UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS
FACULTAD DE MATEMÁTICA, FÍSICA Y COMPUTACIÓN

Extensiones al módulo de validación del agrupamiento en Weka y su evaluación

Trabajo de Diploma en Ciencia de la Computación

Autor: Maikel Castellanos Placer

Tutores: MSc. Leticia Arco García

Dr. Rafael Esteban Bello Pérez

Santa Clara, 2008

A mis padres y a mi hermano por confiar siempre en mí,

A mis abuelos que donde quieran que estén siempre los recordaré, principalmente a mi abuelo Manuel,

A mis tías, tíos y otros familiares que siempre estuvieron al tanto de mis estudios, principalmente a mi tío Héctor por todo lo bueno que ha sido conmigo.

A mis amigos con los cuales he pasado estos largos 5 años y a otros de ellos que se quedaron en el camino, por pasar momentos difíciles y gratos, por soportar mis pesadeces, cueros y recibir algunos a cambio.

A mi tutora por toda su ayuda, por estar siempre dispuesta a atenderme en cualquier momento, incluso cuando estuvo ocupada con su doctorado y por tener paciencia para responder cualquier inquietud o pregunta mía, aunque fuese las más tonta del mundo.

A Omarito por ayudarme cada vez que lo fui a molestar con cualquiera pregunta sobre Weka.

Resumen

En este trabajo se realizaron extensiones al módulo de validación del agrupamiento en Weka. A este módulo se le incorporaron varias medidas internas y externas referenciadas en la literatura para la validación del agrupamiento. Medidas basadas en la Teoría de los Conjuntos Aproximados también se incorporaron a este módulo. Dos motivos principales condujeron al desarrollo de estas extensiones. Uno de ellos, comparar las medidas basadas en la Teoría de Conjuntos Aproximados con las medidas clásicas que se referencian en la literatura. El otro, enriquecer el módulo de validación del agrupamiento en Weka. Se definieron casos de estudio que permitieron evaluar las medidas basadas en la Teoría de los Conjuntos Aproximados, mostrando su confiabilidad y validez.

Abstract

In this work validations of the extentions to the validation module of the clustering in Weka have been made. To this module were incorporated some internal and external measures which are referenciated in literature dealing with clustering validation. Measures based on the Rough Sets Theory were also incorporated to this module. Two main reasons led to the development of this extension: to compare the measures based on the Rough Sets Theory with the classic measures referenciated in literature and to enrich the validation module of the clustering in Weka. Two cases of study were defined wich allowed to evaluate the measures basedon the Rough Sets Theory, Showing its reliability and validity.

Tabla de contenidos

INTRODUCCIÓN **1**

1 Acerca de la validación del agrupamiento **4**

1.1 Agrupamiento 4
1.2 Validación del agrupamiento: Su clasificación 5
1.2.1 Principales medidas externas *7*
1.2.2 Principales medidas internas *9*
1.3 Medidas basadas en la teoría de conjuntos aproximados 15
1.3.1 Fundamentos teóricos *15*
1.3.2 Conjuntos aproximados para evaluar el agrupamiento *17*
1.4 Consideraciones finales del capitulo 22

2 Incorporación de medidas de validación del agrupamiento en Weka **24**

2.1 Weka generalidades 24
2.1.1 Familiarización con el ambiente de Aprendizaje Automatizado Weka *25*
2.1.2 Entrada de datos *27*
2.1.3 Interioridades de Weka *28*
2.1.4 Factibilidad de la implementación de nuevos modelos en Weka *31*
2.2 Algoritmos de agrupamiento en Weka 33
2.2.1 Cobweb *33*
2.2.2 DBScan *34*
2.2.3 EM *35*
2.2.4 FarthestFirst *36*
2.2.5 OPTICS *36*
2.2.6 K-Means *36*
2.3 Validación del agrupamiento en Weka 37
2.4 Diseño e implementación en Weka de las medidas de validación 38
2.4.1 Diagrama de clases *38*
2.4.2 Implementación de las clases *40*
2.4.3 Interfaz de usuarios *43*
2.5 Conclusiones parciales 45

3 Resultados experimentales **46**

3.1 Definición de casos de estudio y herramientas utilizadas 46
3.2 Diseño y aplicación de experimentos 47
3.2.1 Medir confiabilidad *48*

3.2.2 Medir validez ... *48*
3.2.2.1 Validez de contenido ... 49
3.2.2.2 Validez de criterio ... 49
3.2.2.3 Validez de constructo ... 50
3.3 Descripción de los archivos utilizados para evaluar las medidas basadas en RST ... 52
3.4 Conclusiones parciales ... 55

CONCLUSIONES Y RECOMENDACIONES ... 57

Referencias bibliográficas ... 58

Anexos ... 64

Anexo 1. Distancias, similitudes y disimilitudes más usadas para comparar objetos ... 64
Anexo 2. Medidas externas para la evaluación del agrupamiento ... 67
Anexo 3. Medidas internas para la evaluación del agrupamiento ... 70
Anexo 4. Variantes para el cálculo del umbral de similitud entre objetos ... 73
Anexo 5. Comparación de las medidas aplicadas sobre bases de casos con y sin ruido ... 74
Anexo 6. Correlaciones entre medidas basadas en RST e internas referenciadas ... 75
Anexo 7. Correlaciones entre medidas basadas en RST y externas referenciadas ... 78

INTRODUCCIÓN

Una forma de evaluación del agrupamiento muy sencilla, puede ser, por ejemplo, mediante la visualización del conjunto de datos cuando éste es pequeño y los datos son bidimensionales. Sin embargo, muchas técnicas de agrupamiento se han desarrollado especialmente para el reconocimiento de estructuras en espacios de datos altamente dimensionales, esta forma de evaluación puede ser extremadamente difícil, al intentar realizar una visualización efectiva de un conjunto de datos de alta dimensionalidad (Halkidi, Batistakis et al. 2001b). Cuando los datos no pueden representarse gráficamente, o es muy difícil o algunas veces imposible para un observador determinar una partición de los mismos; o no existe una forma simple de decidir si una partición es correcta, se requiere acudir a medidas o índices de validación del agrupamiento.

Muchos algoritmos de agrupamiento varían sus resultados dependiendo de las características de los datos (e.g., geometría y densidad de distribución) (Halkidi, Batistakis et al. 2001b). Otros dependen fuertemente de los valores asignados a los parámetros. Por ejemplo, si hay un parámetro que controle la resolución a la cual los datos son vistos, el algoritmo produce un dendograma en función de ese parámetro. Entonces es necesario decidir cuál nivel del dendograma refleja mejor los grupos naturales presentes en los datos. Algunos algoritmos necesitan que se especifique inicialmente el número de grupos a obtener, otros requieren que se especifique el número de vecinos de cada punto como un parámetro externo. Así, los resultados producidos son en función de los parámetros fijados y se hace necesario verificar cuáles se ajustan a los datos (Levine and Domany 2001).

Variaciones a partir de características de los datos, diferentes técnicas de análisis de grupos y definición de parámetros para el algoritmo a aplicar, indican que una evaluación de los resultados es necesaria para medir la calidad del agrupamiento. Una práctica común, en tal sentido, es aplicar medidas de validación de grupos (Stein, Eissen et al. 2003).

El procedimiento de evaluar los resultados de algoritmos de agrupamiento se conoce por validación del agrupamiento (Theodoridis and Koutroubas 1999, Halkidi, Batistakis et al. 2002). Se dice medida de validación de grupos a una función que hace corresponder a un agrupamiento un número real, indicando en qué grado el agrupamiento es correcto o no

(Höppner, Klawonn et al. 1999). Estas medidas en su mayoría son heurísticas por naturaleza para conservar una complejidad computacional plausible.

Existen varias medidas de validación del agrupamiento, sin embargo éstas no se encuentran integradas en un único sistema que permita comparar sus resultados para distintos métodos de agrupamiento. Por otra parte, existe la herramienta Weka donde se encuentran implementados varios métodos de agrupamiento, pero las técnicas de evaluación presentes en esta herramienta no son suficientes.

Formulación del problema:

Los sistemas existentes que integran varios algoritmos de agrupamiento, por ejemplo el Weka, no tienen incorporado un buen módulo de evaluación, por tanto no existe una herramienta que integre algoritmos de agrupamiento y varias medidas de forma tal que se puedan estudiar las principales medidas de validación del agrupamiento bajo las mismas condiciones.

A partir de problema formulado se plantean las **preguntas de investigación** siguientes:

¿Implementar en Weka las principales medidas de validación permitirá la evaluación de éstas como instrumentos de medición?

¿Es posible evaluar las medidas basadas en la Teoría de los Conjuntos Aproximados para la validación del agrupamiento utilizando varios algoritmos aplicados a diversos conjuntos de datos?

Para dar solución al problema formulado y respuesta a las preguntas planteadas nos proponemos el **objetivo general** siguiente:

Extender el módulo para validar resultados de algoritmos de agrupamiento.

Para cumplimentar el objetivo general, se proponene los **objetivos específicos** siguientes:

1. Incorporar al Weka las medidas internas y externas referenciadas en la literatura, así como las medidas basadas en la Teoría de los Conjuntos Aproximados propuestas por investigadores del Laboratorio de Inteligencia Artificial.

2. Evaluar las medidas internas, externas y basadas en la Teoría de los Conjuntos Aproximados como instrumentos de validación del agrupamiento, utilizando algunos algoritmos de agrupamiento de Weka y varias bases de casos.
3. Comparar los resultados de las medidas basadas en la Teoría de los Conjuntos Aproximados respecto a los resultados de las medidas clásicas.

Justificación de la investigación:

La necesidad del desarrollo de esta investigación está dada porque en el laboratorio de Inteligencia Artificial del CEI se han desarrollado medidas para validar el agrupamiento utilizando la teoría de los conjuntos aproximados, sin embargo la evaluación de las mismas sólo se ha realizado a nivel de resultados de agrupamiento de documentos. Si se desea realizar una evaluación de este instrumento de medición mediante el estudio comparativo con el resultado de las principales medidas publicadas en la literatura científica y a partir del resultado de varios métodos de agrupamiento y aplicados a disímiles conjuntos de datos, se hace necesario implementar todas las medidas en una herramienta que integre métodos de agrupamiento y que tenga un diseño extensible. Estas facilidades las tiene la herramienta Weka y por eso en esta herramienta se van a implementar las principales medidas para realizar un estudio comparativo de las mismas. Además, enriquecer el módulo de validación del agrupamiento en Weka.

1 Acerca de la validación del agrupamiento

El volumen de datos es cada día mayor; una de sus causas es el crecimiento exponencial de las colecciones de datos no estructurados (por ejemplo, textuales). Por tanto, el desarrollo de técnicas que permitan el análisis exploratorio de los datos es fundamental. El análisis de grupos permite descubrir la estructura interna de éstos (grupos, conglomerados, comunidades, subconjuntos, clases) e identificar distribuciones interesantes y patrones subyacentes en ellos, considerando muy poca o ninguna información a priori (Levine and Domany 2001). Los algoritmos para la detección de comunidades juegan un rol similar al análisis de grupos, pero utilizando las propiedades topológicas de las redes complejas. Dos tareas estrechamente relacionadas con los algoritmos de agrupamiento son la evaluación y el etiquetamiento de los grupos encontrados. Éstas permiten conocer con qué grado de certeza los grupos fueron obtenidos y cómo es posible caracterizarlos. A continuación se citarán las principales técnicas de agrupamiento, se particularizará en los llamados algoritmos para la detección de comunidades en redes complejas y se abordarán formas de validación del agrupameinto.

1.1 Agrupamiento

Existen varias definiciones de agrupamiento, en este trabajo se considera la formalizada en (Jain and Dubes 1988): “El análisis de grupos organiza los datos mediante la extracción de la estructura subyacente en ellos como una partición de individuos o como una jerarquía de grupos. La representación puede entonces ser analizada para ver si los datos fueron agrupados acorde a ideas preconcebidas o es necesario sugerir nuevos experimentos”.

Un algoritmo de agrupamiento intenta encontrar grupos naturales de datos basándose principalmente en la similitud y relaciones de los objetos, de forma tal que se obtenga una distribución interna del conjunto de datos mediante su particionamiento en grupos del conjunto de datos. Esta partición debe lograr la homogeneidad dentro de los grupos, i.e., los objetos que pertenecen al mismo grupo deben ser tan similares como se pueda, y la heterogeneidad entre grupos, i.e., los objetos que pertenecen a grupos diferentes deben ser tan diferentes como se pueda (Höppner, Klawonn et al. 1999, Kruse, Döring et al. 2007).

En otras palabras, el agrupamiento es basado en el principio de maximizar la similitud intra-grupo y minimizar la similitud inter-grupo. El concepto de “similitud” tiene que ser

especificado acorde a los datos. En la mayoría de los casos los datos son vectores de valores reales, entonces se requieren algunas medidas (distancias, similitudes, o disimilitudes) para cuantificar el grado de asociación entre ellos. Algunos algoritmos de agrupamiento tienen un requerimiento teórico para el uso de una medida específica, pero lo más común es que el investigador seleccione qué medida utilizará con determinado método. Obsérvese en el Anexo 1 las medidas más usadas para comparar objetos.

1.2 Validación del agrupamiento: Su clasificación

"El agrupamiento es un proceso subjetivo; el mismo conjunto de datos usualmente necesita ser agrupado de formas diferentes dependiendo de las aplicaciones" (Jain, Murty et al. 1999). Esta subjetividad hace el agrupamiento difícil, más aún su validación.

Una forma de evaluación del agrupamiento muy sencilla, puede ser, por ejemplo, mediante la visualización del conjunto de datos cuando éste es pequeño y los datos son bidimensionales. Sin embargo, esta forma de evaluación puede ser extremadamente difícil al intentar realizar una visualización efectiva de un conjunto de datos de alta dimensionalidad (Hansen and Johnson 2005). ¿Qué hacer cuando los datos no pueden representarse gráficamente, o es muy difícil o algunas veces imposible para un observador humano valorar el agrupamiento de los mismos; o no existe una forma simple de decidir si el resultado de un agrupamiento se ajusta a la división que deseamos de los datos?

Por otra parte, muchos algoritmos de agrupamiento varían sus resultados dependiendo de las características de los datos (por ejemplo, geometría y densidad de distribución) (Halkidi, Batistakis et al. 2001b). Otros dependen fuertemente de los valores asignados a los parámetros. Por ejemplo, si hay un parámetro que controle la resolución a la cual los datos son vistos, el algoritmo produce un dendograma en función de ese parámetro. En este caso es necesario decidir cuál nivel del dendograma refleja mejor los grupos según propiedades que se desea que el agrupamiento satisfaga. Algunos algoritmos necesitan que se especifique inicialmente el número de grupos a obtener, otros requieren que se especifique el número de vecinos de cada punto como un parámetro externo. Así, los resultados producidos son en función de los parámetros fijados y se hace necesario verificar cuáles se ajustan a los datos (Levine and Domany 2001).

Variaciones a partir de características de los datos, diferentes técnicas de análisis de grupos y definición de parámetros para el algoritmo a aplicar, indican que una evaluación de los resultados es necesaria para medir la calidad del agrupamiento. Una práctica común, en tal sentido, es aplicar medidas de validación de grupos (Stein, Eissen et al. 2003).

El procedimiento de evaluar los resultados de algoritmos de agrupamiento se conoce por validación del agrupamiento (Theodoridis and Koutroubas 1999, Halkidi, Batistakis et al. 2002). Se dice medida de validación de grupos a una función que hace corresponder a un agrupamiento un número real, indicando en qué grado el agrupamiento es correcto o no (Höppner, Klawonn et al. 1999). Estas medidas en su mayoría son heurísticas por naturaleza para conservar una complejidad computacional plausible.

Las medidas de evaluación del agrupamiento se clasifican en: globales y locales, subjetivas y objetivas, internas, externas y relativas, y supervisadas y no supervisadas (Höppner, Klawonn et al. 1999, Silberschatz and Tuzhilin 1996, Kaufman and Rousseeuw 1990).

Las medidas globales describen la calidad del resultado completo de un agrupamiento usando un único valor real, mientras que las locales evalúan cada grupo obtenido (Höppner, Klawonn et al. 1999). Las medidas objetivas miden propiedades estructurales de los resultados de los agrupamientos, por ejemplo, la separación entre los grupos y la compactación o densidad de los mismos (Halkidi, Batistakis et al. 2001b). La presencia de tales propiedades no garantiza que los resultados sean interesantes para el usuario, estas medidas carecen del enlace con los usuarios, aunque su principal atractivo es que son independientes del dominio (Silberschatz and Tuzhilin 1996). Las medidas subjetivas evalúan considerando la usabilidad de los grupos (Stein, Eissen et al. 2003). Las investigaciones en medidas subjetivas han sido menos intensas que las realizadas en las objetivas (Tuzhilin 2002).

Una clasificación muy usada divide la validación del agrupamiento en: medidas internas, externas y relativas (Theodoridis and Koutroubas 1999, Kaufman and Rousseeuw 1990). Estas últimas tienen un alto costo computacional (Halkidi, Batistakis et al. 2001a). Otra división consiste en medidas supervisadas y no supervisadas, refiriéndose a externas e internas, respectivamente. Las medidas externas se basan en un criterio externo, por ejemplo, una estructura previamente especificada que refleje la intuición que se tenga del agrupamiento de

los datos. No es posible aplicar estas medidas a situaciones del mundo real donde usualmente no está disponible una clasificación de referencia. Las medidas internas evalúan considerando solamente los resultados del agrupamiento en términos de cantidades que involucran los vectores de datos. Observe el Anexo 2 y el Anexo 3. Las medidas relativas se basan en la comparación del agrupamiento a evaluar con otros esquemas de agrupamiento o con resultados del mismo algoritmo con diferentes valores en los parámetros.

1.2.1 Principales medidas externas

Una medida externa es la entropía (Shannon 1948), la cual es una función de la distribución de las clases en los grupos resultantes. La entropía total para un conjunto de grupos es calculada como la suma de las entropías de cada grupo, ponderadas con el tamaño del grupo (Rosell, Kann et al. 2004b). En (Steinbach, Karypis et al. 2000) también usan la entropía como métrica de calidad, la mejor entropía es obtenida cuando cada grupo contiene exactamente un objeto. Otras expresiones para calcular la entropía por grupos y para el resultado de agrupamiento en general se presentan en (Zhao and Karypis 2003).

