distancias entre todos los pares de objetos en el mismo grupo. Determinar la mínima distancia entre dos objetos del mismo grupo de la base de datos. Determinar la máxima distancia entre dos objetos del mismo grupo de la base de datos. For k= 1 to Número de instancias do $S(P) = \sum_{C \in P} \sum_{X_j, X_j \in C} d(X_i, X_j)^2$ Calcular C-Index

Tabla 11 Algoritmo C-Index

2.6.5 Índice Xie-Beni(XB)

El índice de Xie-Beni se define como el cociente entre el error cuadrático medio y el mínimo de las distancias al cuadrado, entre los puntos en los grupos. El objetivo fundamental de este índice es minimizar. Se calcula a través de la fórmula:

$$XB = \frac{1}{N}\frac{WGSS}{\delta(C_k, C_j)} \qquad (21)$$

Dónde:

$$WGSS = \sum_{i=1}^{N}\sum_{j=1}^{M} d\left(x_i, \overline{C_j}\right)^2 \qquad (22)$$

$$\delta(\mathrm{C_k}, \mathrm{C_j}) = \min_{\mathrm{h}\in\mathrm{k},\mathrm{r}\in\mathrm{j}} \mathrm{d}(\mathrm{x_h}, \mathrm{x_r})^2 \qquad (23)$$

Algoritmo 10: Xie-Beni
Entrada: partición Salida: Índice de validación de grupo
Determinar la distancia entre los objetos y el centroide al que pertenecen. Determinar la mínima distancia entre objetos de grupos diferentes. For k= 1 to Número de instancias do $\mathrm{WGSS} = \sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{M}} \mathrm{d}\left(\mathrm{x_i}, \overline{\mathrm{C_j}}\right)^2$ $\delta(\mathrm{C_k}, \mathrm{C_j}) = \min_{\mathrm{h}\in\mathrm{k},\mathrm{r}\in\mathrm{j}} \mathrm{d}(\mathrm{x_h}, \mathrm{x_r})^2$ Calcular Xie-Beni

Tabla 12 Algoritmo Xie-Beni

2.7 Diseño del sistema

2.7.1 Diagrama de paquete

El siguiente diagrama representa el diseño de paquete propuesto para la integración de las métricas de evaluación de los algoritmos de agrupamiento a la herramienta Weka.

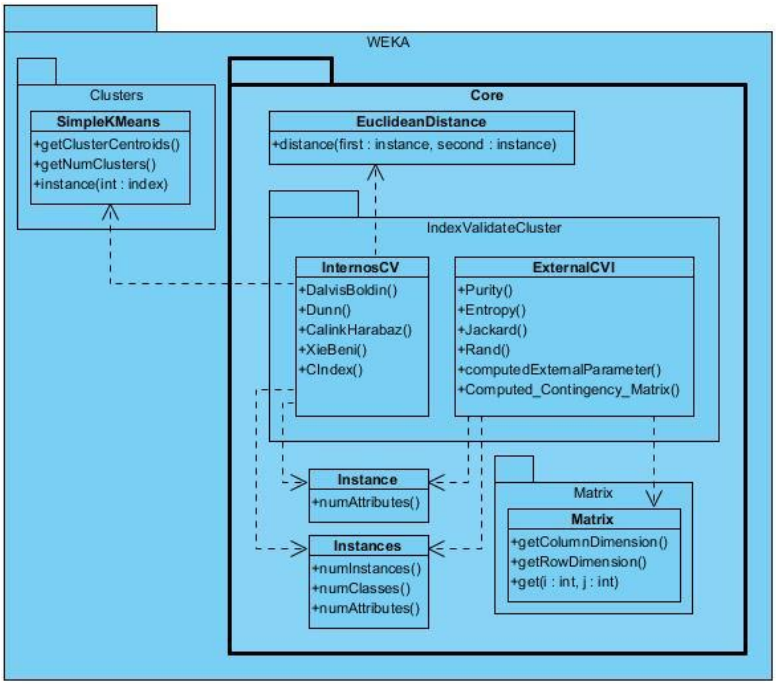

Figura 6 Diagrama de Clases

Se representa el paquete propuesto para dar solución al problema planteado, así como la relación con algunos paquetes de Weka. El paquete Index_Validate_Cluster está compuesto por las clases InternosCV y la clase ExternalCVI, estas contienen los índices de validación internos y externos respectivamente. Mientras que en el paquete Weka se encuentran los paquetes Core y Clusteres, entre otros. El paquete Core está compuesto por las clases e interfaces que conforman la infraestructura de WEKA, define las estructuras que contienen los datos a manejar por los algoritmos de aprendizaje automático, por lo que la propuesta de solución está integrada a este paquete. Contiene además las clases Instance, Instances, EuclideanDistance y el paquete matrix, entre otros.

La clase Instances y la clase Instance permitieron acceder a los datos de las particiones y la base de datos. La clase Instances contiene las bases de datos junto con los métodos para su manejo. Mientras que la clase Instance encapsula cada uno de los ejemplos individuales que forman una base de datos, almacenando los valores de los respectivos atributos. La clase Matrix del paquete Matrix, se utilizó para la implementación del método Computed_Contingency_Matrix, el cual es esencial en la implementación de los índices externos. En la clase EuclideanDistance se encuentra el método distance, que calcula la distancia euclidiana al cuadrado entre dos casos de un agrupamiento. La clase SimpleKMeans está contenida dentro del paquete clusteres. Esta clase contiene los atributos necesarios para la implementación de los índices internos.

2.7.2 Patrones de diseño

Los patrones de diseño brindan una solución a problemas de desarrollo de software que están sujetos a contextos similares. Un patrón de diseño es una descripción de clases y objetos comunicándose entre sí, adaptada para resolver un problema de diseño general en un contexto particular. Son los encargados de identificar Clases, Instancias, Roles, Colaboraciones y la distribución de Responsabilidades (Beck, 1999). El conocimiento de los patrones de diseño es vital para entender el funcionamiento del mismo y por ende lograr los objetivos trazados.