Algunas medidas externas usan las ideas de precision y recall del campo de la recuperación de información. Precision (Pr) y recall (Re) se calculan para un grupo j y una clase i dados usando las expresiones $\mathrm{Pr}(i, j)=n_{ij}/n_j$ y $\mathrm{Re}(i, j)=n_{ij}/n_i$, respectivamente. A partir de los valores de precision y recall se calcula la medida de evaluación *F*-measure. La influencia de precision y recall en su cálculo depende de un umbral α ($0\leq\alpha\leq1$) (Frakes and Baeza-Yates 1992). Un valor global, Overall *F*-measure, se calcula usando el promedio ponderado de los máximos valores de *F*-Measure por clases (Steinbach, Karypis et al. 2000). En (Larsen and Aone 1999) se propuso una variante de *F*-measure para un agrupamiento jerárquico. *F*-measure para una clase es el máximo valor de *F*-measure sobre todos los grupos a todos los niveles de la jerarquía. *F*-measure intenta capturar cuánto los grupos de la partición obtenida se hacen corresponder correctamente con los grupos de referencia (es decir, *F*-measure compara una jerarquía completa con una partición), mientras que la entropía compara particiones (obtenidas en un único nivel de jerarquía) con otra partición (Rosell, Kann et al. 2004a). Variantes de precision y recall, micro-averaged precision y micro-average recall, son utilizadas para evaluar el agrupamiento (Niu, Ji et al. 2004), las expresiones para su cálculo coinciden si cada objeto

pertenece a sólo un grupo y la clasificación de referencia también tiene una única clasificación para cada objeto.

En (Rosell, Kann et al. 2004a) se sugiere el uso del estadístico Kappa para evaluar la concordancia que existe entre dos particiones relativas a una partición de referencia, considerando las dos particiones obtenidas como intentos de clasificaciones siguiendo la referencia. La información mutua entre dos grupos, otra forma de evaluar, toma valores entre cero y el máximo valor de entropía de los grupos (alcanzado cuando ambos grupos son idénticos) (Xu and Gong 2004). Variantes normalizadas existen, pero no se proponen expresiones para valorar globalmente los resultados.

La medida error del agrupamiento utiliza el número de asociaciones incorrectas o ausentes para medir la cercanía que existe entre el resultado del agrupamiento y la clasificación de referencia (Roussinov and Chen 1999). Se considera asociación en una partición a un par de objetos que pertenezcan al mismo grupo, asociación incorrecta a aquella que existe en la partición de referencia y no en el resultado del agrupamiento y asociación ausente a aquella que existe en el resultado del agrupamiento y no en la partición de referencia. Esta medida favorece particiones pequeñas, por tanto una variante normalizada en el intervalo [0, 1] se presenta para proveer una menor dependencia del tamaño de ambas particiones. Siguiendo estas ideas, en (Roussinov and Chen 1999) se redefinen recall (cluster recall) y precision (cluster precision) para reflejar cuán bien el agrupamiento detectó las asociaciones entre los objetos y cuál es la precisión de las asociaciones detectadas, respectivamente.

La distancia Euclideana ha sido utilizada para medir la equivalencia estructural de dos grupos, siendo cero si los grupos son estructuralmente equivalentes y mayor si no lo son (Falkowski, Bartelheimer et al. 2006). El coeficiente de correlación también permite medir la equivalencia estructural de los grupos, mediante la división de la covarianza de la representación vectorial de los objetos por el producto de su desviación estándar. Este coeficiente toma valores entre -1 y +1 (los grupos son estructuralmente equivalentes).

Existen varias medidas basadas en la distribución de pares de objetos, entre ellas: estadístico Rand, coeficiente Jaccard, índice Folkes y Mallows, estadísticos Γ Huberts y Γ normalizado. Estas medidas usualmente se trabajan utilizando las técnicas de Monte Carlo (Theodoridis and

Koutroubas 1999). Esta técnica calcula una función de densidad probabilística de los índices estadísticos definidos. Así, evalúa el agrupamiento obtenido mediante su comparación con una partición de los datos construida por usuarios a partir de su intuición sobre la estructura de los mismos, o la comparación de la partición definida sobre los datos con la matriz de proximidad (Halkidi, Batistakis et al. 2001a). Estas medidas basadas en pruebas estadísticas generalmente tienen un alto costo computacional.

1.2.2 Principales medidas internas

Los algoritmos de agrupamiento generan una estructura espacial y se pueden definir medidas para diferentes aspectos de esta estructura, por ejemplo, densidad (Brun, Sima et al. 2007). Existen varios trabajos encaminados al desarrollo de medidas que validan el agrupamiento de una manera no supervisada. Algunos antiguos como el índice Goodman-Kruskal que tiene una alta complejidad computacional (Goodman and Kruskal 1954), el índice C apropiado cuando los grupos tienen tamaños similares (Hubert and Schultz 1976), los índices propuestos en (Akaike 1974, Schwartz 1978) utilizan criterios de información, seguidos por los propuestos en (Jain and Dubes 1988, Bock 1985). El cálculo de la dispersión intragrupo y la separación entre los grupos ha sido ampliamente trabajado, un ejemplo es el índice Calinski-Harabasz (Calinski and Arabas 1974), utilizado recientemente en (Maulik and Bandyopadhyay 2002) .

La cohesión de los grupos se puede usar como una medida de validación de éstos. Overall similarity es un índice interno que se ha utilizado para medir la cohesión basándose en la similitud de los pares de objetos en un grupo (Steinbach, Karypis et al. 2000).

Los índices para evaluar particiones generalmente se basan en alguna motivación geométrica para estimar cuán compactos y bien separados están los grupos. Un ejemplo son los índices Dunn (Dunn 1974) y sus generalizaciones (Bezdek and Pal 1995). Los índices Dunn varían en función de la medida de distancia entre grupos y el cálculo del diámetro del grupo que se utilice. Originalmente Dunn utilizó el mínimo de todas las distancias entre pares de elementos para calcular la distancia entre los grupos, y consideró el diámetro del grupo como la mayor distancia entre sus miembros (Dunn 1974). Así, las medidas tienden a producir valores elevados en agrupamientos con grupos compactos y muy bien separados.

Bezdek determinó que el índice Dunn es muy sensible al ruido (Bezdek and Pal 1995). Por ejemplo, la distancia entre un par de grupos puede ser menor que el diámetro de un grupo. Bezdek propuso una modificación en el cálculo de la distancia entre grupos mediante la estandarización respecto al tamaño de los mismos y una nueva forma de cálculo del diámetro del grupo mediante el cálculo de la distancia de todos sus elementos al centro del grupo, también estandarizado por su tamaño. Esta variante obtiene mejores resultados para diferentes dominios, pero nótese que se hace referencia a un centro de grupo, y no todos los algoritmos trabajan con prototipos, ni la estructura de todos los datos son grupos con forma esférica. Cinco generalizaciones de los índices Dunn para validar grupos con diferentes formas hiperesféricas y disminuir su sensibilidad al ruido fueron propuestas en (Bezdek and Pal 1998). Éstas abogan por definiciones apropiadas para el cálculo del diámetro de los grupos y la distancia entre los grupos, siguiendo el principio de que todos los datos deben estar explícitamente implicados en el cálculo del índice. Nótese el índice Dunn provee una estructura muy general para definir índices de validación, cada combinación de distancia y tamaño de los grupos define un nuevo índice. Una maximización de estos índices es desea. Ellos requieren una cantidad de tiempo considerable para su cálculo.

La medida Davies-Bouldin es basada en la idea que una buena partición es aquella con gran separación entre grupos y alta homogeneidad y compactación dentro de cada grupo. Esta medida es una proporción de la suma de la dispersión interna del grupo y la separación entre grupos. La dispersión dentro del grupo es relativa a los centroides de éstos y la distancia entre los grupos se basa en la distancia entre sus centros. Una dispersión baja y una distancia grande entre grupos tienden a producir valores bajos, por tanto una minimización de esta medida se desea (Davies and Bouldin 1979). Los índices Dunn y Davies-Bouldin son relativos al análisis geométrico de los grupos: típicamente centroidal y con forma esférica; elementos no presentes en todos los datos.

En (Pal and Biswas 1997) se generalizan los índices Dunn y Davies-Bouldin utilizando las estructuras de grafos GG, RNG y MST, donde los nodos son los objetos que fueron agrupados y las aristas pesadas entre los objetos indican la distancia que existe entre ellos. Se mostró que estas generalizaciones son más robustas a la presencia de ruido utilizando la distancia Euclideana entre los objetos a agrupar. En (Mali and Mitra 2003) transforman los índices

Dunn, Davies-Bouldin y el estadístico normalizado de Hubert a un marco de trabajo simbólico.

El índice I también sigue un esquema general similar a los índices Dunn, pero utiliza la distancia máxima entre grupos y adiciona las distancias (en lugar de promediarlas) multiplicadas por el número de grupos (Maulik and Bandyopadhyay 2002). Por otra parte, el índice v_{SV} calcula la suma pesada de las distancias entre grupos e intragrupos escogiendo la distancia mínima entre grupos y el promedio de las varianzas de los grupos como sus componentes, respectivamente (Kim and Park 2001).

Otras variantes de índices, en su mayoría con un alto costo computacional especialmente cuando el número de grupos y objetos es muy grande (Xie and Beni 1991), se han propuesto en (Dave 1996, Milligan and Cooper 1985). Una medida interna y global es el índice de separación (Höppner, Klawonn et al. 1999). Valores altos de éste indican buenas particiones. Para conjuntos de datos grandes, la determinación del índice de separación es computacionalmente muy costosa porque el número de operaciones para determinar los diámetros y las distancias de los grupos depende cuadráticamente del número de datos, en su defecto es posible utilizar la medida llamada granularidad-disimilitud (Xie, Raghavan et al. 2005). Esta medida es estable al evaluar la granularidad en resultados de agrupamientos basados en prototipos, existen extensiones para validar agrupamientos borrosos. El coeficiente de correlación Cohenetic permite medir el grado de similitud entre un dendograma producido por un algoritmo jerárquico y la matriz de proximidad. Valores del índice cercanos a cero indican que hay gran similitud entre las dos matrices y la técnica de Monte Carlo puede ser usada para la validación (Halkidi, Batistakis et al. 2001a).

En (Halkidi, Vazirgiannis et al. 2000) se presenta el índice de validación SD que suma el promedio de compactación de los grupos y la separación total entre ellos. Un valor pequeño para el primer término indica grupos compactos, su valor es directamente proporcional a la dispersión. El segundo término es sensible a la geometría de grupos e incrementa su valor al crecer el número de grupos, lo que hace que no pueda manipular adecuadamente grupos de formas arbitrarias. Además, requiere la existencia de centros de grupos. S_Dbw es otro índice que evalúa positivamente datos que formen grupos compactos y bien separados (Halkidi and Vazirgiannis 2001). Se basa en la compactación de los grupos (medida como la varianza de

los términos en cada grupo) y la densidad entre los grupos. No debe ser aplicado sobre grupos con formas no convexas, por ejemplo anillos.

Hasta aquí se han mencionado varios índices que calculan la razón entre las distancias intragrupos y las distancias entre grupos, por ejemplo: índices Dunn, Davies-Bouldin e índice I. Otros calculan la suma pesada de esas dos distancias, por ejemplo SD, S_Dbw y v_{SV}. En (Kim and Ramakrishna 2005) incluyen un análisis del diseño y funcionamiento de estos índices, y proponen modificaciones de los mismos a partir nuevas formas de cálculo de las distancias entre grupos e intragrupos para resolver las limitaciones encontradas.

En (Brun, Sima et al. 2007) hacen referencia al índice Silueta que es el promedio, sobre todos los grupos, del ancho de la silueta de sus puntos. Dos cálculos fundamentales intervienen en la silueta de un punto: la distancia promedio entre el punto y todos los otros puntos en el grupo, y el mínimo de la distancia promedio entre el punto y los puntos en otros grupos. Valores altos del índice Silueta global indican grupos más compactos y bien separados. El cálculo de este índice tiene una alta complejidad.

Los índices de validación: RMSSTD, SPR, RS y distancia entre dos grupos, deben ser usados simultáneamente para estimar el número de grupos existente en un conjunto de datos. Estos cuatro índices pueden ser aplicados a cada uno de los pasos de un algoritmo de agrupamiento jerárquico aglomerativo (Sharma 1996). RMSSTD mide la homogeneidad de los grupos formados en cada paso del algoritmo, SPR mide la pérdida de homogeneidad después de combinar dos grupos, RS mide el grado de diferencias entre grupos y la distancia entre grupos se calcula cuando dos grupos son combinados en un paso dado. Estos índices, esencialmente RMSSTD y RS, son usados también para evaluar resultados de agrupamientos no jerárquicos. Por otra parte, en (Jonnalagadda and Srinivasan 2004) proponen un método de validación basado en la distribución de los miembros de los grupos de una generación del agrupamiento a la próxima.

En (Yeung, Haynor et al. 2001) se propuso el método FOM y en (Olex, John et al. 2007) se proponen valores para sus parámetros. Se utiliza para estimar el número de grupos. Su principal limitante es que no puede aplicarse a datos con condiciones experimentales diferentes; además, su puntuación disminuye cuando el número de grupos aumenta.

En (Har-even and Brailovsky 1995) se presenta un método de agrupamiento jerárquico que utiliza dos medidas de validación interna durante la construcción de la jerarquía. Una de ellas compara el agrupamiento con los resultados de una hipótesis nula para todos los objetos y la otra escoge aleatoriamente un número de ejemplos y calcula fronteras estadísticas para determinar si se crearon muchos o pocos grupos.

Varias medidas se han desarrollado para medir la calidad de resultados de agrupamientos sobre datos representados gráficamente (Stein, Eissen et al. 2003, Newman 2003, Newman and Girvan 2004, Kannan, Vempala et al. 2001, Günter and Bunke 2003, Brás-Silva, Brito et al. 2006). Generalmente estas medidas asumen que cada objeto es un nodo del grafo y la ponderación de las aristas indica la similitud entre ellos. La no existencia de aristas representa pares altamente diferentes.

En (Kannan, Vempala et al. 2001) proponen una primera medida que calcula la expansión de un agrupamiento como la expasión mínima resultante de los árboles de expansión mínimos de cada subgrafo (grupo). La segunda utiliza la conductacia, donde subconjuntos de vértices son pesados para reflejar su importancia. Así, le dan más valor a vértices con muchos vecinos similares y le restan a aquellos con pocos vecinos. Como utilizan solamente el valor mínimo de expansión o conductancia entre todos los grupos, pueden evaluar un agrupamiento desfavorablemente cuando la mayoría de los grupos tengan alta calidad. La calidad mínima de los grupos y el peso total de las aristas que no están cubiertas por grupos son criterios para resolver este problema, pero su optimización es costosa.

En (Stein, Eissen et al. 2003) introducen las medidas conectividad parcial pesada Λ y densidad esperada ρ que también interpretan los datos como un grafo de similitud pesado. La maximización de estas medidas se desea. Tener una alta complejidad computacional es una de sus principales desventajas. Además, la medida Λ valida el agrupamiento considerando solamente la relación de los objetos intra-grupo (Stein, Eissen et al. 2003).

Una forma de medir la densidad en representaciones gráficas es mediante la conectividad interna del grupo, calculando la proporción del número de aristas dentro de él respecto al número posible de aristas. Por su parte, la cohesión se puede medir identificando cuán conectados están los miembros de un grupo a nodos que no son miembros de él mediante el

cociente: promedio de la interacción dentro del grupo entre promedio de la interacción de los nodos fuera del grupo. Otras medidas consideran información no-topológica adicional sobre la naturaleza de los grafos para evaluar resultados de métodos jerárquicos (Wilkinson and Huberman 2004, Newman and Girvan 2004, Radicchi, Castellano et al. 2004). Una propuesta específica para algoritmos basados en la intermediación y limitada al más bajo nivel de la estructura de los grupos se presenta en (Wilkinson and Huberman 2004). Otra se introdujo en (Radicchi, Castellano et al. 2004) donde se utilizan las definiciones grupo débil si la suma de todos los grados referentes a conexiones internas de los nodos que pertenecen al grupo es mayor que la suma de todos los grados de conexiones de éstos al resto del grafo y grupo fuerte si cada nodo que pertenece a él tiene más conexiones dentro del grupo que con el resto del grafo. Existen otras definiciones de grupo débil y fuerte (Wasserman and Faust 1994). La modularidad, utilizada para evaluar agrupamientos jerárquicos, mide la fortaleza de los grupos encontrados analizando las interconexiones antes y después del agrupamiento realizado (Newman 2003, Newman and Girvan 2004).

En (Brás-Silva, Brito et al. 2006) proponen el índice de tendencia del agrupamiento (clustering tendency index; IC) que parte de un grafo k-partito (asociado a los k grupos obtenidos por un algoritmo), donde dos vértices que pertenezcan a diferentes conjuntos son adyacentes si la disimilitud entre ellos es mayor que un umbral α. La idea del índice es contar el número de aristas que faltan en el grafo bipartito para que sea completo y sumar cada una de estas diferencias. Una variante normalizada también es presentada. Nótese que este índice depende de la definición de un umbral de corte y no tienen en cuenta las relaciones dentro de los grupos.

Un nuevo índice de validación que mide características geométricas de los datos y la separación entre los grupos es propuesto en (Lam and Yan 2007), basado en trabajos previos (Lam and Yan 2005b, Lam and Yan 2005a). Este índice trabaja bien para datos que contienen grupos con tamaños diferentes y cercanamente distribuidos.

El error de un agrupamiento puede definirse como la diferencia esperada entre sus etiquetas y etiquetas generadas, por ejemplo, con una distribución Gausiana. Las etiquetas son referidas a la asignación de las instancias a los grupos (Brun, Sima et al. 2007) .

Algunos métodos para evaluar se basan en la estabilidad de los grupos y controlan y alteran el ruido mediante el remuestreo del conjunto de datos original (Levine and Domany 2001, Borgelt and Kruse 2006). En (Lange, Braun et al. 2003) proponen una medida basada en la estabilidad del agrupamiento, pero tiene una alta complejidad computacional.

1.3 Medidas basadas en la teoría de conjuntos aproximados

A continuación se describirán medidas para la validación del agrupamiento basadas en la teoría de los conjuntos aproximados. Pero antes se detallarán los fundamentos teóricos sobre los conjuntos aproximados.

1.3.1 Fundamentos teóricos

La teoría de conjuntos aproximados fue introducida por Z. Pawlak en 1982 (Pawlak 1982). La descripción más general es que se basa en aproximar cualquier concepto, un subconjunto duro del dominio (por ejemplo, una clase de un problema de clasificación o un grupo resultante de un proceso de agrupamiento), por un par de conjuntos exactos, llamados aproximación inferior y superior del concepto aproximado.

Los conjuntos aproximados consideran que a todo objeto *x* de un universo *U* está asociada una cierta cantidad de información, expresada por medio de algunos atributos que describen el objeto (Komorowski, Pawlak et al. 1999, Bazan, Nguyen et al. 2004). La estructura de información básica de esta teoría es el sistema de información; par (*U, A*) donde $A = \{a_1, a_2, ..., a_m\}$ es el conjunto de atributos y *U* es un conjunto no vacío llamado universo de objetos descritos usando los atributos a_i (Komorowski, Pawlak et al. 1999)[1].

Los objetos que tienen la misma descripción son inseparables (similares) con respecto al conjunto de atributos considerados (*B*). Esta relación de inseparabilidad constituye la base matemática de la teoría, y la misma induce una partición del universo *U* en bloques de objetos inseparables (similares) (Komorowski, Pawlak et al. 1999). Cualquier subconjunto *X* (concepto) del universo *U* se puede expresar en términos de estos bloques de forma exacta o aproximada. La vaguedad es una propiedad de los conceptos y puede ser atribuida a los límites

[1] Esta definición es independiente a la definición de sistema de información de Shannon

del conjunto, mientras que la incertidumbre es una propiedad de los elementos del concepto y tiene que ver con la pertenencia o no a éste (Pawlak, Grzymala-Busse et al. 1995). Cuando un concepto es vago, los elementos del universo no pueden ser identificados con certeza como elementos del concepto.

Algunas extensiones de la teoría clásica de los conjuntos aproximados no requieren que se cumpla la transitividad ni la simetría, tales como las relaciones llamadas de tolerancia o similitud. La extensión de RST clásico a relaciones de similitud R'_B acepta que objetos que no son inseparables pero sí suficientemente cercanos o similares puedan pertenecer a la misma clase (Slowinski and Vanderpooten 1997). Varias medidas de similitud entre objetos o de comparación de atributos pueden ser utilizadas, obsérvese Anexo 1.

El objetivo es construir relaciones de similitud R'_B a partir de relaciones de inseparabilidad, relajando las condiciones iniciales de inseparabilidad. No obstante a la flexibilización, si R_B es una relación de inseparabilidad definida en U, R'_B es una relación de similitud extendida de R_B sí y sólo sí $\forall x \in U, \quad R_B(x) \subseteq R'_B(x)$ y $\forall x \in U$, $\forall y \in R'_B(x)$, $R_B(y) \subseteq R'_B(x)$, donde $R'_B(x)$ es la clase de similitud de x, es decir, $R'_B(x)=\{y \in U: yR'x\}$.

R'_B no tiene que ser necesariamente simétrica, aunque la mayoría de las definiciones de similitud usualmente lo son. No se impone, tampoco, que R'_B sea transitiva. El único requerimiento es la reflexividad. R'_B puede ser siempre vista como una extensión de la relación de inseparabilidad trivial R_B definida por $R_B(x) = \{x\}$, $\forall x \in U$. En RST extendida a relaciones de similitud, cada objeto puede pertenecer a más de una clase de similitud, por lo que el cubrimiento inducido por R'_B sobre U no es necesariamente una partición. En esta investigación es de interés considerar todos los atributos que caracterizan los objetos, ya que ellos fueron sometidos a un proceso de reducción de la dimensionalidad previo a la aplicación de esta teoría, así $B=A$ y se excluye B de la notación. Dos conceptos básicos fueron introducidos a partir de las relaciones de inseparabilidad: aproximaciones inferiores ($R'_*(X)$) y superiores ($R'^*(X)$) de un concepto X ($X \subseteq U$). Observe en las expresiones **(1.1)** y **(1.2)** sus cálculos a partir de relaciones de similitud.

$$R'_*(X) = \{x \in X : R'(x) \subseteq X\} \tag{1.1}$$

$$R'^{*}(X) = \bigcup_{x \in X} R'(x) \tag{1.2}$$

La región límite o frontera ($BN(X)$) de X para la relación R' se calcula considerando las expresiones (1.1) y (1.2) (Deogun, Raghavan et al. 1994). Observe la expresión (1.3).

$$BN(X) = R'^{*}(X) - R'_{*}(X) \tag{1.3}$$

Si el conjunto BN es vacío entonces el conjunto X es exacto respecto a la relación R'. En caso contrario, $BN(X) \neq \varnothing$, el conjunto X es inexacto o aproximado con respecto a R'.

Usar relaciones de similitud permite representar naturalmente varios problemas y además ofrece mayores posibilidades para la construcción de las aproximaciones; sin embargo, al trabajar en un espacio mayor, resulta más complejo computacionalmente buscar las aproximaciones relevantes (Pal and Skowron 1999).

1.3.2 Conjuntos aproximados para evaluar el agrupamiento

Aplicar RST a la evaluación del agrupamiento permite realizar una validación no supervisada y con bajo costo computacional. El cálculo común inicial de las relaciones y aproximaciones inferiores y superiores puede ser reutilizado por varias medidas de calidad, inclusión y proximidad de conceptos.

Objetos descritos por rasgos constituyen un sistema de información y adicionalmente se trabaja con conceptos definidos sobre ese sistema de información. Cada concepto X_i se corresponde con un grupo resultante de un proceso de agrupamiento al cual dichos objetos fueron sometidos. Obsérvese la Tabla . Sólo se considera en esta investigación resultados de agrupamientos duros y deterministas, donde los conceptos forman una partición.

	Rasgo_1	Rasgo_2	…	Rasgo_m
Objeto_1	Valor_{11}	Valor_{12}	…	Valor_{1m}
Objeto_2	Valor_{21}	Valor_{22}	…	Valor_{2m}
…	…	…	…	…
Objeto_n	Valor_{n1}	Valor_{n2}	…	Valor_{nm}

Tabla 1.1 Sistema de información.

Así, es posible calcular, para cada objeto agrupado, el conjunto de objetos relacionados con él, siguiendo la relación definida por la expresión (1.4), donde $s(x, y)$ retorna un valor de

similitud entre los objetos x e y, y ξ es el umbral de similitud que será considerado. La forma de medir la similitud y qué umbral utilizar para formar los conjuntos de relaciones depende del dominio donde fue aplicado el agrupamiento, cómo fueron descritos los objetos y qué nivel de granularidad se desea evaluar en los resultados. En el Anexo 4 se muestran posibles variantes para el cálculo del umbral. Adicionalmente, se pueden calcular las aproximaciones inferiores y superiores de cada grupo (concepto) usando (1.1) y **(1.2)**, respectivamente.

$$R'(x)=\{y\in U : yR'x, \text{ es decir } y \text{ está relacionado con } x \text{ si y sólo si } s(x,y)>\xi\} \qquad (1.4)$$

A partir del cálculo de las aproximaciones inferiores y superiores por grupos, se propone validar el agrupamiento y cada grupo mediante la aplicación de medidas ofrecidas por RST para evaluar los conceptos definidos sobre sistemas de información. Éstas permiten tener una noción de la proximidad de los conceptos y pueden ser aplicadas en varios esquemas de razonamiento (Zhong, Skowron et al. 1999). Trabajos previos de este enfoque se presentan en (Arco, Bello et al. 2006a, Arco, Bello et al. 2006c, Arco, Bello et al. 2006b).

Una medida que permite evaluar cada concepto es la precisión de la aproximación. Un concepto aproximado X puede ser caracterizado numéricamente por el coeficiente (1.5) llamado precisión de la aproximación, donde $|X|$ denota la cardinalidad de X, $X\neq\emptyset$. Obviamente, $0\leq\alpha(X)\leq 1$. Si $\alpha(X)=1$, X es duro (exacto), si $\alpha(X)<1$, X es aproximado (vago, inexacto), siempre respecto al conjunto de atributos considerado (Skowron 2000).

$$\alpha(X)=\frac{\left|R'_*(X)\right|}{\left|R'^*(X)\right|} \qquad (1.5)$$

La calidad de la aproximación es otra medida que permite evaluar conceptos. El coeficiente (1.6) expresa el porciento de objetos que pueden ser correctamente asignados a X. Además, $0 \leq\alpha(X) \leq \gamma(X) \leq 1$, y $\gamma(X)=0$ si $\alpha(X)=0$, mientras $\gamma(X)=1$ si $\alpha(X)=1$ (Skowron 2000).

$$\gamma(X)=\frac{\left|R'_*(X)\right|}{\left|X\right|} \qquad (1.6)$$

Las medidas precisión y calidad de la aproximación están asociadas a cada concepto, por tanto, ofrecen una valoración local de cada grupo obtenido. Sin embargo, en muchos casos es

necesario medir la calidad y la precisión como un todo considerando el sistema de información y los conceptos X_1, …, X_l definidos sobre él, con l total de conceptos. El coeficiente calidad del agrupamiento[2], expresión (1.7), describe la inexactitud de los conceptos, expresando la proporción de los objetos que pueden estar correctamente asignados a los grupos en el sistema. Si ese coeficiente es 1, el sistema de información según los conceptos definidos es consistente, en otro caso es inconsistente (Pawlak 1991).

$$\Gamma(DS) = \frac{\sum_{i=1}^{l} \left| R'_{*}(X_i) \right|}{\left| U \right|} \tag{1.7}$$

La precisión del agrupamiento[3] expresa las posibles asignaciones correctas a grupos. Su esencia es mostrar la proporción entre la cantidad de objetos que pudieran estar bien agrupados y la cantidad de objetos que pudieran o no pertenecer a los grupos del sistema de información (Liang, Shi et al. 2003). Observe la expresión (1.8).

$$\mathrm{A}(DS) = \frac{\sum_{i=1}^{l} \left| R'_{*}(X_i) \right|}{\sum_{i=1}^{l} \left| R'^{*}(X_i) \right|} \tag{1.8}$$

Si bien las medidas calidad y precisión del agrupamiento logran medir globalmente el nivel de inconsistencia, calidad y precisión de los conceptos en un sistema de información dado, consideran que cada grupo tiene igual repercusión en la evaluación. Sin embargo, no todos los grupos deben tener igual influencia al evaluar el agrupamiento, una ponderación de los mismos se desea. En esta investigación, inicialmente, se obtuvieron expresiones generalizadas de precisión y calidad del agrupamiento considerando la ponderación de los grupos por su cardinalidad, calculando el peso w_i asociado al grupo X_i como $w_i = \left|X_i\right| / \left|U\right|$ (Arco, Bello et al. 2006c, Arco, Bello et al. 2006a). En trabajos posteriores se consideraron otras formas de ponderación (Arco, Bello et al. 2006b, Caballero, Arco et al. 2007b). Investigaciones realizadas por miembros del Laboratorio de Inteligencia Artificial del Centro de Estudios de

[2] Nombrado en la literatura calidad de la aproximación de la clasificación o también calidad de la clasificación. En la literatura utilizan la palabra clasificación porque asumen que los conceptos coinciden con clases de un atributo de decisión. El uso en este trabajo es sobre los conceptos asociados a cada grupo resultante de un proceso de agrupamiento.
[3] Nombrado en la literatura precisión de la aproximación de la clasificación o también precisión de la clasificación.

Informática proponen la calidad y precisión generalizadas del agrupamiento, expresiones (1.9) y (1.10), respectivamente. El peso asociado a un grupo X_i se representa por w_i, cumpliéndose las restricciones $w_i \geq 0$ y $\sum_{i=1}^{l} w_i = 1$.

$$\Gamma_G(DS) = \frac{\sum_{i=1}^{l} \left(\left| R'_*(X_i) \right| \cdot w_i \right)}{\left| U \right|} \tag{1.9}$$

$$\mathrm{A}_G(DS) = \frac{\sum_{i=1}^{l} \left(\left| R'_*(X_i) \right| \cdot w_i \right)}{\sum_{i=1}^{l} \left(\left| R'^*(X_i) \right| \cdot w_i \right)} \tag{1.10}$$

Varios criterios pueden ser empleados para ponderar los grupos y así captar mejor propiedades deseadas (por ejemplo, similitud intragrupo, pertenencia de los objetos al grupo y cardinalidad del grupo).

Una forma de medir la pertenencia de un objeto a un grupo es la función de pertenencia aproximada. Ésta cuantifica el grado de solapamiento relativo entre $R'(x)$ y el concepto al cual x pertenece. Esta función se interpreta como una estimación basada en frecuencias de la probabilidad condicional de que el objeto y pertenezca al conjunto X, dados los valores del objeto x con respecto al conjunto de atributos. El valor $\mu_X(x)$ mide el grado de inclusión del objeto x en el grupo X (Grabowski 2004, Skowron 2000). Observe la expresión (1.11).

$$\mu_X(x) = \frac{\left| X \cap R'(x) \right|}{\left| R'(x) \right|} \tag{1.11}$$

La media de la pertenencia aproximada de los objetos a cada grupo puede también ser empleada para ponderar los grupos (Arco, Bello et al. 2006b, Caballero, Arco et al. 2007a, Caballero, Arco et al. 2007b, Caballero 2007). Observe la expresión (1.12). Sin embargo, esta ponderación puede fallar en algunos casos. Por tanto, investigaciones realizadas en el Laboratorio de Inteligencia Artificial proponen nuevas formas para el cálculo de la pertenencia aproximada de los objetos a los grupos.

$$w_i = \frac{\sum_{x \in X_i} \mu_{X_i}(x)}{|X_i|} \tag{1.12}$$

Miden la pertenencia aproximada de un objeto *x* a un grupo *X*, cuantificando en qué grado la clase de similitud de *x* (*R*'(*x*)) cubre el grupo *X*. Observe la expresión (1.13). En la expresión (1.14) se muestra una variante de ponderación utilizando el cálculo de la pertenencia según (1.13). Esta nueva propuesta ya ha sido utilizada en (Arco, Bello et al. 2006b, Caballero, Arco et al. 2007a, Caballero 2007) con el nombre función de compromiso aproximado.

$$\nu_X(x) = \frac{|X \cap R'(x)|}{|X|} \tag{1.13}$$

$$w_i = \frac{\sum_{x \in X_i} \nu_{X_i}(x)}{|X_i|} \tag{1.14}$$

Otra propuesta que permite calcular la pertenencia aproximada de los objetos a las clases se presenta en la expresión (1.15). Con la expresión (1.16) es posible ponderar los grupos.

$$\varpi_X(x) = \frac{|X \cap R'(x)|}{|X \cup R'(x)|} \tag{1.15}$$

$$w_i = \frac{\sum_{x \in X_i} \varpi_{X_i}(x)}{|X_i|} \tag{1.16}$$

Otras formas de ponderación son posibles. Pesar los grupos considerando su cohesión y densidad puede hacer la evaluación más precisa y cercana a la realidad. Una medida que se puede utilizar para ponderar la precisión y la calidad de los grupos es overall similarity (Steinbach, Karypis et al. 2000), o la varianza de las similitudes entre los objetos en los grupos (en esta medida una minimización es deseada), otras variantes también pueden ser consideradas.

A partir de los conceptos de RST a aplicar y las nuevas medidas definidas, las investigaciones en el tema han guían la aplicación de RST en la validación del agrupamiento mediante el Algoritmo 1.

Algoritmo 1. Aplicación de RST en la validación del agrupamiento

Entrada: Colección de objetos (sistema de información), resultado del agrupamiento (conceptos), medida y umbral de similitud, y formas de ponderación de los grupos.
Salida: Valores de las medidas de precisión y calidad aplicadas a los grupos y al agrupamiento en general.

1. Obtener las clases de similitud para cada objeto en el sistema de información; expresión (1.4).
2. Calcular las aproximaciones inferiores y superiores por grupo; expresiones (1.1), respectivamente.
3. Calcular la calidad y precisión por grupo; expresiones (1.5) y (1.6), respectivamente.
4. Calcular la calidad y precisión del agrupamiento; expresiones (1.7) y (1.8), respectivamente.
5. Para cada variante de cálculo de peso especificada
   a. Calcular los pesos por grupos.
   b. Calcular la calidad y precisión generalizadas del agrupamiento; expresiones (1.9) y (1.10), respectivamente.

De esta forma es posible medir la vaguedad o imprecisión de cada grupo obtenido y del agrupamiento en su totalidad. Si la región límite es pequeña, entonces se obtendrán mejores resultados de las medidas utilizadas en la evaluación (valores cercanos a 1). Valores altos de las medidas indican un mejor agrupamiento.

La complejidad computacional de esta propuesta es $O(mn^2)$, ya que calcular la matriz de similitud es $O(n^2)$, calcular los objetos relacionados con cada objeto es $O(n^2)$, porque calcular los objetos relacionados con un objeto específico es $O(n)$ y es necesario hacerlo para los n objetos del sistema de decisión y el cálculo de las aproximaciones inferiores y superiores por cada clase es $O(mn^2)$ en cada caso, donde m es el número de rasgos que describen los objetos, según estudios realizados en (Deogun, Choubey et al. 1998, Bell and Guan 1998).

1.4 Consideraciones finales del capitulo

Validar el agrupamiento es una etapa fundamental del post-agrupamiento. Una medida de validación no logra captar todas las buenas propiedades deseadas en un proceso de agrupamiento, por tanto, la aplicación de varias medidas es necesaria para lograr caracterizar correctamente el resultado a evaluar. Existen muchas medidas que permiten la validación

interna y externa del agrupamiento. Variantes de validación del agrupamiento utilizando conjuntos aproximados también fue incluida en este estudio del arte.

2 Incorporación de medidas de validación del agrupamiento en Weka

En este capítulo se da una breve introducción de Weka sobre cuando fue creado y de sus primeras versiones, que ventajas tienes, cuales son sus principales clases, que conjunto de paquetes que lo conforman, los diferentes tipos de ficheros con los que interactua y en el caso del *.arff* los tipos de datos que pueden tomar los atributos. También se hace una pequeña descripción de cómo funciona cada algoritmo de agrupamiento y el tipo de validación con que cuenta cada una para validar el resultado de un agrupamiento. Para finalizar se habla sobre el diseñó e implementación de los distintos tipos de medidas que seran utlilizadas para la validación del agrupamiento en Weka.