Patrones para Asignar Responsabilidades GRASP

Los patrones generales de software para asignar responsabilidades, más conocidos como GRASP[7], "describen los principios fundamentales de la asignación de responsabilidades a objetos, expresados en forma de patrones" (Larman, 2003).

Experto en información: Este patrón GRASP es el responsable de asignar, la responsabilidad a aquellas clases que contenga la información. De esta forma el diseño obtendrá una mayor cohesión y conservará el encapsulamiento de la información, propiciando el decremento del acoplamiento.

Un ejemplo donde se evidencia este patrón es en la clase ExternalCVI, esta clase contiene la información necesaria para implementar cada uno de los índices de validación de grupo externa, de igual manera sucede con la InternosCV.

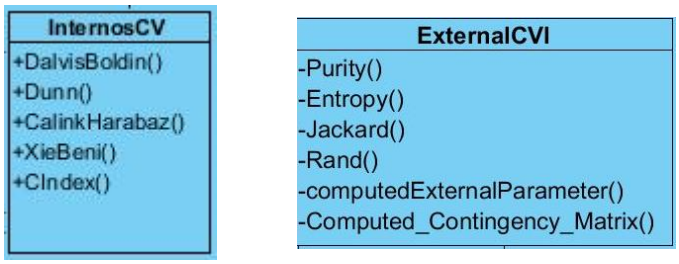

Figura 7 Ejemplo de patrón experto

Bajo acoplamiento: El acoplamiento nos presenta la dependencia de una clase con otras. Una clase tiene alto acoplamiento cuando depende de muchas clases, en caso contrario se estará en presencia de una clase con bajo acoplamiento. Lo que propicia que no afecte el cambio de estas clases por otros componentes, es fácil de entender y de reutilizar.

[7]GRASP: General Responsibility Assignment Software Patterns

Como se ve representado en el diagrama de clases, la mayoría de estas entidades están vinculadas solamente por una relación de uso. Por lo que de realizarse una modificación en una de las clases, esto no afectará a las restantes, asegurando así el bajo acoplamiento. Ejemplo de esto se muestra en la siguiente figura.

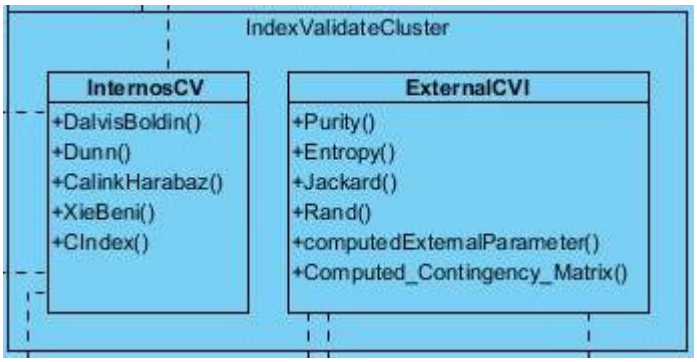

Figura 8 Ejemplo de patrón bajo acoplamiento

Alta Cohesión: Se aplica para realizar un diseño que evite contener clases con un alto grado de abstracción, que asuman responsabilidades que podían haber delegado a otros objetos o que tengan responsabilidades muy complejas.

2.8 Conclusiones del capítulo

La aplicación de los patrones alta cohesión y bajo acoplamiento durante la implementación del paquete de métricas, garantiza que puedan ser incorporados al mismo nuevos índices, tanto internos como externos, sin que esto implique un alto costo en el desarrollo.

Se incorporó el paquete de métricas al paquete Core de la herramienta WEKA, como parte del proceso de integración. De esta forma puede ser utilizado por los algoritmos de agrupamiento disponibles en WEKA.

CAPÍTULO 3. DISEÑO DE LA SOLUCIÓN PROPUESTA

Introducción

En el presente capítulo se diseñan un conjunto de experimentos que permiten validar la solución propuesta. Se propone el test de Friedman, con el objetivo de identificar si existen diferencias entre los índices de validación de agrupamiento analizados. Adicionalmente se realizan experimentos para validar el funcionamiento de los índices de validación de grupo propuestos a través de un caso de estudio. Por último se discuten los resultados obtenidos producto de estos experimentos.

3.1 Validación de los índices de validación de agrupamiento

3.1.1 Base de datos

Para validar el funcionamiento de los índices de validación de agrupamiento, se propone la realización de experimentos, utilizando como caso de estudio al algoritmo SimpleKmeans existente en Weka. Se emplearon 13 bases de datos estándares extraídas del UCI Repository (Iris, breast, diabetes, ecoli, banknote, glass, ionosphere, transfusión, vehicle, vertebral_columm, user_modeling, segment, messidor_features). A continuación se realiza una descripción de estas bases de datos, las cuales se relacionan en la siguiente tabla, donde la primera columna indica el nombre de la base de datos, la segunda se refiere a la cantidad de objetos y la tercera se refiere a la cantidad de rasgos:

Bases de datos	Cantidad de objetos	Cantidad de rasgos
banknote	1372	5
breast	106	10
diabetes	768	9
ecoli	336	8
glass	214	10
ionosphere	351	35
segment	2002	4
iris	150	5
vehicle	846	19
transfusion	748	5
user_modeling	258	6
vertebral_columm	310	7
messidor_features	1151	20

Tabla 13 Descripción de las base de datos

Banknote: Los datos fueron obtenidos a partir de imágenes de billetes auténticos y falsificados. Para la digitalización, se utilizó una cámara industrial normalmente utilizada para la inspección de impresión. Las imágenes finales tienen 400x 400 píxeles.

Breast: Los conjuntos de datos se derivan de un estudio realizado sobre el cáncer de mama en la Universidad de Wisconsin y en el Hospital Madison del Dr. William H. Wolberg.