2.1 Weka generalidades

Según (González, 2005-2006). En el año 1992 Ian Witten, profesor del Departamento de Ciencia de la Computación de la Universidad de Waikato, Nueva Zelanda, creó una aplicación que más tarde, en 1993, recibiría el nombre de Weka, creándose con ello una interfaz e infraestructura para la misma. Su nombre: Weka (*Gallirallus australis*) se debe a un ave endémica de este país, de aspecto pardo y tamaño similar a una gallina, se encuentra en peligro de extinción y es famosa por su curiosidad y agresividad.

En 1994 fue terminada la primera versión de Weka, aunque no publicada, montada en una interfaz de usuario con varios algoritmos de aprendizaje escritos en lenguaje C. Muchas versiones de Weka fueron construyéndose hasta que en octubre de 1996 es publicada la primera versión (2.1).

En julio de 1997 se publica la versión 2.2 con la inclusión de nuevos algoritmos de aprendizaje y con la facilidad de configurar la corrida de una gran escala de experimentos. Cerca de esta fecha es tomada la decisión de rescribir Weka en lenguaje Java. A mediados de 1999 es liberada la versión 3 de Weka con todo su código implementado en Java. Después de la versión 3 se han publicado varias versiones, cada una de ellas ofreciendo esencialmente como mejora la adición de nuevos algoritmos.

La aplicación comenzada por Ian Witten en 1992 aún continúa ampliándose. El desarrollo de mejores extensiones al sistema ha dejado de concentrarse en el lugar donde fue creado, para dispersarse por todo el mundo, donde cantidad de científicos incluyen y validan sus propios modelos.

Este ambiente de Aprendizaje Automatizado está contenido por una extensa colección de algoritmos de la Inteligencia Artificial, útiles para ser aplicados sobre datos mediante las interfaces gráficas de usuario (GUI: Graphical User Interface) que ofrece o para usarlos dentro de cualquier aplicación. Contiene las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, agrupamiento, asociación y visualización.

2.1.1 Familiarización con el ambiente de Aprendizaje Automatizado Weka

Según (González, 2005-2006). Weka es un sistema multiplataforma y de amplio uso probado bajo sistemas operativos Linux, Windows y Macintosh. Puede ser usado desde la perspectiva de usuario mediante las seis interfaces que brinda, a través de la línea de comando desde donde se pueden invocar cada uno de los algoritmos incluidos en la herramienta como programas individuales y mediante la creación de un programa Java que llame a las funciones que se desee.

Weka (versión 3.5.2) dispone de seis interfaces de usuario diferentes que pueden ser accedidas mediante la ventana de selección de interfaces (GUI Chooser), que constituye la interfaz de usuario gráfica (GUI: Grafic User Interface) como se muestra en la Figura 2.1.

Interfaz para línea de comando (Simple CLI: Command Line Interface). Permite invocar desde la línea de comandos cada uno de los algoritmos incluidos en Weka como programas individuales. Los resultados se muestran únicamente en modo texto. A pesar de ser en apariencia muy simple es extremadamente potente porque permite realizar cualquier operación soportada por Weka de forma directa; no obstante, es muy complicada de manejar ya que es necesario un conocimiento completo de la aplicación. Su utilidad es pequeña desde que se fue recubriendo Weka con interfaces. Actualmente ya prácticamente sólo es útil como una herramienta de ayuda a la fase de pruebas. Es muy beneficiosa principalmente para los sistemas operativos que no proporcionan su propia interfaz para línea de comandos.

Explorador (Explorer). Interfaz de usuario gráfica para acceder a los algoritmos implementados en la herramienta para realizar el aprendizaje automatizado. Es el modo más usado y descriptivo. Permite realizar operaciones sobre un sólo archivo de datos.

Experimentador (Experimenter). Facilita la realización de experimentos en lotes, incluso con diferentes algoritmos y varios conjuntos de datos a la vez.

Flujo de conocimiento (KnowledgeFlow). Proporciona una interfaz netamente gráfica para el trabajo con los algoritmos centrales de Weka. Esencialmente tiene las mismas funciones del Explorador aunque algunas de ellas aún no están disponibles. El usuario puede seleccionar los componentes de Weka de una barra de herramientas, y conectarlos juntos para formar un "flujo del conocimiento" que permitirá procesar y analizar datos.

Visualizador de Arff (ArffViewer). Interfaz para la edición de ficheros con extensión arff.

Log. Muestra la traza de la máquina virtual de acuerdo a la ejecución del programa.

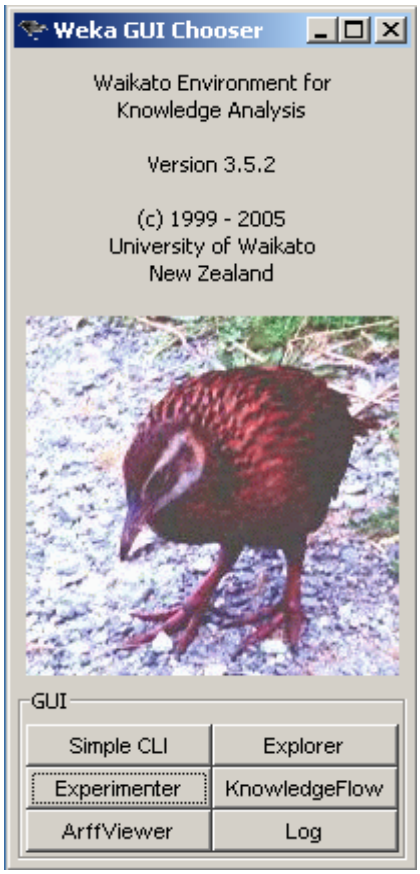

Figura 2.1 **Ventana de selección de interfaces (GUI Chooser)**

2.1.2 Entrada de datos

Según (González, 2005-2006). Weka denomina a cada uno de los casos proporcionados en el conjunto de datos de entrada instancias, cada una de las cuales posee propiedades o rasgos que las definen. Los rasgos presentes en cada conjunto de datos son llamados atributos.

El formato de archivos con el que trabaja Weka es denominado *arff*, acrónimo de **A**ttribute-**R**elation **F**ile **F**ormat. Este formato está compuesto por una estructura claramente diferenciada en tres partes:

Sección de encabezamiento: se define el nombre de la relación.

Sección de declaración de atributos: se declaran los atributos a utilizar especificando su tipo. Los tipos aceptados por la herramienta son:

a) Numérico (*Numeric*): expresa números reales.

b) Entero (*Integer*): expresa números enteros.

c) Fecha (*Date*): expresa fechas.

d) Cadena (*String*): expresa cadenas de texto, con las restricciones del tipo String.

e) Enumerado: expresa entre llaves y separados por comas los posibles valores (caracteres o cadenas de caracteres) que puede tomar el atributo.

Sección de datos: se declaran los datos que componen la relación

A continuación se especifica la sintaxis a utilizar para declarar cada una de estas secciones que forman el fichero con extensión *arff*.

El formato del encabezamiento es el siguiente:

@relation <nombre-de-la-relación>, donde: <nombre-de-la-relación> es de tipo cadena. Si dicho nombre contiene algún espacio será necesario expresarlo entrecomillado.

La sintaxis para declarar los atributos es:

@attribute <nombre-del-atributo> <tipo>, donde: <nombre-del-atributo> es de tipo cadena, y al igual que el nombre de la relación, si hay algún espacio, es necesario poner comillas.

La sección de datos se encabeza con @data <conjunto-de-datos>, donde en <conjunto-de datos> se especifican todas las instancias; separando los valores de los atributos para una misma instancia entre comas y las instancias (relaciones entre los atributos) con saltos de línea. En el caso de que algún dato sea desconocido se expresará con un símbolo "?". Es posible añadir comentarios con el símbolo " %", que indicará que desde ese símbolo hasta el final de la línea es todo un comentario. Los comentarios pueden situarse en cualquier lugar del fichero.

El formato por defecto de los ficheros que usa Weka es el *arff,* pero eso no significa que sea el único que admita. Esta herramienta tiene intérpretes de otros formatos como CSV y C4.5. Además, las instancias pueden leerse también de un URL (Uniform Resource Locator) o de una base de datos en SQL usando JDBC.

2.1.3 Interioridades de Weka

Según (González, 2005-2006). Weka se implementa en Java, lenguaje de alto nivel ampliamente difundido y multiplataforma, lo que posibilita el Weka puedan ejecutarse en cualquier plataforma sobre la que haya una máquina virtual Java disponible. Esta herramienta sigue los preceptos del código abierto (open source), por lo su código fuente está totalmente disponible, permitiendo la modificación del mismo. Solo es necesario recompilarlo para posteriormente agregar extensiones al sistema. Además es un software de distribución gratuita lo que posibilita su uso, copia, estudio, modificación y redistribución sin restricciones de licencias.

Las clases de Weka están organizadas en paquetes. Un paquete es la agrupación de clases e interfaces donde lo habitual es que las clases que lo formen estén relacionadas y se ubiquen en un mismo directorio. Esta organización de la estructura de Weka hace que añadir, eliminar o modificar elementos no sea una tarea compleja. Está formado por 10 paquetes globales, y dentro de ellos se agrupan otros paquetes que aunque su contenido se ajusta al paquete padre ayudan a organizar aun mejor la estructura de clases e interfaces.

Los paquetes globales son:

- "*associations*": contiene las clases que implementan los algoritmos de asociación.

- "*attributeSelection*": contiene las clases que implementan técnicas de selección de atributos.
- "*classifiers*": agrupa todas las clases que implementan algoritmos de clasificación y estas a su vez se organizan en subpaquetes de acuerdo al tipo de clasificador.
- "*clusterers*": contiene las clases que implementan algoritmos de agrupamiento.
- "*core*": paquete central que contiene las clases controladoras del sistema.
- "*datagenerators*": paquete que contiene clases útiles en la generación de conjuntos de datos atendiendo al tipo de algoritmo que será usado.
- "*estimators*": clases que realizan estimaciones (generalmente probabilísticas) sobre los datos.
- "*experiment*": contiene las clases controladoras relacionadas con el modo Experimentador.
- "*filters*": está constituido por las clases que implementan algoritmos de preprocesamiento.
- "*gui*": contiene todas las clases que implementan los paneles de interacción con el usuario, agrupadas en subpaquetes correspondientes a cada una de las interfaces.

El paquete "*core*" constituye el centro del sistema Weka. Es usado en la mayoría de las clases existentes. Las clases principales del paquete "core" son: *Attribute*, *Instante*, e *Instances*. Mediante un objeto de la clase *Attribute* podremos representar un atributo. Su contenido es el nombre, el tipo y en el caso de que sea un atributo nominal, los posibles valores.

Los métodos más usados de la clase *Attribute* son:

- *enumerateValues*: retorna una enumeración de todos los valores de un atributo si es de tipo nominal, cadena, o una relación de atributos, o valor nulo en otro caso.
- *index*: retorna el índice de un atributo.
- Los métodos *isNominal, isNumeric, isRelationValued, isString, isDate* retornan verdadero si el atributo es del tipo especificado en el nombre del método.
- *name*: retorna el nombre del atributo.

- *numValues*: retorna el número de valores del atributo. Si este no es nominal, cadena o relación de valores retorna cero.
- *toString*: retorna la descripción de un atributo de la misma manera en que puede ser declarado en un archivo .arff.
- *value*: retorna el valor de los atributos nominales o de cadena.

La clase *Instance* se utiliza para manejar una instancia. Un objeto de esta clase contiene el valor del atributo de la instancia en particular. Todos los valores (numérico, nominal o cadena) son internamente almacenados como números en punto flotante. Si un atributo es nominal (o cadena), se almacena el valor del índice del correspondiente valor nominal (o cadena), en la definición del atributo. Se escoge esta representación a favor de una elegante programación orientada a objetos, incrementando la velocidad de procesamiento y reduciendo el consumo de memoria. Esta clase es *serializable* por lo que los objetos pueden ser volcados directamente sobre un fichero y también cargados de uno. En Java un objeto es serializable cuando su contenido (datos) y su estructura se transforman en una secuencia de bytes al ser almacenado. Esto hace que los objetos puedan ser enviados por algún flujo de datos con comodidad. Un objeto de la clase *Instances* contiene un conjunto ordenado de instancias, lo que conforma un conjunto de datos. Las instancias son almacenadas, al igual que la clase Instante, como números reales, incluso las nominales, que como se explicó anteriormente, utilizando como índice la declaración del atributo e indexándolas según este orden.

Los métodos más útiles de la clase *Instance* son:

- *classAttribute*: devuelve la clase de la que procede el atributo.
- *classValue*: devuelve el valor de la clase del atributo.
- *value*: devuelve el valor del atributo que ocupa una posición determinada.
- *enumerateAttributes*: devuelve una enumeración de los atributos que contiene esa instancia.
- *weight*: devuelve el peso de una instancia en concreto.

Los métodos más útiles de la clase *Instances*:

- *numInstances*: devuelve el número de instancias que contiene el conjunto de datos.
- *instance*: devuelve la instancia que ocupa la posición i.
- *enumerateInstances*: devuelve una enumeración de las instancias que posee el conjunto de datos.
- *attribute*: existen dos métodos con este nombre, la diferencia es que uno recibe como parámetro el índice del atributo, y el otro el nombre, ambos retornan el atributo.
- *enumerateAttributes*: retorna una enumeración de todos los atributos.
- *numAttributes*: retorna el número de atributos como un entero.
- *attributeStats*: calcula estadísticos de los valores un atributo especificado.

Las implementaciones de los esquemas reales de aprendizaje son el recurso más valioso que Weka proporciona.

El paquete *filters* contiene las clases que implementan los algoritmos de reprocesamiento de datos presentes en Weka. Agrupados en subpaquetes según su tipo.

Una de las características más interesantes de Weka es la posibilidad de modificar su código y obtener versiones adaptadas con funcionalidades que no contengan las versiones oficiales, ampliando así sus posibilidades de uso. Desde el inicio se diseñó como una herramienta orientada a la extensibilidad, por lo que se creó una estructura factible para ello.

El diseño e implementación de las clases que componen el sistema se conciben para que la adición de un nuevo algoritmo sea una tarea relativamente fácil. Para realizarla no es necesario tener en cuenta detalles que se relacionan indirectamente con la implementación del algoritmo en cuestión, tales como: la lectura del fichero de datos, el cálculo de valores estadísticos, entre otros.

2.1.4 Factibilidad de la implementación de nuevos modelos en Weka

Según (González, 2005-2006). La extensibilidad de Weka hace que sea más factible adicionar un nuevo algoritmo a esta herramienta, y no implementarlo como un programa independiente, a los efectos de su validación. Algunas razones son:

- La implementación de un algoritmo se simplifica utilizando Weka. Implementar un nuevo modelo consiste en redefinir métodos ya existentes en Weka o crear otros nuevos con las funcionalidades específicas del algoritmo a adicionar.
- No se requiere programar la interfaz gráfica de usuario para utilizar el nuevo algoritmo.
- No se necesita programar lo referente a la entrada de los datos. Se reutilizaría la manera de hacerlo en Weka, que está independiente a la implementación de los algoritmos.
- Se facilita la validación del algoritmo implementado. No es necesario preocuparse por implementar las variantes de validación, ni las medidas de desempeño a emplear; se reutilizan las existentes en la herramienta.
- Se propicia y facilita la comparación del nuevo algoritmo con otros ya reportados en la literatura e implementados en la herramienta, facilitando el análisis de la factibilidad de este último, algo que sería más costoso en tiempo si se hubiera implementado como un modelo aislado.
- Se facilita la generalización de un nuevo algoritmo. Por ejemplo, un nuevo criterio para calcular distancia pudiera ser fácilmente validado con los algoritmos implementados.
- Facilita probar un conjunto de datos con todos los algoritmos disponibles en la herramienta a los efectos de seleccionar el adecuado.
- El tiempo de desarrollo del prototipo de un software a la medida utilizando un algoritmo implementado en Weka disminuye a partir de reutilizar su código.
- Es posible hacer corridas en lotes y en varias terminales, sin esfuerzos adicionales de programación. Utilizando el modo Experimentador que Weka tiene implementado se pueden realizar experimentos con varios modelos y conjuntos de datos a la vez, utilizando varias terminales; lo cual es muy recomendable para algoritmos que consuman cantidad de tiempo de corrida. Los resultados se almacenan en ficheros, para su posterior análisis estadístico.

- Se propicia el uso y divulgación de los nuevos modelos implementados. El hecho de queden incorporados a Weka los hace disponibles para la comunidad de científicos y usuarios de este campo.
- Facilita el preprocesamiento de los datos. Un algoritmo implementado en Weka pudiera requerir previamente una transformación a los datos originales almacenados en el fichero con extensión *.arff*. La herramienta cuenta con una serie de algoritmos de preprocesamiento (filtros) implementados a tales efectos.

Nótese que el hecho de que los filtros estén separados de los algoritmos que usan los datos es una ventaja. De esta manera, se facilita la implementación de un nuevo modelo, si este requiere realizar un tipo de preprocesamiento ya implementado entre los tantos filtros disponibles en la herramienta.

2.2 Algoritmos de agrupamiento en Weka

Los algoritmos de agrupamiento en Weka se encuentran dentro del paquete *clusterers*, estos heredan de la superclase abstracta *Clusterer*. Además, en este paquete se encuetra la clase *ClusterEvaluation*, que es la encargada de la evaluación de un algoritmo de agrupamiento en específico. Esta clase incluye la función *evaluateClusterer* que dado un conjunto de datos (instancias) devuelve el resultado del agrupamiento para éste. Para ello, esta función llama a la función *clusterInstance* de la clase *Clusterer*, la cual devuelve el número del grupo al que pertenece una instancia del conjunto de datos.

A continuación se describirán los algoritmos de agrupamiento incluidos en Weka y que fueron utilizados en este trabajo.