Pima diabetes: Es creada por el Instituto Nacional de Diabetes, Enfermedades Digestivas y Enfermedades Renales. Basadas en detectar si el paciente muestra signos de diabetes según los criterios establecidos por la Organización Mundial de la Salud. Todos los pacientes evaluados son mujeres de 21 años, que tiene como herencia la enfermedad Diabetes.

Ecoli: Creada por Kenta Nakai en el Instituto de Biología Molecular y Celular, en Japón. Sus datos se basan en la predicción del nivel celular de las proteínas. Está conformado por 336 datos y 8 atributos (7 predictivo, 1 nombre).

Glass: Trata sobre el tamaño de una muestra de 214 tipos de vidrio, sobre los cuales se realizan comparaciones en dependencia de la norma establecida.

Ionosphere: Estos datos se recogieron en La Base de la Fuerza Aérea Canadiense de Goose Bay, mediante un sistema. Este sistema consta de una red en fase de 16 antenas de alta frecuencia, con una potencia de transmisión total del orden de 6,4 kilovatios. Las instancias en esta base de datos son descrito por 2 atributos.

Segement: El conjunto de datos realiza un muestreo de la piel al azar, sobre diferentes grupos de edad (joven, de mediana edad y adulto) y grupos de raza (blanco, negro y asiático). El tamaño de la muestra de aprendizaje es 245.057; de los cuales 50.859 es la muestra de piel y 194.198 es la muestra de la raza.

Iris: Esta base de datos consta de 3 clases de 50 ejemplos cada una, donde cada clase se refiere a un tipo de planta Iris. Una clase está separada linealmente de las otras dos, mientras que dos no son linealmente separables. Consta de cuatro atributos: longitud y ancho del sépalo de la flor y longitud y ancho de los pétalos en centímetros. Los tres tipos de planta Iris son: Iris Setosa, Iris Virginica e Iris Versicolor.

Vehicle: Se basa en la clasificación de 846 vehículos, en dependencia de sus características fundamentales.

Transfusión: Datos extraídos del Centro de Servicios de Transfusión de Sangre en Hsin-Chu City en Taiwán. Se seleccionaron 748 donantes al azar de la base de datos de donantes. Cada dato de los donantes, incluye R (fecha reciente desde el pasado mes donación), F (Frecuencia el número total de la donación), M (Monetario total en sangre donado en cc), T (Tiempo meses desde primera donación), y una variable binaria que representa si él / ella donó sangre en marzo de 2007 (1 soporte para la donación sangre; 0 significa que no realizó la donación de sangre).

Vertebral_columm: Conjunto de datos biomédicos integrado por el Dr. Henrique da Mota, durante un período de residencia médica, en el Grupo de Investigación Aplicada en Ortopedia. Los datos se han organizado en dos tareas de clasificación diferentes, pero relacionados. La primera tarea consiste en clasificar a los pacientes como pertenecientes a una de tres categorías: Largo (100 pacientes), disco Hernia (60 pacientes) o espondilolistesis (150 pacientes). Para la segunda tarea, las categorías de disco Hernia y espondilolistesis se fusionaron en una sola categoría denominada como "anormal". Por lo tanto, la segunda tarea consiste en clasificar a los pacientes como pertenecientes a una de dos categorías: Largo (100 pacientes) o anormales (210 pacientes).

User_modeling: Creado por Hamdi Tolga Kahraman en la Facultad de Tecnología, en el Departamento de Ingeniería de Software de la Universidad Técnica de Karadeniz, contiene 258 atributos y rasgos.

Messidor_features: Este conjunto de datos contiene características extraídas del conjunto de imágenes Messidor[8], con el objetivo de predecir si una imagen contiene signos de retinopatía diabética o no.

3.1.2 Experimentos

Para cada resultado del algoritmo SimpleKmeans, sobre estas bases de datos se calculan los valores de los índices de validación de agrupamiento, tanto internos como externos. Estos valores son tomados para diferentes particiones (Número de grupo = K), (Ver Anexos), de modo que se pueda verificar cuál de las particiones es la mejor. Las siguientes tablas muestran los resultados obtenidos de estos experimentos, para K = 3 en ambos índices de validación, denotando esta partición como la mejor, tomando al índice Dunn como pivote.

[8] Messidor: Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology.

Base de datos	C.I	D.B	C.H	X.B	D
banknote	0,20	0,67	0,01	0,37	1,41
breast	0,21	0,51	0,008	0,60	0,01
diabetes	0,16	1,71	9,61	1,04	0,04
ecoli	0,16	0,85	0,001	0,38	0,004
glass	0,25	0,56	0,01	0,25	0,004
ionosphere	0,20	2,47	0,01	2,09	0,003
segment	0,14	0,14	4,21	0,09	0,002
iris	0,44	0,36	0,001	0,21	0,004
vehicle	0,23	0,97	1,64	0,64	0,005
transfusion	0,041	0,7	5,29	0,41	3,60
user_modeling	0,29	1,5	0,004	0,84	0,01
vertebral_columm	0,08	1,00	0,005	0,50	0,001
messidor_features	0,12	0,64	8,92	0,39	9,01

Tabla 14 Resultado de los indices de validación internos para K=3.

Base de datos	E	J	P	R
banknote	0.85	0,31	0,69	0,56
breast	1,63	0,31	0,50	0,56
diabetes	0,84	0,31	0,66	0,56
ecoli	0,99	0,31	0,75	0,56
glass	1,70	0,31	0,45	0,56
ionosphere	0,81	0,31	0,68	0,56
segment	0,26	0,31	0,93	0,56
iris	0,41	0,31	0,88	0,56
vehicle	1,80	0,31	0,38	0,56
transfusion	0,78	0,31	0,76	0,56
user_modeling	1,26	0,31	0,56	0,56
vertebral_columm	0,77	0,31	0,69	0,56
messidor_features	0,99	0,31	0,53	0,56

Tabla 15 Resultado de los indices de validación externa para K=3.