2.2.1 Cobweb

Según (Garre, 2005). Cobweb es un algoritmo de agrupamiento jerárquico. Se caracteriza porque utiliza aprendizaje incremental, esto es, realiza los agrupamientos instancia a instancia y sigue una estrategia jerárquica. Durante la ejecución del algoritmo se forma un árbol donde las hojas representan los objetos o grupos de objetos y el nodo raíz engloba por completo el conjunto de datos de entrada. Al principio, el árbol consiste en un único nodo raíz. Las instancias se van añadiendo una a una y el árbol se va actualizando en cada paso. La

actualización consiste en encontrar el mejor sitio donde incluir la nueva instancia, operación que puede necesitar de la reestructuración de todo el árbol (incluyendo la generación de un nuevo nodo anfitrión para la instancia y/o la fusión/partición de nodos existentes) o simplemente la inclusión de la instancia en un nodo que ya existía. La clave para saber cómo y dónde se debe actualizar el árbol la proporciona una medida denominada *utilidad de categoría*, que mide la calidad general de una partición de instancias en un segmento. La reestructuración que mayor utilidad de categoría proporcione es la que se adopta en ese paso. El algoritmo es muy sensible a otros dos parámetros:

a) Acuity: Este parámetro es muy necesario, ya que la utilidad de categoría se basa en una estimación de la media y la desviación estándar del valor de los atributos, pero cuando se estima la desviación estándar del valor de un atributo para un nodo en particular, el resultado es cero si dicho nodo sólo contiene una instancia. Así pues, el parámetro acuity representa la medida de error de un nodo con una sola instancia, es decir, establece la varianza mínima de un atributo.

b) Cut-off: Este valor se utiliza para evitar el crecimiento desmesurado del número de segmentos. Indica el grado de mejoría que se debe producir en la utilidad de categoría para que la instancia sea tenida en cuenta de manera individual. En otras palabras: cuando no es suficiente el incremento de la utilidad de categoría en el momento en el que se añade un nuevo nodo, ese nodo se corta, conteniendo la instancia otro nodo ya existente.

Además, Cobweb pertenece a los métodos de aprendizaje conceptual o basado en modelos. Esto significa que cada cluster se considera como un modelo que puede describirse intrinsicamente, más que un ente formado por una colección de puntos.

Al algoritmo COBWEB no hay que proporcionarle el número exacto de clusters que queremos, sino que en base a los parámetros anteriormente mencionados encuentra el número óptimo.

2.2.2 DBScan

Los principales algoritmos basados en densidad son: el algoritmo para el agrupamiento espacil basado en densidad para aplicaciones con ruido (Density-Based Spatial Clustering of Applications with Noise; DBSCAN) (Ester, Kriegel et al. 1996) y el algoritmo basado en

densidad (DENsity-based CLUstEring; DENCLUE) (Hinneburg and Keim 1998). Ambos tienen una complejidad $O(n \log n)$, con n número de nodos. No funcionan correctamente con datos de alta dimensionalidad y dependen altamente de los parámetros iniciales.

DBSCAN trata con datos ruidosos y un grupo se define como un conjunto maximal de puntos densamente conectados. Los grupos son identificados mediante la detección de la densidad de los puntos. Regiones con alta densidad de puntos describen la existencia de grupos mientras que regiones con una baja densidad de puntos indican grupos ruidosos o puntos fuera de la curva. Este algoritmo es particularmente usado para agrupar grandes conjuntos de datos y es capaz de identificar grupos con diferentes tamaños y formas.

La idea clave de DBSCAN es, para cada punto de un grupo, la vecindad de un radio dado tiene que contener al menos un número mínimo de puntos, tal que, la densidad en la vecindad no excesa algún umbral predefinido. Este algoritmo necesita tres parámetros de entrada: el tamaño de la vecindad, el radio que delimita el área de la vecindad de un punto y el número mínimo de puntos que pueden existir en la vecindad.

2.2.3 EM

Expectación-maximización (Expectation Maximization; EM) (Bradley, Fayyad et al. 1998) asigna a cada instancia una distribución de probabilidad de pertenencia a cada cluster. Aunque manipula datos de alta dimensionalidad, realiza un refinamiento muy costoso.

El algoritmo EM puede decidir cuántos clusters crear basado en validación cruzada o se le puede especificar a priori cuantos debe generar. Utiliza el modelo Gaussiano finito de mezclas, asumiendo que todos los atributos son variables aleatorias independientes.

EM se usa en estadística para encontrar estimadores de parámetros de máxima verosimilitud en modelos probabilísticos que dependen de variables no observables. El algoritmo EM alterna pasos de expectación (paso E), donde se computa la expectación de la verosimilitud mediante la inclusión de variables latentes como si fueran observables, y un paso de maximización (paso M), donde se computan estimadores de máxima verosimilitud de los parámetros mediante la maximización de la verosimilitud esperada del paso E. Los parámetros que se encuentran en el paso M se usan para comenzar el paso E siguiente, y así el proceso se repite.

- **Expectation:** Utiliza los valores de los parámetros, iniciales o proporcionados por el paso *Maximization* de la iteración anterior, obteniendo diferentes formas de la FDP (Función de Densidad de Probabilidad) buscada.
- **Maximization:** Obtiene nuevos valores de los parámetros a partir de los datos proporcionados por el paso anterior.

2.2.4 FarthestFirst

El algoritmo Primero el más lejano (Farthest First) mejora el *k*-medias (Hochbaum and Shmoys 1985). Parte de una selección aleatoria de los centros de grupos, calcula la distancia de cada instancia al centroide más cercano, y la instancia que quede más lejana del centroide más cercano es seleccionada como el centroide de un grupo. Este proceso es repetido hasta que el número de grupos sea mayor que un umbral especificado. Este algoritmo es utilizado en el agrupamiento de documentos (Liu, Cai et al. 2006) .

Este algoritmo parte seleccionando aleatoriamente una instancia que será el centro del grupo. Se calcula la distancia entre cada instancia restante y su centro más cercano. La instancia que más alejada está del centro es seleccionada como un centro de grupo. Este proceso es repetido hasta que el número de grupos sea mayor que un umbral especificado. Ésta es posiblemente la mejor heurística para el problema de los *k*-centros.

2.2.5 OPTICS

Ordenamiento de puntos para identificar la estructura del agrupamiento (Ordering Points To Identify the Clustering Structure; OPTICS) (Ankerst, Breunig et al. 1996) es una variante mejorada de DBSCAN. Produce un orden especial del conjunto de datos con relación a su estructura de agrupamiento basada en densidad. Este método es bueno tanto para el análisis de grupos automático e interactivo, incluyendo el hallazgo de la estructura intrínseca de los grupos.

2.2.6 K-Means

K-medias (k-means) (MacQueen, 1967) es uno de los algoritmos más sencillos de aprendizaje no supervisado. Éste es uno de los algoritmos que crean particiones más usado. *k*-medias que

tiene una complejidad temporal $O(Ikn)$[4] (Jain and Dubes 1988, Kaufman and Rousseeuw 1990, McQueen 1967, Xiong, Wu et al. 2006). Este algoritmo no funciona bien con grupos que no tengan forma convexa y requiere que el número de grupos a obtener sea especificado a priori, por tanto requiere un cierto conocimiento del dominio, ya que es sensible a cómo se hizo inicialmente la partición. A partir de él se han derivado varios como el *x*-medias (*x*-means) para una estimación eficiente del número de grupos, el conjunto *k*-medias (batch *k*-means) y el *k*-medias incremental (incremental *k*-means), y su variante mejorada medias (means) (Berry 2004).

Este algoritmo se caracteriza por su sencillez. En primer lugar se debe especificar por adelantado cuantos grupos se van a crear, éste es el parámetro *k*, para lo cual se seleccionan *k* elementos aleatoriamente, que representarán el centro o media de cada grupo.

A continuación cada una de las instancias, ejemplos, es asignada al centro del grupo más cercano de acuerdo con la distancia que le separa de él. Si se utilizaran similitudes, se asignara al centro con mayor similitud.

Para cada uno de los grupos así construidos se calcula su centro y estos centros son tomados como los nuevos centros de sus respectivos grupos. Finalmente se repite el proceso completo con los nuevos centros de los grupos. La iteración continúa hasta cierto criterio de convergencia.

El algoritmo Simple k-means en Weka utiliza el método *k*-medias para el agrupamiento. XMeans extiende *k*-medias con una estimación eficiente del número de grupos.

2.3 Validación del agrupamiento en Weka

En el desarrollo de este trabajo existieron dos razones fundamentales para incluir medidas o índices de validación del agrupamiento en Weka. Una de ellas es comprobar que las medidas de RST son un buen instrumento para la validación del agrupamiento y por tanto, establecer las bases para el desarrollo de experimentos que pudieran comparar el desempeño de las medidas basadas en RST y medidas clásicas a partir del resultado de varios algoritmos de agrupameinto. La otra, es agregar a Weka un conjunto de medidas que sirvan para validar el

[4] Se utiliza *n* número de objetos y *k* número de grupos.

agurpamiento, ya el Weka ofrece muy pocas facilidades para evaluar resultados de agrupamientos. De manera general, Weka solo ofrece la matriz de confusión y el porciento de instancias bien y mal clasificadas al concluir cada agrupamiento. Esta forma de evaluación es primitiva y requiere de la clasificación de referencia, por tanto son medidas externas. Cada algoritmo en específico adiciona algún otro elemento que contribuye a la validación:

EM:

- Log likelihood: es el resultado de calcular la suma de la densidad de cada instancia entre la suma de las instancias que pertenecen a cada grupo.

SimpleKMeans:

- Media o moda de cada atributo por grupo.
- Desviación estandar de cada atributo por grupo.

XMeans:

- Distorción: es la suma por grupos de la suma de la distancia de cada instancia dentro de un grupo con su centro.

2.4 Diseño e implementación en Weka de las medidas de validación

A continuación se describirá cómo fueron incluídas en Weka las medidas internas y externas que permiten enriquecer la validación del agrupamiento. Se especificará cómo las medidas se integran al diseño ya establecido en Weka, y aquellos elementos adicionados debido a los requerimientos de las medidas implementadas.

2.4.1 Diagrama de clases

Se diseñaron e implementaron seis clases, tres relacionadas con las medidas (*MedidasInternas*, *MedidasExternas* y *RST*) y tres relacionadas con las distancias (*Jaccard*, *Dice* y *Cosine*). Estas últimas implementan la interfaz *DistanceFunction* de Weka. Estas clases se relacionan con la clase *ClusterEvaluation* de Weka mediante una relacion de asociación. Observe en la Figura 2.2 que se interactua con las salidas de los algoritmos de agrupamiento mediante la clase *Clusterer*, ya que todos los algoritmos considerados en el estudio heredan de esta clase.

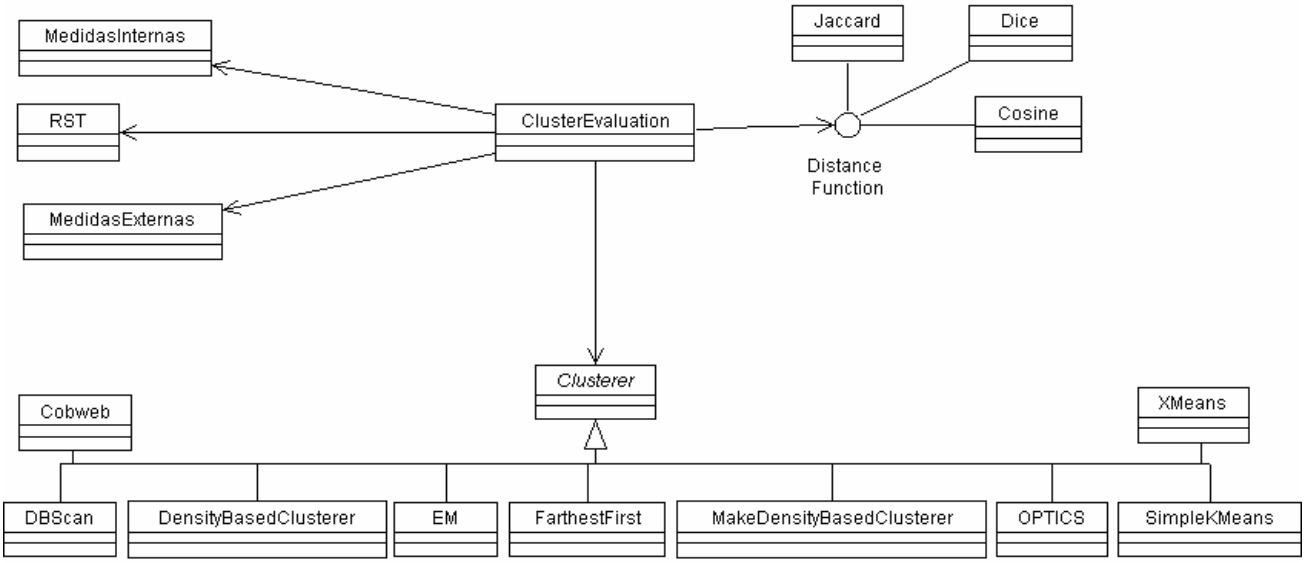

Figura 2.2 Diagrama de clases.

De las medidas implementadas, el índice de Bezdek (*IndicesBezdek_Dunn*) y la medida de Davies-Bouldin (*IndicesDavies_Bouldin*) requieren trabajar con los centros de cada grupo para validar el agrupamiento. Sin embargo, hay algoritmos de agrupamiento incorporados en Weka que no devuelven centros de los grupos, porque éstos no constituyen parte de su procesamiento. Por eso, en el diseño se consideró utilizar los centros de grupos proporcionados por los algoritmos de agrupamiento que los calculen y calcularlos para aquellos que no los devuelvan. El centro de cada grupo se calculó como el vector promedio de las instancias que lo forman. Así, fue incorporado en *ClusterEvaluation* (ver Figura 2.3) el método *CrearCentro*, que permite crear los centros cuando éstos no son proporcionados por los algoritmos. La distancia de cada instancia al centro del grupo a que ella pertenece se almacena en la diagonal principal de la matriz de distancias. La matriz de distancias es una matriz simétrica (*Matrix*) que fue incluida como un atributo en *ClusterEvaluation*. Esta matriz se calcula una sola vez y es reutilizada por cada una de las medidas internas implementadas.

Las salidas de los resultados del agrupamiento en Weka no favorecen el procesamiento necesario para el cálculo de las medidas implementadas. Por un lado, Weka guarda el arreglo *m_clusterAssignments* donde por cada instancia se especifica a que grupo fue asignada. Por el otro, *instanceStats* es un arreglo donde el Weka especifica una descripción de los grupos, con datos como el número de instancias incluidas en ellos. Por tal motivo, en este trabajo se ha

incluido la matriz *Grupos* en *ClusterEvaluation* que permite almacenar y acceder rápidamente a los miembros de cada grupo. Cada fila de esta matriz representa un grupo y las columnas aquellas instancias miembros del grupo.

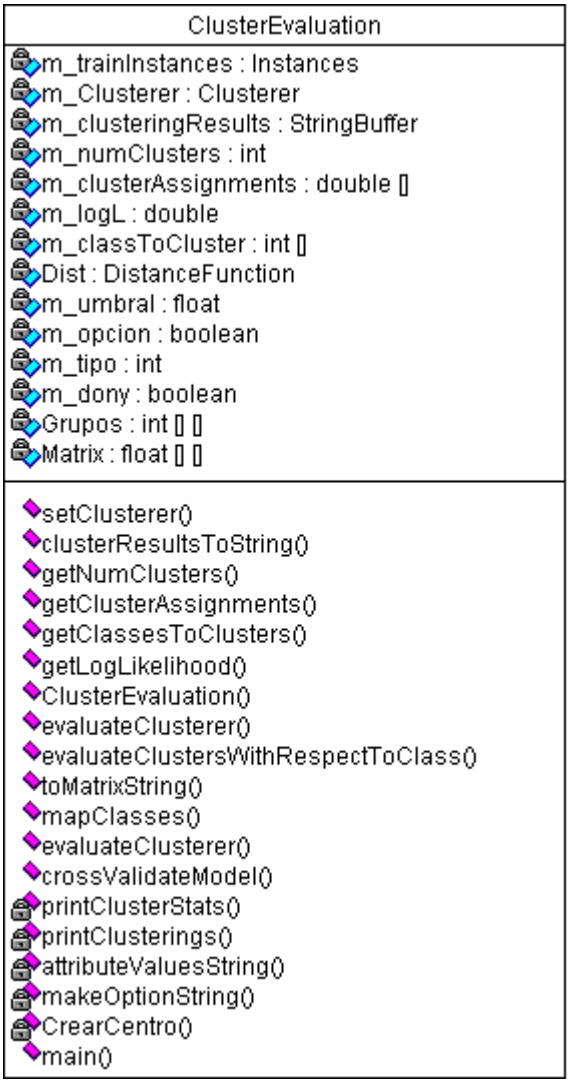

Figura 2.3 Clase ClusterEvaluation

2.4.2 Implementación de las clases

Observe en la Figura 2.4 la implementación de las clases *MedidasInternas*, *MedidasExternas* y *RST*. Se desarrollaron estas tres clases, porque la naturaleza de las medidas internas difiere de las medidas externas porque estas últimas requieren una clasificación de referencia y las primeras no. Las medidas basadas en RST son también medidas internas, pero llevan un tratamiento especial para el desarrollo de la Teoría de los Conjuntos Aproximados.

- **Medidas Internas**

Esta clase implementa algunas de las principales medidas internas referidas en la literatura para la validación del agrupamiento. Las medidas incluidas son: Overall Similarity (*OverallSimilarity*), índice de Dunn (*IndicesDunn*), índice de Bezdek (*IndicesBezdek_Dunn*), medida de Davies-Bouldin (*IndicesDavies_Bouldin*), la densidad parcial pesada (*DensidadParcialPesadaA*) y densidad esperada (*DensidadEsperadaP*). Estas medidas no necesitan una clasificación de referencia externa, por tanto, para su procesamiento no se considera el atributo clase del archivo *.arff*. Estas medidas sólo basan sus cálculos en la relación que existe entre los atributos predictores y los grupos resultantes de un proceso de agrupamiento.

Estas medidas fueron implementadas y agregadas al Weka dentro de la clase *MedidasInternas.java*, incluida en el *clusterers*.

- **Medidas externas**

Esta clase implementa algunas de las principales medidas externas referidas en la literatura para la validación del agrupamiento. Las medidas incluidas son: Entropía (*Entropia*) y Overall F-Measure (*OverallF_Measure*). La medida Overall F-Measure depende del umbral alfa para ponderar Precision y Recall. Siempre se publican los resultados globales de Precision (alfa = 1), Recall (alfa = 0) y Overall F-Measure con igual ponderación para precision y recall (alfa = 0.5).

Estas medidas utilizan una clasificación de referencia externa. Por tal motivo ellas requieren el uso del atributo clase del archivo *.arff* para realizar sus cálculos. Este atributo aparece la mayoría de las veces al final de cada caso para una mayor comodidad. Weka toma por defecto el último atributo, aunque es posible especificar qué atributo es la clase. En Weka estas se encuentran dentro de una clase llamada *MedidasExternas.java* la cual se encuentra dentro del paquete *clusterers*.

- **Medidas basadas en teoría de los conjuntos aproximados (RST).**

Las medidas basadas en RST forman parte de las medidas internas ya que éstas no necesitan de una clasificación de referencia externa en el archivo *.arff* para realizar sus cálculos.

Esta clase implementa algunas de las principales medidas basadas en RST para la validación interna del agrupamiento. Las medidas globales incluidas son: precisión del agrupamiento (*PrecisionCalif*), calidad del agrupamiento (*CalidadCalif*), precisión generalizada del agrupamiento (*PrecisionClasifGeneralizada*) y calidad generalizada del agrupamiento (*CalidadClasifGeneralizada*). Las medidas locales por grupos incluidas son: precision aproximada (*PrecisionAprox*) y calidad aproximada (*CalidadAprox*). Adicionalmente, se ofertan cinco variantes para ponderar las medidas globales precisión y calidad generalizadas. Se incluyeron tres formas de peso propias de RST a partir del cálculo de la precisión aproximada de las objetos a los grupos: *MediaPertenenciaAprox1*, *MediaPertenenciaAprox2* y *MediaPertenenciaAprox3*. Además, el diseño permite la ponderación por la cardinalidad de los grupos y por los resultados de Overall Similarity por grupo.

Observe en la que las medidas basadas en RST parten de cálculos que son comunes a todas. La clase RST calcula todos los objetos relacionados con cada objeto del conjunto de instancias y las aproximaciones inferior y superior de cada grupo. Se implementó RST extendido a relaciones de similitud, por tanto, es posible especificar qué umbral de similitud será utilizado para construir las relaciones.

En Weka estas se encuentran dentro de la clase *RST.java* perteneciente al paquete *clusterers* para que luego pueda ser reutilizada por otras clases.

Figura 2.4 Clases MedidasInternas, MedidasExternas y RST.

- **Distancias**

Las nuevas funciones de distancia implementadas son: Jaccard, Dice y Coseno, las cuales se encuentra cada una dentro de una clase con su mismo nombre e implementan la interfaz *DistanceFuction* de Weka y se diferencian entre sí sólo en el algoritmo de distancia que usan. Observe la Figura 2.5. Todas las medidas internas utilizan estas distancias u otras ya incorporadas en Weka.

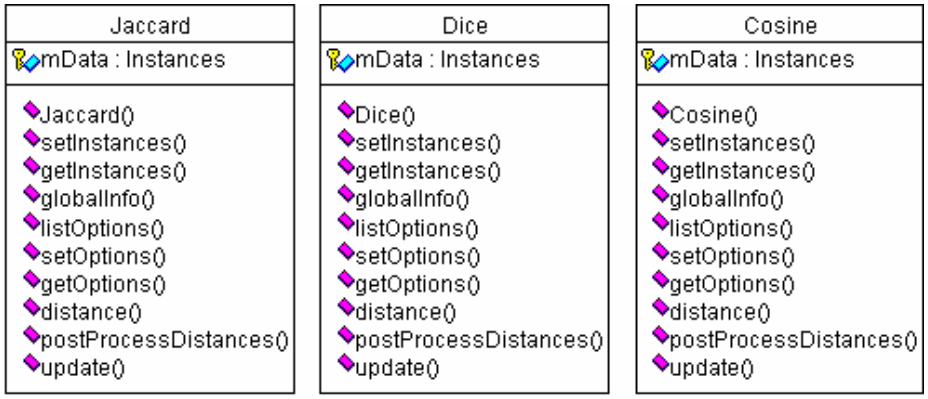

Figura 2.5 Distancias incorporadas al Weka.

2.4.3 Interfaz de usuarios

Observe en la Figura 2.6 las adecuaciones realizadas en la interfaz del Weka prevista para el agrupamiento y su validación. En la parte visual de Weka, específicamente en la pestaña → *Cluster*, es donde se encuentran los algoritmos de agrupamiento y los diferentes modos de agrupar los datos. Después de seleccionado un algoritmo de agrupamiento, es posible especificar cómo trabajar con el conjunto de datos (*use training set, supplied test set*, *percentage split*) y si se van a utilizar las clases para la validación del agrupamiento o no, y qué atributo considerarlo como clase (*classes to clusters evaluation*). Esta última opción (*classes to clusters evaluation*) es necesario seleccionarla cuando se quieren ver los resultados de las medidas externas, las medidas internas siempre se publican porque sólo consideran grupos e instancias en su procesamiento y no requieren de la especificación del atributo clase.

En esta pestaña, *cluster*, se añadieron otras componentes necesarias para especificar los valores de los parámetros requeridos por las medidas de validación:

- Cuadro de chequeo (CheckBox) → *Validaty indices* para que el usuario elija si quiere o no aplicar las medidas de validación al algoritmo seleccionado.
- Cuadro combinado (ComboBox) → *Distance measures* para que el usuario elija qué tipo de medidas de distancia (*Jaccard*, *Dice*, *Cosine*, *Euclidean*, *Heterogeneous Euclidean*) quiere aplicar.
- Cuadro combinado → *Criteria for computing threshold* para que el usuario escoja una de las funciones (*Mean of distances*, *Mean of min distances*, *Mean of max distances*, *Combination mean&max distances*) que tiene para calcular el umbral para obtener las relaciones de cada objeto en RST.
- Cuadro de textos (TextField) → *Threshold for RST* para que el usuario especifique un umbral por defecto. En este caso, si el usuario entra un valor, la opción anterior no se realiza (no se calcula el umbral mediante la función seleccionada), porque esta tiene prioridad y en caso de que no entre un valor se realiza opción previa.

Observe en el panel de la derecha de la Figura 2.6 como se muestran los resultados de las medidas internas y externas para un agrupamiento realizado. Las formas de validación del Weka se mantienen para cada algoritmo.

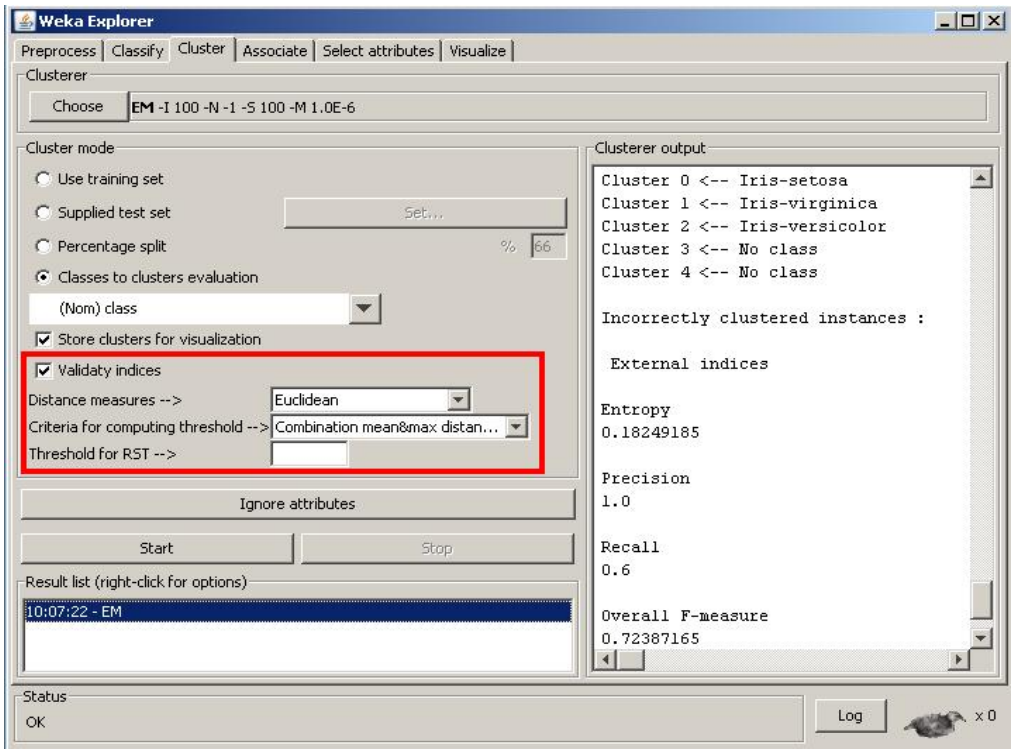

Figura 2.6 Interfaz del agrupamiento y su validación en Weka.

2.5 Conclusiones parciales

Se diseñó e implementó un módulo de evaluación del agrupamiento en Weka, ya que Weka, no tienen incorporado un buen módulo de evaluación, por tanto no existe una herramienta que integre algoritmos de agrupamiento y varias medidas de forma tal que se puedan estudiar las principales medidas de validación del agrupamiento bajo las mismas condiciones. En dicho módulo se añadieron un conjunto de medidas para chequear la validación del agrupamiento, estas medidas se clasifican en: medidas internas y externas, dentro de las internas se incluyen las medidas basadas en RST propuestas en el Laboratorio de IA.

3 Resultados experimentales

El objetivo de este capítulo es utilizar los resultados del capítulo 2 para evaluar las medidas de validación basadas en RST, utilizando varios algoritmos de agrupamiento de Weka y las medidas internas y externas que como resultado de esta tesis fueron incorporadas.

La evaluación de medidas de validación del agrupamiento es una tarea ardua. Para chequear la confiabilidad y validez de las medidas basadas en RST se han diseñado experimentos, aplicados posteriormente a los casos de estudio que se definen, que permiten un análisis sobre la evaluación de los conjuntos aproximados como instrumento de medición.

3.1 Definición de casos de estudio y herramientas utilizadas

El uso de RST para validar mediante la determinación local y global de la precisión, calidad e inconsistencia de los resultados de agrupamientos es posible en diversos dominios. Es por eso que se definieron casos de estudio que recopilan bases provenientes de diferentes áreas de aplicación.

El primer caso de estudio definido constituye una recopilación de bases de casos internacionales donde aparecen objetos de diversa naturaleza descritos por rasgos en su mayoría numéricos, aunque también hay presencia de rasgos simbólicos en algunos de ellos. La selección, para este primer caso de estudio, ha incluido bases con la presencia o no de valores ausentes y no ha sido un requerimiento la presencia de rasgos objetivos. Obsérvese en la Tabla 3.1 la descripción y la fuente de cada una de las bases de casos que conforman el primer caso de estudio. El conjunto de bases de casos descritos en la Tabla 3.2 constituye el segundo caso de estudio definido. Todas estas bases constan de rasgos predictores y objetivos, por tanto existe la clasificación de referencia para cada una de ellas.

Weka permite considerar los resultados de varios algoritmos de agrupamiento y utilizarlos en el estudio experimental. La similitud utilizada en Weka para aplicar las medidas basadas en RST fue el dual de la distancia HEOM. Obsérvese el Anexo 1.

Las medidas externas consideradas en el estudio son: entropía (Rosell, Kann et al. 2004a, Steinbach, Karypis et al. 2000, Zhao and Karypis 2003), precision (P), recall (R) y OFM (Frakes and Baeza-Yates 1992). Medidas internas consideradas en el estudio son: overall

similarity (OS) (Steinbach, Karypis et al. 2000), los índices Dunn (DD) (Dunn 1974) y su generalización (DB) (Bezdek and Pal 1995), la medida Davies-Bouldin (IDB) (Davies and Bouldin 1979), la conectividad parcial pesada (CPP) y la densidad esperada (DE) (Stein, Eissen et al. 2003).

Los algoritmos EM, FarthestFirst, DBSCAN, SimpleKMeans (algoritmo *k*-medias) y XMeans[5], incluidos en Weka, fueron aplicados a los casos de estudio.

3.2 Diseño y aplicación de experimentos

El enfoque para validar basado en RST, instrumento de medición a evaluar, se basa en el Algoritmo 1, que incluye la aplicación de las medidas de RST. Las tres variantes de cálculo de la media de la pertenencia de los objetos a los grupos, expresiones (1.12), (1.14) y (1.16), se consideran pesos en las expresiones de calidad y precisión generalizadas. Otras dos ponderaciones que se incluyen son: la cardinalidad normalizada de los grupos y la medida overall similarity por grupo. Al interpretar las tablas resultantes de los experimentos considere la notación siguiente: precisión aproximada (PA), calidad aproximada (CA), precisión generalizada (PGRM1, PGRM2, PGRM3, PGC y PGOS) y calidad generalizada (CGRM1, CGRM2, CGRM3, CGC y CGOS) con ponderaciones según (1.12), (1.14), (1.16), cardinalidad normalizada y overall similarity por grupo, respectivamente.

La Figura 3.1 muestra el esquema utilizado para chequear la confiabilidad y validez del instrumento (Grau, Correa et al. 2004, Sampieri 2007, Sampieri, Collado et al. 2006).

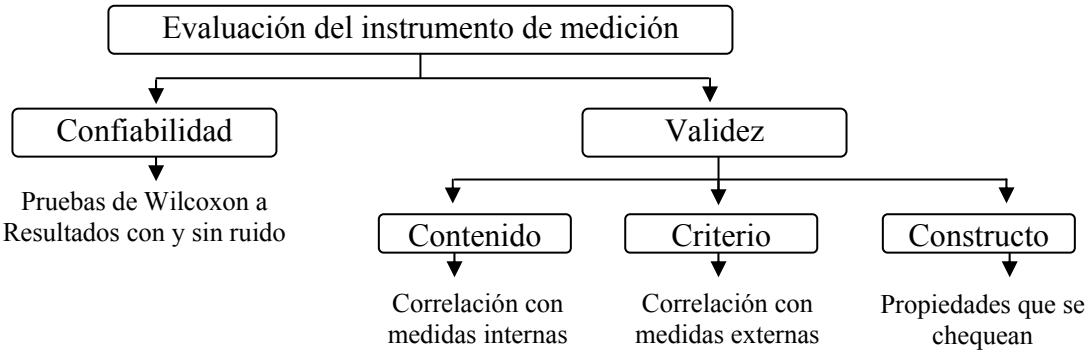

Figura 3.1 Esquema general para evaluar la propuesta de validación de agrupamiento basada en RST.

[5] Se han utilizados los mismos identificadores que utiliza Weka para referirse a los algoritmos

3.2.1 Medir confiabilidad

La confiabilidad se midió comparando los valores arrojados por el instrumento de medición a los resultados del agrupamiento con y sin ruido. Todos los casos recopilados en los dos primeros casos de estudio fueron utilizados en este experimento. Cada base de casos original fue agrupada mediante los algoritmos incluidos en Weka: EM, FarthestFirst, DBSCAN, SimpleKMeans y XMeans. Cada resultado de agrupamientos fue evaluado con las medidas internas basadas en RST. Así, ya se tenían los valores de las medidas para cada agrupamiento aplicado a cada base sin ruido. Posteriormente, se introdujo ruido a cada base de casos. Se modificaron un 5%, 10%, 15%, 20% y 25% de los valores de los rasgos que describen los objetos en cada base. La alteración, también aleatoria, fue de a lo sumo un 10% del valor original del rasgo. A cada una de las bases con ruido introducido desde un 5% hasta un 25%, se le aplicaron los algoritmos de agrupamiento ya mencionados y los resultados fueron evaluados utilizando las medidas internas basadas en RST. Así, ya se tienen resultados de las medidas para cada base sin ruido y con varios niveles de ruido después de aplicado cada algoritmo de agrupamiento. La prueba no paramétrica de Wilcoxon[6] fue aplicada entre los valores de las medidas resultantes de agrupamientos con y sin ruido; sus valores de significación se muestran en el Anexo 5. Obsérvese que incluso alterando hasta un 25% de los valores de las bases de casos, los valores de significación siempre son superiores a 0.05, indicando que no existen diferencias significativas entre las poblaciones comparadas horizontalmente (medidas aplicadas sobre bases con y sin ruido). Es importante señalar que los resultados de las medidas están incluso influenciados por la sensibilidad al ruido de cada uno de los métodos de agrupamiento utilizados. Esto evidencia que, para las bases consideradas, las medidas son confiables porque produjeron valores similares en condiciones similares, el ruido no cambió los resultados.

3.2.2 Medir validez

Para medir la validez hay que tener evidencias de la validez de contenido, de criterio y de constructo (Grau, Correa et al. 2004). La primera se refiere al grado en que el instrumento refleja un dominio específico del contenido de lo que mide. La segunda compara los

[6] Se utilizó SPSS 13.0 para Windows

resultados del instrumento con un criterio externo. La última, se refiere al grado en que una medición se relaciona consistentemente con otras mediciones de acuerdo con hipótesis derivadas teóricamente y que conciernen a los conceptos o constructos.

3.2.2.1 Validez de contenido

Se verifica que el instrumento al menos cubre las propiedades que las medidas internas utilizadas en el estudio cubren, considerando todos los casos recopilados en los dos primeros casos de estudio. Cada base de casos fue agrupada mediante los algoritmos ya mencionados incluidos en Weka. Se utilizaron estos algoritmos porque consideran en su mayoría prototipos y forman grupos compactos y bien separados, lográndose de esta forma que las medidas internas implementadas tengan un buen comportamiento en la evaluación. Se aplicó el método de correlación de Pearson entre los resultados de las medidas basadas en RST y las medidas internas seleccionadas. Los resultados de la correlación se muestran en el Anexo 6 donde cada fila corresponde a las medidas internas referenciadas y cada columna a las medidas basadas en RST propuestas. La primera subfila corresponde al coeficiente de correlación y la segunda muestra la significación de la correlación entre cada par de medida referenciada y propuesta. No todas las medidas basadas en RST correlacionan con todas las medidas internas utilizadas en el análisis, pero se observa que las medidas basadas en RST en su conjunto logran cubrir todas las propiedades que cubren las medidas ya publicadas, para las bases de casos consideradas. Las medidas basadas en RST tienen un cálculo común inicial, sin embargo, las medidas citadas requieren estructuras y cálculos diversos y en la mayoría de los casos, costosos. En la

Tabla A6.2 se observa que ninguna de las medidas basadas en RST es capaz de correlacionar con overall similarity y con conectividad parcial pesada, incluso en el resto de las tablas del Anexo 6 pocas medidas basadas en RST son capaces de correlacionar con estas dos medidas internas referenciadas. Esta situación se debe esencialmente a que estas dos medidas sólo tienen en cuenta el comportamiento interno de cada grupo a evaluar pero no consideran en su análisis la integración entre los grupos y por tanto, pueden dar una valoración positiva del agrupamiento cuando en realidad no lo es. Las medidas basadas en RST siempre tienen en cuenta la integración de los grupos en el análisis.