3.1.3 Caso Estudio para los índices de validación de agrupamiento

La herramienta Weka permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, sobre cualquier conjunto. Para ello únicamente se requiere que los datos a analizar se almacenen con formato ARFF[9]. Usando como base de dato transfusion, la cual contiene 748 objetos con 5 atributos. Se explica el proceso de validación, primeramente en necesario cargar la base de dato con el formato establecido.

[9] ARFF: Attribute-Relation File Format

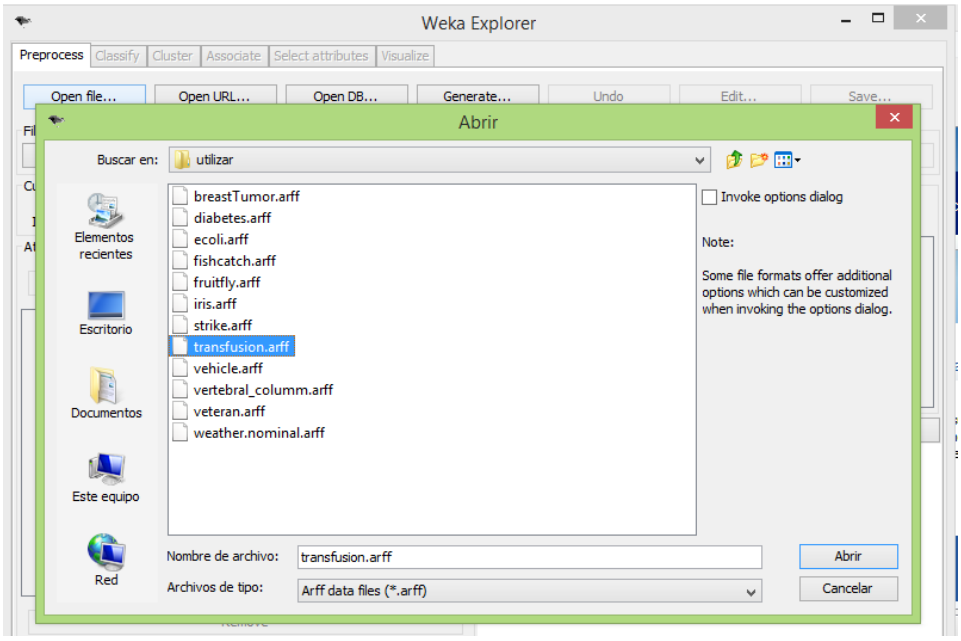

Figura 9 Acceso desde la herramienta WEKA a una base de datos.

Pulsando la tercera pestaña, llamada Cluster, en la parte superior de la ventana se accede a la sección dedicada al agrupamiento. Se elige el algoritmo de agrupamiento SimpleKMeans, el cual será utilizado para evaluar las diferentes bases de datos, con los respectivos índices de validación internos y externos. Es importante dar valor True a las opciones displayStdDevs y preserverInstancesOrder señaladas en la figura 10 y con el botón **Start** empieza el funcionamiento (Figura 11).

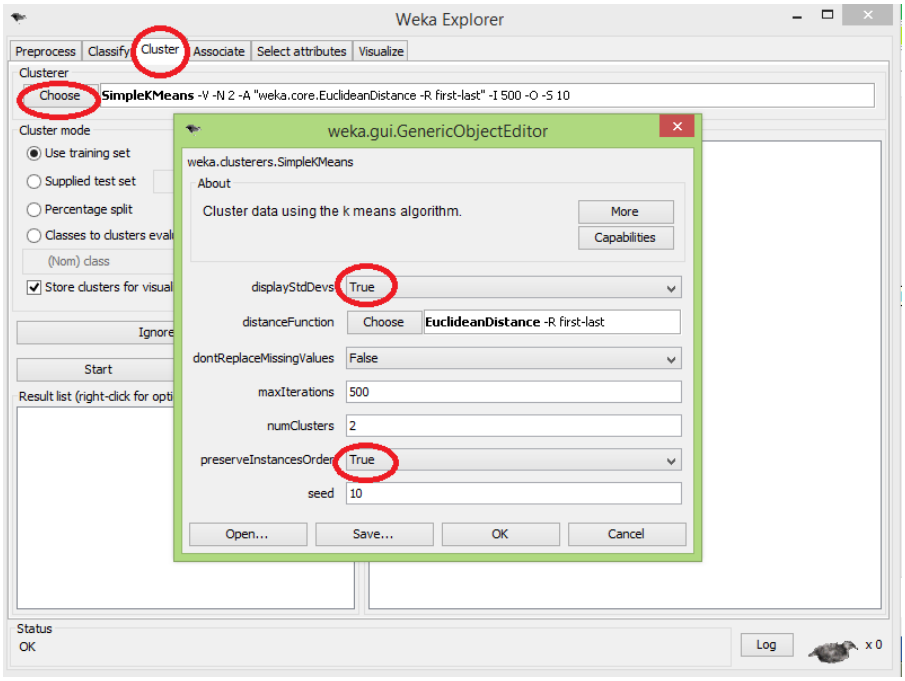

Figura 10 El modo Cluster dentro del explorer de WEKA.

Luego se obtiene el resultado de evaluar el algoritmo SimpleKmeans sobre la base de datos transfusion, el cual se muestra en la figura 11.