3.2.2.2 Validez de criterio

El criterio externo que se ha utilizado en esta evaluación del instrumento es el resultado producido por las principales medidas externas referenciadas. Por tanto, sólo ha sido posible utilizar el segundo caso de estudio, para él existe la clasificación de referencia. Las bases de este caso de estudio fueron sometidas a cada uno de los algoritmos de agrupamiento

seleccionados de Weka. A los resultados de los agrupamientos se les aplicaron las medidas basadas en RST y las medidas externas referenciadas. El objetivo de este experimento es verificar, para los casos considerados, que las medidas basadas en RST logran tener un comportamiento similar a las medidas externas, sin requerir conocimiento adicional de los datos a agrupar. Por tanto, se sugiere su utilidad en la práctica donde no existan clasificaciones de referencia. Para ello, al igual que al validar contenido, se ha aplicado el método de correlación de Pearson entre los resultados de validación de la nueva propuesta y las medidas externas referenciadas. El Anexo 6 muestra los resultados de la correlación. Aquí cada fila corresponde a las medidas externas referenciadas. No todas las medidas basadas en RST correlacionan con todas las medidas externas utilizadas en el análisis, igual situación ocurrida en la comparación anterior; pero sí es posible observar que todas las medidas externas correlacionan con al menos una de las medidas basadas en RST. Esto muestra que, para las bases consideradas, las medidas basadas en RST tienen un comportamiento similar a las medidas externas aplicadas sin utilizar una clasificación humana de referencia.

3.2.2.3 Validez de constructo

Se han propuesto dos criterios principales para la evaluación del agrupamiento (Halkidi, Batistakis et al. 2001b): la compactación de los grupos y la separación entre ellos; constituyendo dos de los constructos principales en esta validación. Los elementos de la propuesta encaminados a chequear la compactación de los grupos son las expresiones que permiten la ponderación de las variantes generalizadas: las tres formas de cálculo de la pertenencia aproximada y overall similarity. Todas las variantes de precisión y calidad aproximadas chequean la separación entre los grupos, y sus variantes ponderadas logran chequear ambos constructos. Así lo muestran las correlaciones realizadas entre los resultados de las medidas basadas en RST y la selección de medidas internas y externas tomadas de la literatura científica. Observe el Anexo 6 y el Anexo 6. Adicionalmente, las medidas basadas en RST permiten medir la precisión y calidad de los grupos, así como el nivel de inconsistencia de los mismos. Estos resultados se logran por la propia definición de las medidas.

Los experimentos realizados no sólo han permitido mostrar que las medidas son un instrumento adecuado para validar el agrupamiento, sino que ha permitido reducir el número

de medidas a considerar en la evaluación. Note en el Anexo 6 y el Anexo 6 que la ponderación con la cardinalidad de los grupos, overall similarity y la tercera expresión de la pertenencia aproximada son las medidas que reflejan las mejores correlaciones; indicando de esta forma que este subconjunto de medidas debe ser el más utilizado para validar agrupamientos utilizando RST.

Después de los experimentos realizados es posible preguntarse: ¿es esta nueva forma de evaluación mejor que las anteriores? No existe una medida que sea válida para todos los tipos de agrupamientos y que pueda captar todas sus propiedades, por tanto, es mejor preguntarse ¿cuáles son las ventajas de estas medidas respecto a las anteriores? Con este nuevo enfoque para la validación es posible:

- Calcular el nivel de inconsistencia, la precisión y calidad local de cada grupo, y globalmente de los grupos respecto al sistema de información dado.
- Validar agrupamientos con independencia de la forma de los grupos resultantes.
- Incluir en la validación el análisis de propiedades estructurales de los grupos.
- Variar la granularidad con que se desea evaluar el agrupamiento mediante la especificación del umbral de similitud entre los objetos.
- Brindar una variedad de medidas que requieren un cálculo inicial común para todas y el cálculo de las particularidades de ellas es poco costoso.
- Validar sin tener en consideración centros de los grupos, por tanto, se puede evaluar resultados de un mayor número de tipos de agrupamientos.
- Utilizar similitudes asimétricas entre los objetos, en aplicaciones donde sea requerido.
- Considerar elementos como pertenencia de los objetos a los grupos y cardinalidad de los grupos en el proceso de evaluación.
- Utilizar en la validación la misma función de similitud entre objetos que fue utilizada en el agrupamiento.

3.3 Descripción de los archivos utilizados para evaluar las medidas basadas en RST

Tabla 3.1 Descripción de los archivos sin clasificación de referencia

No.	Nombre del archivo	Cantidad de instancias	Cantidad de rasgos		Valores ausentes
			Numéricos	Simbólicos	
Conjuntos de datos para técnicas de agrupamiento publicados por la Universidad de Köln http://www.uni-koeln.de/themen/statistik/data/cluster					
1	Achieve	25	4	0	No
2	Birth	70	2	0	No
3	Dentitio	66	8	0	No
4	Milk	25	4	0	No
5	Nutrient	27	5	0	No
Conjunto de datos para técnicas de aprendizaje automático y descubrimiento del conocimiento en bases de datos, publicados por la Universidad de California, Irvine. http://archive.ics.uci.edu/ml y http://kdd.ics.uci.edu					
6	AutoPrice	159	16	0	No
7	Baskball	96	5	0	No
8	Bodyfat	252	15	0	No
9	Bolts	40	8	0	No
10	Colesterol*	297	9	5	No
11	Cleveland*	297	8	6	No
12	Cloud*	108	4	0	No
13	Cpu*	209	7	0	No
14	Detroit	13	14	0	No
15	EchoMonths	130	7	3	Si
16	Elusage	55	2	1	No
17	Fishcatch	158	6	2	Si
18	Fruitfly	125	3	2	No
19	Gascons	27	5	0	No
20	Housing	506	13	1	No
21	Longley	16	7	0	No
22	Lowbwt	189	5	5	No
23	Mbagrade	61	2	1	No
24	Pbc*	312	12	7	Si
25	Pollution	60	16	0	No
26	PwLinear	200	11	0	No
27	Quake	2178	4	0	No
28	Schlyote	37	5	1	No
29	Sleep	62	8	0	Si
30	Strike	625	6	1	No
31	Veteran	137	4	4	No
32	Vineyard*	52	3	0	No

Archivos de datos del libro DATA por Andrews y Herzberg http://lib.stat.cmu.edu/datasets/Andrews/					
33	T05.1a	73	34	0	No
34	T05.1b	296	8	0	No
35	T06.1a	40	9	0	No
36	T06.1b	40	9	0	No
37	T06.2	125	12	0	No
38	T08.1	190	3	0	No
39	T09.1	109	3	2	No
40	T10.1	72	12	0	No
41	T11.1	235	12	0	No
42	T12.1	39	12	0	No
43	T14.1	732	8	0	No
44	T15.1	135	61	0	No
45	T16.1	60	12	0	No
46	T17.1	127	14	0	No
47	T21.1	33	7	0	No
48	T28.1	39	8	0	No
49	T28.2	53	8	0	No
50	T28.3	122	8	0	No
51	T30.1	19	6	0	No
52	T33.1	100	5	0	No
53	T33.2	11	13	0	No
54	T35.1	53	12	0	No
55	T35.2	52	6	0	No
56	T36.1	145	5	0	No
57	T38.1	209	9	0	Si
58	T40.2	18	5	0	Si
59	T41.1	10	14	0	No
60	T44.1	77	6	0	No
61	T47.1	96	0	24	No
62	T47.2	96	0	24	No
63	T48.1a	500	16	0	No
64	T48.1b	500	16	0	No
65	T48.1c	500	16	0	No
66	T48.1d	500	16	0	No
67	T48.1e	500	16	0	No
68	T48.1f	500	16	0	No
69	T48.1g	679	16	0	No
70	T48.3a	499	16	0	No
71	T48.3b	500	16	0	No
72	T48.3c	543	16	0	No
73	T49.1	30	7	0	Si
74	T50.1	64	16	0	No

75	T53.1	302	9	0	Si
76	T59.1	42	8	0	No
77	T60.1	95	4	0	No
78	T62.1	48	5	0	No
79	T64.1	20	12	0	No
80	T65.1	34	12	0	No
81	T65.2	34	12	0	No
82	T65.3	34	12	0	No
83	T65.4	41	8	0	No
84	T67.1	47	8	0	No
85	T70.1	32	10	0	No

Tabla 3.2 Descripción de los archivos con clasificación de referencia

No.	Nombre del archivo	Cantidad de instancias	Cantidad de clases	Cantidad de rasgos		Valores ausentes
				Numéricos	Simbólicos	
Conjunto de datos para técnicas de aprendizaje automático y descubrimiento del conocimiento en bases de datos, publicados por la Universidad de California, Irvine. http://archive.ics.uci.edu/ml y http://kdd.ics.uci.edu						
1	Balance-scale	625	3	4	0	No
2	Credit-g	1000	2	7	13	No
3	Diabetes	768	2	8	0	No
4	Echocardiogram*	61	2	8	3	No
5	Glass	214	7	9	0	No
6	Heart-statlog	270	2	7	6	No
7	Ionosphere	351	2	34	0	No
8	Iris	150	3	4	0	No
9	Setter*	20000	26	16	0	No
10	Segment	2310	7	19	0	No
11	Sonar	208	2	60	0	No
12	Vehicle	946	4	18	0	No
13	Vowel*	990	11	10	0	No
14	WaveForm*	5000	3	40	0	No
Conjunto de datos para validar algoritmos sobre agrupamiento y series de tiempo, publicados por el Dr. Eamonn Keogh http://www.cs.ucr.edu/~eamonn/time_series_data/#Control_chart						
15	50Words*	564	15	254	0	No
16	Adiac*	411	10	176	0	No
17	Beef	60	5	254	0	No
18	Cbf	930	3	128	0	No
19	Coffee	56	2	255	0	No
20	CoverType*	731	7	54	0	No
21	Ecg200	200	2	96	0	No
22	FaceFour	111	4	254	0	No

23	Gun-Point	200	2	150	0	No
24	Lighting7	143	7	254	0	No
25	Monk2	432	2	6	0	No
26	OliveOil	60	4	255	0	No
27	Swedishleaf	1125	15	128	0	No
28	SyntheticControl*	600	6	60	0	No
29	Trace	199	4	254	0	No
30	Two-Patterns*	1244	4	128	0	No
31	Wafer*	144	2	152	0	No
32	Wbcd	699	2	9	0	No
33	Yoga*	1088	2	254	0	No
34	Zoo	101	7	16	0	No

* Bases de casos modificadas para facilitar su procesamiento

3.4 Conclusiones parciales

Las medidas basadas en RST para validar resultados del agrupamiento logran correlaciones altamente significativas con las principales medidas internas y externas referenciadas en la literatura. Así, se muestra que esta propuesta logra captar las mismas buenas propiedades que aquellas ya existentes de igual naturaleza y que tiene un alto valor práctico, ya que logra valoraciones similares a las externas sin requerir clasificaciones de referencia. Estos resultados se evidenciaron con bases de casos provenientes de disímiles dominios.

El agrupamiento basado en RST es objetivo, confiable y válido, tiene baja complejidad computacional, es independiente de la forma de los grupos resultantes, permite la inclusión de propiedades estructurales y de medidas de otras naturalezas en la validación, posibilita la variación de la granularidad con que se desea evaluar el agrupamiento mediante la especificación del umbral de similitud entre los objetos, no depende de la existencia de centros de grupos para la evaluación por lo que permite validar el resultado proveniente de un gran número de métodos de agrupamiento.

La medida calidad generalizada con las ponderaciones: cardinalidad, overall similarity y tercera expresión de la pertenencia aproximada por grupos, logra captar todas las propiedades deseadas de los agrupamientos, mostrando una correlación altamente significativa con los resultados de las medidas internas y externas aplicadas. Así se demuestra que las variantes generalizadas introducidas, así como las formas de ponderación propuestas, particularmente la

nueva expresión para medir la pertenencia aproximada, arrojan mejores resultados que las medidas de calidad y precisión ya establecidas en RST.

Todo lo anterior muestra que el empleo de la Teoría de los Conjuntos Aproximados, concretamente el Algoritmo 1 y las medidas basadas en RST, permite valorar los grupos y los resultados generales de agrupamientos; a través de la validación de las principales propiedades deseadas del mismo.

CONCLUSIONES Y RECOMENDACIONES

Como resultado de tesis se extendió el módulo para validar los resultados de los algoritmos de agrupamiento, cumpliéndose de esta forma el objetivo general planteado, ya que:

- Se incorpó al Weka las medidas internas y externas referenciadas en la literatura, así como las medidas basadas en la Teoría de los Conjuntos Aproximados propuestas por investigadores del Laboratorio de Inteligencia Artificial. Estas se encuentran dentro del paquete *clusterer* de Weka con sus respectivos nombres de clases *MedidasInternas.java, MedidasExternas.java y RST.java.*
- Se evaluó las medidas internas, externas y basadas en la Teoría de los Conjuntos Aproximados como instrumentos de validación del agrupamiento, utilizando algunos algoritmos de agrupamiento de Weka y más de 100 bases de casos.
- Comparó los resultados de las medidas basadas en la Teoría de los Conjuntos Aproximados respecto a los resultados de las medidas clásicas mediante métodos estadísticos. Comprobándose que las medidas basadas en la Teoría de los Conjuntos Aproximados dan resultados muy similares a las restantes medidas referenciadas en la literatura.

Teniendo en consideración que el Weka es extensible se recomienda:

- Incorporarle al módulo de validación del agrupamiento en Weka otras medidas internas y externas referenciadas en la literatura; para contribuir de esta forma a una mejor validación de los resultados del agrupamiento.
- Agregarle al Weka la clase *RST.java* dentro del paquete *core* la cual sea genérica, para que pueda ser utilizada un disversas implementaciones dentro Weka donde se necesite trabajar con Teoría de los Conjuntos Aproximados.

Referencias bibliográficas

Akaike, H. (1974) A new look at the statistical model identification. IEEE Trans. Automat. Control, 19: 716-723.

Ankerst, M., Breunig, M. M., Kriegel, H.-P. and Snader, J. (1996) OPTICS: Ordering points to identify the clustering structure. In International Conference on Management of Data. Vol. 28 ACM Press, Philadelphia, PA, USA, pp. 49-60.

Arco, L., Bello, R. and Artiles, M. (2006a) New clustering validity measures based on rough set theory. In Proceedings of International Symposium on Fuzzy and Rough Sets (ISFUROS'06). (Eds, Falcón, R. and Bello, R.) Santa Clara, Cuba.

Arco, L., Bello, R. and Artiles, M. (2006b) Un nuevo enfoque del uso de los conjuntos aproximados en la solución de problemas de la minería de textos. In VII Conferencia Científica Internacional de la Universidad de Ciego de Ávila (UNICA2006). Ciego de Ávila, Cuba.

Arco, L., Bello, R. and García, M. M. (2006c) On clustering validity measures and the rough set theory. In Proceedings of the Fifth Mexican International Conference on Artificial Intelligence (MICAI'06). IEEE Computer Society, Apizaco, México, pp. 168-177.

Batchelor, B. (1978) Pattern Recognition: Idead in Practice, Plenum Press, New York.

Bazan, J., Nguyen, H. S. and Szczuka, M. (2004) A view on rough set concept approximations. Fundamenta Informatica, 59(2-3): 107-118.

Bell, D. and Guan, J. (1998) Computational methods for rough classification and discovery. Journal of the American Society for Information Science (ASIS), 49(5): 403-414.

Berry, M. W. (2004) Survey of Text mining: Clustering, Classification, and Retrieval, Springer Verlag, New York, NY, USA.

Bezdek, J. and Pal, N. (1995) Cluster validation with generalized Dunn's indices. In Proceedings of the 2nd International two-stream Conference on ANNES. (Eds, Kasabov, N. and Coghill, G.) IEEE Press, Piscataway, NJ, pp. 190-193.

Bezdek, J. C. and Pal, N. R. (1998) Some new indexes of cluster validity. IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics, 28(3): 301-315.

Bock, H. (1985) On significance tests in cluster analysis. J. Classification, 2: 77-108.

Borgelt, C. and Kruse, R. (2006) Finding the number of fuzzy clusters by resampling. In IEEE 2006 Proceedings of International Conference on Fuzzy Systems. pp. 48-54.

Bradley, P. S., Fayyad, U. and Reina, C. (1998) Scaling clustering algorithms to large databases. In 4th International Conference on Knowledge Discovery and Data Mining. AAAI Press, New York, NY, USA, pp. 9-15.

Brás-Silva, H., Brito, P. and Costa, J. P. d. (2006) A partitional clustering algorithm validated by a clustering tendency index based on graph theory. Pattern Recognition, 39: 776-788.

Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E. and Dougherty, E. R. (2007) Model-based evaluation of clustering validation measures. Pattern Recognition, 40: 807-824.

Caballero, Y. (2007) Aplicación de la teoría de los conjuntos aproximados en el proprocesamiento de los conjuntos de entrenamiento para algoritmos de aprendizaje. In Departamento de Ciencia de la Computación. Vol. Doctor en Ciencias Técnicas Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Villa Clara.

Caballero, Y., Arco, L., Bello, R. and Marx-Gómez, J. (2007a) New measures for evaluationg decision systems using rough set theory: the application in seadonal weather forecasting. In Proceedings of the Third International ICSC Symposium on Information Technologies in Environmental Engineering (ITEE'07). (Eds, Marx-Gómez, J., Sonnenschein, M., Müller, M., Welsch, H. and Rautenstrauch, C.) Springer Verlag, Carl von Ossietzky Universität Oldenburg, Alemania, pp. 161-174.

Caballero, Y., Arco, L., Bello, R., Salgado, Y., Márquez, Y., León, P. and Álvarez, D. (2007b) Nuevas medidas de la teoría de los conjuntos aproximados para la evaluación de sistemas de información en Bioinformática. In II Congreso Internacional de Bioinformática y Neuroinformática. XII Convención y Expo Internacional Informática'07. La Habana, Cuba.