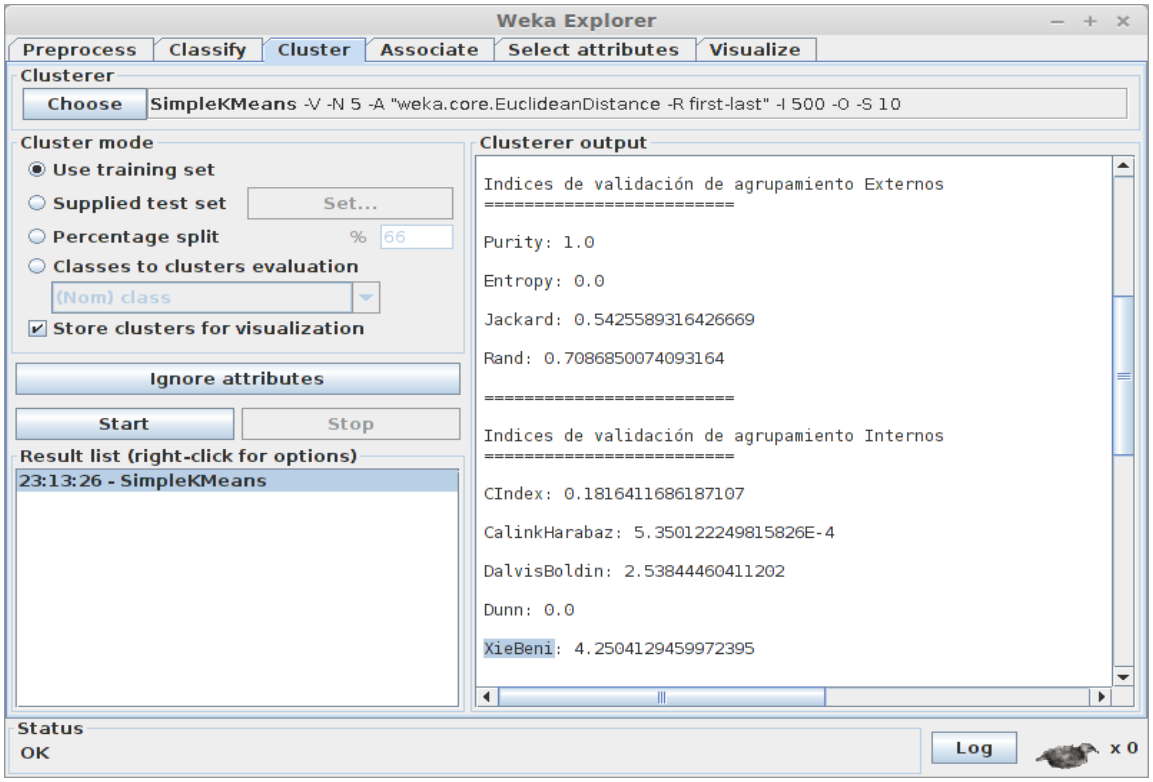

Figura 11 Evaluación del algoritmo SimpleKmeans sobre un conjunto de datos.

3.2 Test de Friedman.

La prueba Friedman es un equivalente no paramétrico de ANOVA[10] (Friedman, 1937), su objetivo fundamental es determinar si existen o no diferencia entre los algoritmos analizados. Ordena los algoritmos para cada conjunto de datos separadamente, donde el algoritmo con mejor desempeño obtiene la jerarquía de 1, el segundo mejor la jerarquía 2 y así sucesivamente. La prueba Friedman compara las jerarquías comunes entre los algoritmos a través de la siguiente fórmula:

$$R_j = \frac{1}{N}\sum_i r_i^j \qquad (24)$$

Dónde:

[10] Análisis de varianza

r_i^j: Es el valor del algoritmo j en la base de dato i.

N: El número de bases de datos.

Friedman declara como hipótesis nula que los algoritmos son equivalentes y sus rankings iguales. Si esta hipótesis nula es negada, es necesario realizar una prueba post-hoc (Nemenyi) que se basa en el cálculo de la diferencia crítica, comparando los algoritmos entre sí. A partir de esta comparación establece que los algoritmos son significativamente diferentes, si el ranking resultante de su comparación es mayor o igual a esta diferencia, calculada de la siguiente forma (Demsar, 2006):

$$DC = q_\sigma \sqrt{\frac{K(K+1)}{6N}} \qquad (25)$$

Siendo K el número de algoritmos.

Mientras que q_σ es un constante, dada por K y el valor de σ , en la siguiente tabla se muestra esta relación.

σ \ K	2	3	4	5	6	7	8	9	10
0.05	1.960	2.343	2.569	2.728	2.850	2.949	3.031	3.102	3.164
0.10	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

Tabla 16 Valores de q_σ .

3.2.1 Resultados del Test de Friedman.

En la siguiente tabla se muestra el ranking obtenido por el test de Friedman para K=3.

Algoritmo	**Ranking**
Pureza	2,80
Entropía	3,61
Jaccard	1,07
Rand	2,50

Tabla 17 Valores obtenidos por el test de Friedman para los índices externos.

Algoritmo	**Ranking**
CIndex	1,69
Calinski-H	3,23
Davies-B	4,03

Dunn	2,26
Xie-B	3,76

Tabla 18 Valores obtenidos por el test de Friedman para los índices internos.

Al analizar los resultados se concluye que existen diferencias entre los algoritmos. Se determinó que el mejor algoritmo para K = 3, K=5, K=7 y K=9 en los índices externos es Jaccard. Por otra parte para K = 3, K=5 y K=9 en los índices internos es CIndex, mientras que para K = 7 es Calinski-Harabaz (Ver Anexos). A raíz de que no se cumple la hipótesis nula se hace necesario aplicar la prueba post-hoc de Nemenyi, se calcula el valor de la distancia crítica para los algoritmos analizados anteriormente. El valor obtenido es de 2.35 para $q_\sigma = 0.05$ en los índices externos, mientras que para los internos obtiene un valor de 5.47. A continuación se muestran los resultados obtenidos por el test de Friedman de las comparaciones entre los algoritmos de la presente investigación, para los valores de $q_\sigma = 0.05$.

Figura 12 Comparación entre los índices externos.

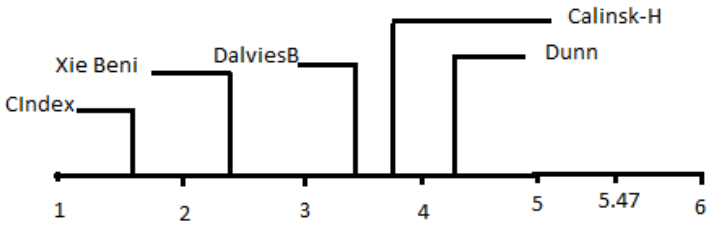

Figura 13 Comparación entre los índices internos.

Como se evidencia en los índices Pureza y Entropía presentan diferencias significativas con el resto de los índices externos. Por otra parte los índices internos no presentan diferencias significativas entre ellos para K = 3 y $q_\sigma = 0.05$, esto sucede también para K = 5 y $q_\sigma = 0.05$. (Ver Anexos)

3.3 Conclusiones del capítulo

Se utilizó un caso de estudio para evaluar el funcionamiento de los métodos de validación de grupo sobre varias bases de datos, evaluadas por el algoritmo SimpleKmeans. En general los resultados de este experimento mostraron como mejor partición para las bases de datos K = 3, mostrando como mejor índice Dunn.

Con el test de Friedman aplicado se determinó que existen diferencias entre los algoritmos analizados, por lo que se hizo necesario aplicar una prueba post-hoc de Nemenyi, obteniendo como resultado que los índices Pureza y Entropía presentan diferencias significativas respecto a los otros algoritmos externos, mientras que los internos no presentan diferencias significativas para K = 3 y K = 5.

CONCLUSIONES GENERALES:

El estudio de los métodos de validación de grupo permitió identificar dos enfoques: índices de validación externa e índices de validación interna, en el caso de la validación externa se identificaron los índices a partir de las métricas analizadas y en el caso de la validación interna se basa en el cálculo de distancias utilizando la distancia euclidiana.

La aplicación de los patrones alta cohesión y bajo acoplamiento durante la implementación del paquete de métricas, garantiza que puedan ser incorporados al mismo nuevos índices de validación, tanto internos como externos, sin que esto implique un alto costo en el desarrollo. Se garantiza que el mismo pueda ser utilizado por los algoritmos de agrupamiento disponibles en WEKA al incorporarlo al paquete Core de dicha herramienta.

Los resultados del caso de estudio para evaluar el funcionamiento de los métodos de validación de grupo sobre varias bases de datos, evaluadas por el algoritmo SimpleKmeans, mostraron como mejor partición para las bases de datos K = 3.
Con el test de Friedman aplicado se determinó que existen diferencias entre los algoritmos analizados, por lo que se hizo necesario aplicar una prueba post-hoc de Nemenyi, obteniendo como resultado que los índices Pureza y Entropía presentan diferencias significativa respecto a los otros algoritmos externos, mientras que los internos no presentan diferencias significativas para K = 3 y K = 5.

Con la integración del paquete de métricas implementado a la herramienta WEKA permite la evaluación de la calidad de las particiones resultantes de evaluar los algoritmos de agrupamiento.

REFERENCIAS

A. González, C. (s.f.). Comparación de dos métodos de aprendizaje sobre el mismo problema.

Almeida, Fernando Antonio de. 2014. Family clustering of secondary chronic kidney disease with hypertension or diabetes mellitus. A case-control study. 2014.

Alpaydin, E. 1999. Combined 5 × 2 F test for comparing supervised classification learning algorithms. Neural Computation. 1999. 1885–1892.

Amigo, Enrique. 2009. A comparison of Extrinsic Clustering Evaluation. Madrid, Spain: UNED. : Departamento de Lenguajes y Sistemas Informáticos,, 2009.

Beck, K. 1999. Extreme Programming Explained. Embrace Change. s.l. : s.l. : Pearson Education, 1999.

Bengio, Y., & Grandvalet, Y. (. 2004. No unbiased estimator of the variance of k-fold cross-validation. Journal of Machine Learning Research,. 2004.

Berzal, Fernando. 2012. Clustering. s.l. : Universidad de Granada, 2012.

Bouckaert, R. . 2003. Choosing between two learning algorithms based on calibrated tests. Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003). Washington DC : s.n., 2003.

Bouckaert, R., & Frank, E. 2004. Evaluating the replicability of significance tests for comparing. Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference. Sydney: D. Honghua, R. Srikant, and C. Zhang. 2004.

Bouckaert, Remco R. 2010. WEKA Manual for version 3-7-2. 2010.

Bouguessa, W. 2006. An objective approach to cluster validation. 2006.

Bouldin. 1979. A clustering separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979.

Bradzil, P. B., & Soares, C. 2000. A comparison of ranking methods for classification algorithm selection. Proceedings of 11th European Conference on Machine Learning. 2000.

Buuren, Stef van. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. 2007.

Calinski. 1974. A dendrite method for cluster analysis. Communications in Statistics. 1974.

Caparrini, Fernando Sancho. 2013. Introducción al Aprendizaje Automático. s.l. : Dpt. Ciencias de la Computación e Inteligencia Artificial. Univ. de Sevilla , 2013.

Defays, D. 1977. An efficient algorithm for a complete linkmethod. s.l. : The Computer Journal, 20, 364 – 366 , 1977.

Demsar, J. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. Jurnal of Machine Learning Research 7,. 2006. 1-30.

Dietterich, T. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, . 1998. 1895–1924.

Dubes. 1988. Algorithms for Clustering Data. Prentice-Hall. N.J, USA : s.n., 1988.

Duda, and Hart. 1973. Pattern Classification, and Scene Analysis. New York : John Wiley & Sons, 1973.

E. Rendón. 2011. Internal versus External cluster validation indexes. 2011. págs. Issue 1, Volume 5.

Ester, M. 1996. A density-based algorithm for discovering clusters in large special databases with noise. 1996.

Estrada, José Antonio Cecchini. 2013. Aplicaciones del modelo de autodeterminación en la educación física de primaria. 2013.

Fowlkes y Mallows. 1983. A method for comparing two hierarchical clusterings. s.l. : Journal of the American Statistical Association, 1983. 78(383):553-569.

Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association. 1937.

Goikoetxea, Ibai Gurrutxaga. 2010. Aportaciones a la clasicación no supervisada y a su validación. Aplicación a la seguridad informática. Donostia : s.n., 2010.

González, Damaris. 2010. Algoritmos de Agrupamiento basados en densidad y Validación de clusters. s.l. : Universitat Jaume I, 2010.

Gurrutxaga, I. 2008. Evaluation of Malware clustering based on its dynamic behaviour. Adelaide, Australia, 1 : Proceedings of the 7th Australasian Data Mining Conference (AusDM), 2008. 163-170.

Halkidi, M. 2002a. Cluster validity methods part I. s.l. : SIGMOD Rec, 2002a. 40 – 45.

Halkidi, Maria. 2002b. Cluster validity methods part II. 2002b.

HALKIDI, MARIA. 2001. On Clustering Validation Techniques. 2001.

Hinneburg. 1998. An efficient Approach to Clustering in Large Multimedia Databases with Noise.Knowledge. 1998.

Hull. 1994. Information Retrieval Using Statistical Classification. 1994.

I. Perona, Iñaki Albisua, Olatz Arbelaitz, Ibai Gurrutxaga, Jose I.Martin, Javier Muguerza, y Jesus M. Perez. 2009. Histogram based payload processing for unsupervised anomaly detection systems in network intrusion. s.l. : Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA),, 2009. pag. 329-340.

Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. Bulletin de la Societ´e Vaudoise de Sciences Naturelles . 1908.

Jain. 1986. Bootstrap technique in cluster analysis. ,. s.l. : Pattern Recognition, 1986. 20, 547 – 569.

Jain, A.K. 1999. Data Clustering: A Review. ACM Computing Surveys, 31(3), 264–323. 1999.

Jímenez, M. G. 2003. Análisis de Datos en WEKA– Pruebas de Selectividad. 2003.

Larman, Craig. 2003. UML y patrones: una introducción al análisis y diseño orientado a objetos y al proceso unificado. [trad.] Begoña Moros Valle. Segunda. s.l. : Pearson Educación, 2003, 2003. 9788420534381.

Levine. 2001. Resampling method for unsupervised estimation of cluster validity. s.l. : Neural Computation, 2001. 13, 2573-2593.

Métricos, Espacios. 2011. 2011.

N. Bolshakova, Azuaje. 2003. Cluster Validation Techniques for genome expresión data. s.l. : Signal Processing, 83, 825 – 833, 2003.

Nadeau, C., & Bengio, Y. 2000. Inference for the generalization error. Advances in Neural Information Processing Systems 12. 2000.

Netbeans. Netbeans. [En línea] [Citado el: 20 de abril de 2015.] https://netbeans.org.

Olvera, J.A. and Martinez, F.T. 2005. Edition schemes based on BSE, Lectura Note in Computer Science, Progress in Pattern Recognition Image Analysis and Applications,10th Iberoamerican Congress on Pattern Recognition, CIARP. 2005.

Oñate, Silvia Fabiola Cujano. 2012. Análisis e Implementación de Alta Disponibilidad Mediante Clustering en Sistemas de Call Center Basados en VoIP. 2012.

OPTICS. 1999. Ordering Points To Identify the Clustering Structure. 1999.

Orallo, Jose Ramirez Quintana, Cesar Ferri. 2001. Introducción a la Minería de Datos. Madrid : s.n., 2001.

Otero, Abraham. 2007. Tutorial Básico de Java. 2007.

Piper, I. 2003. The BrainIT group: concept and core dataset. 2003.

Pizarro, J., Guerrero, E., & Galindo, P. 2002. Multiple comparison procedures applied to model selection. 2002. Neurocomputing 48 155–173..

Programming. 2010. Programming. [En línea] 2010. [En línea] http://mscerts.programming4.us/es/398166.aspx..

Ramakrishna. 2005. New indices for cluster validity assessment. 2005.

Rand. 1971. Objective criteria for the evaluation of clustering methods. s.l. : Journal of American Statistical Association, 1971. 66:846-850.

Raymond. 2013. Eficient and Efective Clustering Methods for Spatial Data Mining. 2013.

Rodriguez, Luz Maria Hernandez. 1999. UML Y PATRONES. Introduccion al analisis y diseño orientado a objetos. 1999.

Rousseeuw. 1987. Silhouttes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. s.l. : Journal of Computational and Applied Mathematics, 1987. 20, 53 – 65.

Salzberg, S. 1997. On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Mining and Knowledge Discovery,. 1997.

Sánchez, Sergio Luengo. 2014. Clustering basado en redes bayesianas con predictoras continuas: aplicaciones en neurociencia. 2014.

Sibson, R. SLINK. 1973. An optimally efficient algorithm for the single link cluster method. . s.l. : Computer Journal, 16, 30– 34, 1973.

Stanfill, C. and Waltz, D. 1986. Toward memory-based reasoning. 1986.

Vazquez, E., Escolano, A., Junquera, J., & Riaño, P. 2001. Repeated measures multiple comparison procedures applied to model selection in neural networks. Proc. of the 6th Intl. Conf. On Artificial and Natural Neural Networks,. 2001. (págs. 88-95)..

VisualParadigm. Visual Paradigm. Sitio web oficial de Visual Paradigm. [En línea] http://www.visual-paradigm.com/.

W. J Matthijs. 2008a. On the equivalence of Cohen's Kappa and the Hubert Arabie Adjusted Rand index. s.l. : Journal of Classification, 2008a. 25(2):177-183.

Weiss, Indurkhya. 1998. “Predictive Data Mining”. San Francisco : Morgan Kaufmann, 1998.

Wilson, D.R. and T.R. Martinez. 2000. Reduction techniques for instance based learning algorithms. 2000.

WUNSCH, DONALD C. 2009. Clustering. s.l. : IEEE Press, 2009.

ANEXOS

Anexo 1

Base de datos	E	J	P	R
banknote	0,59	0,28	0,82	0,60
breast	1,51	0,28	0,49	0,60
diabetes	0,80	0,28	0,67	0,60
Ecoli	0,89	0,28	0,79	0,60
glass	1,69	0,28	0,49	0,60
ionosphere	0,57	0,28	0,85	0,60
segment	0,23	0,28	0,93	0,60
iris	0,32	0,28	0,89	0,60
vehicle	1,75	0,28	0,41	0,60
transfusion	0,74	0,28	0,76	0,60
user_modeling	1,22	0,28	0,59	0,60
vertebral_column	0,70	0,28	0,71	0,60
messidor_features	0,93	0,28	0,60	0,60

Tabla 19 Resultado de los indices de validación externa para K=5.

Anexo 2

Base de datos	E	J	P	R
banknote	0,46	0,22	0,86	0,58
breast	1,37	0,22	0,55	0,58
diabetes	0,77	0,22	0,70	0,58

Ecoli	0,74	0,22	0,81	0,58
glass	1,38	0,22	0,54	0,58
ionosphere	0,63	0,22	0,82	0,58
segment	0,03	0,22	0,99	0,58
iris	0,21	0,22	0,88	0,58
vehicle	1,64	0,22	0,43	0,58
transfusion	0,74	0,22	0,76	0,58
user_modeling	1,15	0,22	0,63	0,58
vertebral_column	0,63	0,22	0,78	0,58
messidor_features	0,92	0,22	0,61	0,58

Tabla 20 Resultado de los indices de validación externa para K=7.

Anexo 3

Base de datos	**E**	**J**	**P**	**R**
banknote	0,46	0,17	0,85	0,56
breast	1,29	0,17	0,62	0,56
diabetes	0,76	0,17	0,70	0,56
Ecoli	0,69	0,17	0,79	0,56
glass	1,18	0,17	0,66	0,56
ionosphere	0,54	0,17	0,85	0,56
segment	0,02	0,17	0,99	0,56
iris	0,27	0,17	0,91	0,56
vehicle	1,47	0,17	0,52	0,56

transfusion	0,72	0,17	0,76	0,56
user_modeling	1,00	0,17	0,65	0,56
vertebral_column	0,65	0,17	0,76	0,56
messidor_features	0,92	0,17	0,61	0,56

Tabla 21 Resultado de los indices de validación externa para K=9.

Anexo 4

Base Datos	**C.I**	**D.B**	**C.H**	**X.B**	**D**
banknote	0,29	0,53	0,001	0,39	4,11
breast	0,24	0,73	0,009	0,69	0,002
diabetes	0,20	1,58	0,001	0,90	0,002
Ecoli	0,15	1,30	0,002	1,83	0,002
glass	0,13	0,65	0,01	0,75	0,001
ionosphere	0,19	2,01	0,01	1,63	0,001
segment	0,09	0,39	7,46	8,83	5,57
iris	0,31	0,7	0,001	0,84	0,002
vehicle	0,18	1,27	2,68	0,72	0,005
transfusion	0,05	0,61	6,65	0,4	8,47
user_modeling	0,26	1,66	0,007	1,00	0,009
vertebral_columm	0,07	1,07	0,007	0,61	0,001
messidor_features	0,18	1,22	0,001	0,70	7,62

Tabla 22 Resultado de los indices de validación interna para K=5.

Anexo 5

Base Datos	C.I	D.B	C.H	X.B	D
banknote	0,17	0,56	0,001	0,45	2,09
breast	0,07	0,73	0,01	0,78	0,003
diabetes	0,14	1,66	0,001	1,35	0,003
Ecoli	0,06	1,16	0,002	1,04	0,001
glass	0,09	0,90	0,01	0,55	0,001
ionosphere	0,20	0,1	0,02	2,68	0,002
segment	0,20	0,23	2,8	2,72	1,32
iris	0,32	0,60	0,002	0,45	0,008
vehicle	0,23	1,35	3,55	0,81	0,004
transfusion	0,04	0,63	8,15	0,47	8,49
user_modeling	0,39	1,20	0,007	0,84	0,008
vertebral_columm	0,17	1,15	0,009	0,90	0,001
messidor_features	0,12	1,01	0,001	0,73	8,01

Tabla 23 Resultado de los indices de validación interna para K=7.

Anexo 6

Base Datos	C.I	D.B	C.H	X.B	D
banknote	0,26	0,60	0,001	0,49	1,72
breast	0,12	0,80	0,01	2,53	0,003
diabetes	0,21	1,60	0,002	1,41	0,002
Ecoli	0,08	1,15	0,002	1,45	0,001
glass	0,26	1,40	0,01	3,44	0,001

ionosphere	0,19	2,83	0,02	2,26	0,003
segment	0,20	0,21	8,52	0,62	2,59
iris	0,32	0,67	0,002	0,54	0,01
vehicle	0,23	1,36	4,19	0,93	0,005
transfusion	0,06	0,49	7,5	0,61	1,77
user_modeling	0,36	1,14	0,008	0,75	0,01
vertebral_column	0,18	1,14	0,01	0,90	0,001
messidor_features	0,14	1,04	0,001	0,88	6,24

Tabla 24 Resultado de los indices de validación interna para K=9.

Anexo 7

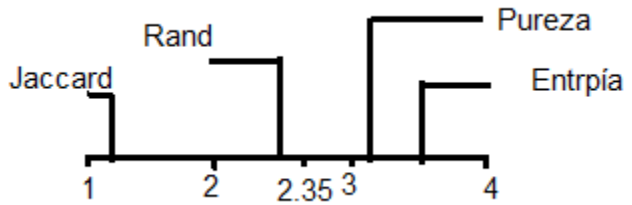

Figura 14 Comparación entre los índices externos para K = 5

Anexo 8

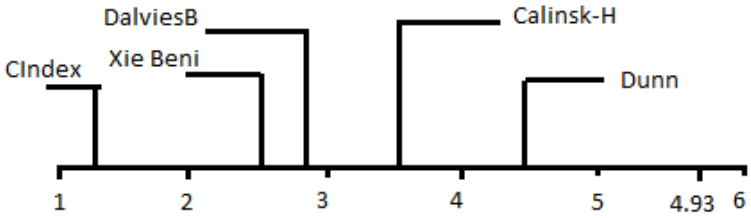

Figura 15 Comparación entre los índices internos para K = 5