Calinski, R. B. and Arabas, J. (1974) A dendrite method for cluster analysis. Comm. in Statistics, 3: 1-27.

Dave, R. N. (1996) Validating fuzzy partitions obtained through c-shells clustering. Pattern Recognition Letters, 17: 613-623.

Davies, D. L. and Bouldin, D. W. (1979) A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Learning, 1(4): 224-227.

Deogun, J. S., Choubey, S. K., Raghavan, V. V. and Sever, H. (1998) Feature selection and effective classifiers. Journal of the American Society for Information Science (ASIS), 49(5): 423-434.

Deogun, J. S., Raghavan, V. V. and Server, H. (1994) Rough set based classification methods and extended decision tables. In Proceedings of the International Workshop on Rough Sets and Soft Computing. San José, California, pp. 302-309.

Diday, E. (1974) Recent prograss in distance and similarity measures in pattern recognition. In Second International Joint Conference on Pattern Recognition. pp. 534-539.

Duch, W. (2002) Similarity-based methods: a general framework for classification. Control and Cybernetics, 29(4): 937-968.

Dunn, J. (1974) A fuzzy relative isodata process and its use in detecting compact well-separated clusters. J. Cybernetics, 3: 32-57.

Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In 2nd International Conference on Knowledge Discovery and Data Mining. AAAI Press, Portland, OR, USA.

Falkowski, T., Bartelheimer, J. and Spiliopoulou, M. (2006b) Mining and visualizing the evolution of subgroups in social networks. In WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society, Washington, DC, USA, pp. 52-58.

Frakes, W. B. and Baeza-Yates, R. (1992) Information Retrieval. Data Structure & Algorithms, Prentice Hall, New York.

García, M. (1999) Monografía de reconocimientos de patrones.

Garre, M., Cuadrado, J. J. and Sicilia, M. A. (2005) Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software.

Goodman, L. and Kruskal, W. (1954) Measures of associations for cross-validations. J. Am. Stat. Assoc., 49: 732-764.

Grabowski, A. (2004) Basic properties of Rough Sets and Rough Membership Function. Formalized Mathematics, 12(1): 21-28.

Grau, R., Correa, C. and Rojas, M. (2004) Metodología de la investigación, El Poira, Ibagué, Colombia.

González, H. M. and Pérez, L. I. A. (2005-2006) Extensiones al Ambiente de Aprendizaje Automatizado Weka. Departamento de Inteligencia Artificial, Tesis de Pregrado.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001a) Clustering algorithms and validity measures. In Proceedings of the 13th International Conference on Scientific and Statistical Database Management. IEEE Computer Society, pp. 3-22.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001b) On clustering validation techniques. Journal of Intelligent Information Systems, 17(2/3): 107-145.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2002) Clustering validity checking methods: Part II. ACM SIGMOD Record, 31(3): 19-27.

Halkidi, M. and Vazirgiannis, M. (2001) Clustering validity assessment: finding the optimal partitioning of a data set.

Halkidi, M., Vazirgiannis, M. and Batistakis, Y. (2000) Quality scheme assessment in the clustering process. In Proceedings of PKDD. Lyon, France.

Hand, D. J. (1981) Discrimination and classification, John Wiley and Sons.

Hansen, C. D. and Johnson, C. R. (Eds.) (2005) The Visualization handbook, Elsevier Academic press.

Har-even, M. and Brailovsky, V. L. (1995) Probabilistic validation approach for clustering. Pattern Recognition Letters, 16: 1189-1196.

Hinneburg, A. and Keim, D. A. (1998) An efficient approach to clustering in large multimedia databases with noise. In 4th International Conference on Knowledge Discovery and Data Mining. AAAI Press, New York, NY, USA, pp. 58-65.

Hochbaum, D. S. and Shmoys, D. (1985) A best possible heuristic for the k-center problem. Mathematics of Operations Research, 10(2): 180-184.

Höppner, F., Klawonn, F., Kruse, R. and Runkler, T. (1999) Fuzzy cluster analysis: methods for classification, data analysis and image recognition., John Wiley & Sons Ltd., West Sussex, England.

Hubert, L. and Schultz, J. (1976) Quadratic asignment as a general data-analysis strategy. Br. J. Math. Stat. Psicol., 29: 190-241.

Jain, A. K. and Dubes, R. C. (1988) Algorithms for clustering data, Prentice Hall College Div, Englewood Cliffs, NJ.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999) Data clustering: a review. ACM Computing Surveys, 31(3): 264-323.

Jonnalagadda, S. and Srinivasan, R. (2004) An information theory approach for validating clusters in microarray data.

Kannan, R., Vempala, S. and Vetta, A. (2001) On clusterings: good, bad and spectral.

Kaufman, L. and Rousseeuw, P. J. (1990) Finding groups in data: an introduction to cluster analysis, John Wiley and Sons.

Kim, D. J. and Park, Y. W. (2001) A novel validity index for determination of the optimal number of clusters. IEEE Trans. Inform. Syst., E84-D(2): 281-285.

Kim, M. and Ramakrishna, R. S. (2005) New indices for cluster validity assessment. Pattern Recognition Letters, 26: 2353-2363.

Komorowski, J., Pawlak, Z. and Polkowski, L. (1999) In Rough-Fuzzy Hybridization: A New Trend in Decision Making(Eds, Pal, S. K. and Skowron, A.) Springer-Verlag, Singapore, pp. 3-98.

Kruse, R., Döring, C. and Lesor, M.-J. (2007) In Advances in Fuzzy Clustering and its Applications(Eds, Oliveira, J. V. d. and Pedrycz, W.) John Wiley and Sons, Est Sussex, England, pp. 3-27.

Lam, B. S. Y. and Yan, H. (2007) Assessment of microarray data clustering results based on a new geometrical index for cluster validity. Soft Computing, 11: 341-348.

Lange, T., Braun, M. L., Roth, V. and Buhmann, J. M. (2003) In Advances in neural information processing systems(Eds, Becker, S., Thrun, S. and Obermayer, K.), pp. 617-624.

Levine, E. and Domany, E. (2001) Resampling method for unsupervised estimation of cluster validity. 2001, 13(11): 2573-2593.

Liang, J., Shi, Z. and Li, D. (2003) Applications of inclusion degree in rough set theory. International Journal of Computational Cognition, 1(2): 67-78.

Liu, Y., Cai, J., Yin, J. and Huang, Z. (2006) An efficient clustering algorithm for small text documents. In Seventh International Conference on Web-Age Information Management (WAIM 2006). IEEE Commnunications Society.

McQueen, J. (1967) Some methods for classification and analysis of multivariate observations. In 5th Berkeley Symposium on Mathematics.

Mali, K. and Mitra, S. (2003) Clustering and its validation in a symbolic framework. Pattern Recognition Letters, 24: 2367-2376.

Maulik, U. and Bandyopadhyay, S. (2002) Performance evaluation of some clustering algorithms and validity indices. IEEE Trans. Pattern Anal Mach Intell, 24(12): 1650-1654.

Michalski, R. S., Stepp, R. E. and Diday, E. (1981) A recent advance in data analysis: clustering objects into classes characterized by conjuntive concepts. Progress in Pattern Recognition, 1: 33-56.

Milligan, G. W. and Cooper, M. C. (1985) An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50: 159-179.

Newman, M. E. J. (2003a) Mixing patterns in networks. Physical Review E, 67(026126).

Newman, M. E. J. and Girvan, M. (2004) Finding and evaluating community structure in networks. Physical Review E, 69(026113).

Niu, Z.-Y., Ji, D.-H. and Tan, C.-L. (2004) Document clustering based on cluster validation. In CIKM'04 Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. ACM Press, pp. 501-506.

Olex, A. L., John, D. J., Hiltbold, E. M. and Fetrow, J. S. (2007) Additional limitations of the clustering validation method figure of merit. In Proceedings of ACM Southeast Regional Conference. (Eds, John, D. and Kerr, S. N.) ACM Press, pp. 238-243.

Pal, N. R. and Biswas, J. (1997) Cluster validation using graph theoretic concepts. Pattern Recognition, 30(6): 847-857.

Pal, S. K. and Skowron, A. (Eds.) (1999) Rough Fuzzy Hybridization: A New Trend in Decision Making, Springer.

Pawlak, Z. (1982) Rough sets. International Journal of Computer and Information Sciences, 11: 341-356.

Pawlak, Z. (1991) Rough Sets: Theoretical Aspects of Reasoning about Data, Academic Publishers, Dordrecht.

Pawlak, Z., Grzymala-Busse, J. W., Slowinski, R. and Ziarko, W. (1995) Rough sets. Communications ACM, 38(11): 89-95.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. and Parisi, D. (2004a) Defining and identifying communities in networks. PNAS Proc. National Academy of Science USA, 101: 2658-2663.

Rosell, M., Kann, V. and Litton, J.-E. (2004a) Comparing comparisons: document clustering evaluation using two manual classifications. In Proceedings of ICON 2004 International Conference on Natural Language Processing Hyderabad, India.

Rosell, M., Kann, V. and Litton, J. E. (2004b) Comparing comparisons: document clustering evaluation using two manual classifications. In Proceedings of International Conference on Natural Language processings ICON. Hyderabad, India.

Roussinov, D. G. and Chen, H. (1999) Document clustering for electronic meetings: an experimental comparison of two techniques. Decision Support Systems, 27: 67-79.

Sampieri, R. H. (2007) Fundamentos de la metodología de la investigación, McGraw-Hill Interamericana de España, Madrid.

Sampieri, R. H., Collado, C. F. and Lucio, P. B. (2006) Metodología de la investigación, McGraw-Hill Interamericana de México, Madrid.

Schwartz, G. (1978) Estimation the dimension of a model. Ann Statu, 6: 461-464.

Shannon, C. E. (1948) A mathematical theory of communications. The Bell System Technical Journal, 27: 379-423, 623-656.

Sharma, S. C. (1996) Applied Multivariate Techniques. John Wiley and Sons.

Silberschatz, A. and Tuzhilin, A. (1996) What makes patterns interesting in knowledge discovery systems. IEEE Trans. on Knowledge and Data Engineering, 8(6).

Skowron, A. (2000) Rough sets in KDD. In Proceedings of Conference on Intelligent Information Processing (IIP2000). (Eds, Shi, Z., Faltings, B. and Musen, M.) Publishing House of Electronic Industry, Beijing, pp. 1-17.

Slowinski, R. and Vanderpooten, D. (1997) In Advances in Machine Intelligence & Soft-Computing, Vol. IV (Ed, Wang, P. P.), pp. 17-33.

Stein, B., Eissen, S. M. z. and Wißbrock, F. (2003) On clustering validity and the information need of users. In Proceedings of 3rd IASTED Int. Conference on Artificial Intelligence and Applications (AIA 03). (Ed, Hanza, M. H.) ACTA Press, Benalmádena, Spain, pp. 216-221.

Steinbach, M., Karypis, G. and Kumar, V. (2000) A comparison of document clustering techniques. In Proceedings of KDD Workshop on Text Mining.

Strehl, A., Ghosh, J. and Mooney, R. (2000) Impact of similarity measures on Web-page clustering. In 17th NAtional Conference on Artificial Intelligence (AAAI-2000): Workshop of Artificial Intelligence for Web Search. Austin, Texas.

Theodoridis, S. and Koutroubas, K. (1999) Pattern Recognition, Academic Press.

Tuzhilin, A. (2002) In Handbook of Data Mining and Knowledge DiscoveryOxford University Press.

Wasserman, S. and Faust, K. (1994b) Social network analysis: methods and applications, Cambridge University Press, Cambridge.

Wilkinson, D. M. and Huberman, B. A. (2004) A method for finding communities of related genes. PNAS Proc. National Academy of Science USA, 10(1073).

Wilson, D. R. and Martínez, T. R. (1997) Improved heterogeneous distance functions. Journal of Artificial Intelligence Research, 6: 1-34.

Xie, X. L. and Beni, G. (1991) A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Learning, 13(4): 841-846.

Xie, Y., Raghavan, V. V., Dhatric, P. and Zhao, X. (2005) A new fuzzy clustering algorithm for optimally finding granular prototypes. International Journal of Approximate Reasoning, 40: 109-124.

Xiong, H., Wu, J. and Chen, J. (2006) K-means clustering versus validation measures: a data distribution perspective. In KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, Philadelphia, PA, USA, pp. 779-784.

Xu, W. and Gong, Y. (2004) Document clustering by concept factorization. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, Sheffield, United Kingdom.

Yeung, K. Y., Haynor, D. R. and Ruzzo, W. L. (2001) Validating clustering for gene expression data. Bioinformatics, 17(4): 309-318.

Zhao, Y. and Karypis, G. (2003) Criterion functions for document clustering: experiments and analysis. Machine Learning.

Zhong, N., Skowron, A. and Ohsuga, S. (Eds.) (1999) New Directions in Rough Sets, Data Mining, and Granular Soft-Computing, 17th International Workshop, RSFDGrC'99, Yamaguchi, Japan, November 8-11, 1999, Proceedings, Springer.

Anexos

Anexo 1. Distancias, similitudes y disimilitudes más usadas para comparar objetos

Sean los objetos O_i y O_j descritos por k rasgos, donde $O_i=(o_{i1}, \ldots, o_{ik})$ y $O_j=(o_{j1}, \ldots, o_{jk})$

Distancia Euclideana

$$D_{Euclideana}(O_i, O_j) = \sqrt{\sum_{h=1}^{k} (o_{ih} - o_{jh})^2} \qquad \text{(A1.1)}$$

Distancia Minkowski (Batchelor 1978)

$$D_{Minkowski}(O_i, O_j) = \left(\sum_{h=1}^{k} |o_{ih} - o_{jh}|^{\gamma} \right)^{\frac{1}{\gamma}} \text{ donde } \gamma \geq 1 \qquad \text{(A1.2)}$$

La distancia Minkowsky es equivalente a la distancia Manhattan o city-block, función alternativa que requiere menos esfuerzo computacional, y a la distancia Euclideana cuando γ es 1 y 2, respectivamente (Batchelor 1978) . Para los valores de $\gamma \geq 2$, la distancia Minkowsky es equivalente a la distancia Supermum (Hand 1981, Reed 1972) .

Distancia Euclideana heterogénea (Heterogenous Euclidean – Overlap Metric; HEOM)

$$D_{HEOM}(O_i, O_j) = \sqrt{\sum_{h=1}^{k} d_{local}(o_{ih}, o_{jh})^2} \text{ , donde}$$

$$d_{local}(o_{ih}, o_{jh}) = \begin{cases} d_{Overlap}(o_{ih}, o_{jh}) & si\ h\ simbólico \\ d_{NormEuclidean}(o_{ih}, o_{jh}) & si\ h\ numérico \end{cases} \qquad \text{(A1.3)}$$

$$d_{Overlap}(o_{ih}, o_{jh}) = \begin{cases} 0, & si\ o_{ih} = o_{jh} \\ 1, & en\ otro\ caso \end{cases} \quad \text{y} \quad d_{NormEuclidean}(o_{ih}, o_{jh}) = \frac{|o_{ih} - o_{jh}|}{\max_h - \min_h}$$

Distancia Camberra (Michalski, Stepp et al. 1981, Diday 1974).

$$D_{Camberra}(O_i, O_j) = \sum_{h=1}^{k} \frac{|o_{ih} - o_{jh}|}{|o_{ih} + o_{jh}|} \quad \text{(A1.4)}$$

Correlación de Pearson (Wilson and Martínez 1997).

$$D_{Pearson}(O_i, O_j) = \frac{\sum_{h=1}^{k} \left(o_{ih} - \overline{atributo_h}\right)\left(o_{jh} - \overline{atributo_h}\right)}{\sqrt{\sum_{h=1}^{k} \left(o_{ih} - \overline{atributo_h}\right)^2 \sum_{h=1}^{k} \left(o_{jh} - \overline{atributo_h}\right)^2}} \quad \text{(A1.5)}$$

donde $\overline{atributo_h}$ es el valor promedio que toma el $atributo_h$ en el conjunto de datos.

Las expresiones de Chebychev, Mahalanobis, distancia de Hamming y la máxima distancia son otras variantes de cálculo de distancias entre objetos (Wilson and Martínez 1997). En (Duch 2002) se presentan formas de medir la similitud considerando una relación asimétrica entre los objetos y teniendo en cuenta la probabilidad condicional de un objeto respecto al otro.

Existen varias formas de medir la similitud entre objetos, sin embargo, existen pocos estudios comparativos de ellas y sus efectos en el agrupamiento. En la minería de textos, al comparar documentos con el objetivo de agruparlos, la determinación de la similitud entre documentos depende de la representación del documento y de los pesos que se le asignen al caracterizarlo. A partir de estudios realizados, se ha demostrado que al agrupar documentos, los coeficientes Dice, Jaccard y Coseno, han reportado los mejores resultados (Frakes and Baeza-Yates 1992). Una valoración del imparto de la distancia Euclideana y los coeficientes Dice, Jaccard y Coseno en dominios textuales fue presentada en (Strehl, Ghosh et al. 2000).

Coeficiente Dice

$$S_{Dice}(O_i, O_j) = \frac{2\sum_{h=1}^{k} (o_{ih} \cdot o_{jh})}{\sum_{h=1}^{k} o_{ih}^2 + \sum_{h=1}^{k} o_{jh}^2} \quad \text{(A1.6)}$$

Coeficiente de Jaccard

$$S_{Jaccad}(O_i, O_j) = \frac{\sum_{h=1}^{k}(o_{ih} \cdot o_{jh})}{\sum_{h=1}^{k} o_{ih}^2 + \sum_{h=1}^{k} o_{jh}^2 - \sum_{h=1}^{k}(o_{ih} \cdot o_{jh})} \quad \text{(A1.7)}$$

Coeficiente Coseno

$$S_{Coseno}(O_i, O_j) = \frac{\sum_{h=1}^{k}(o_{ih} \cdot o_{jh})}{\sqrt{\sum_{h=1}^{k} o_{ih}^2 \cdot \sum_{h=1}^{k} o_{jh}^2}} \quad \text{(A1.8)}$$

Anexo 2. Medidas externas para la evaluación del agrupamiento

Entropía (Shannon 1948, Steinbach, Karypis et al. 2000, Rosell, Kann et al. 2004a) (Zhao and Karypis 2003)

$$E_{CS} = \sum_{j=1}^{m} \frac{n_j * E_j}{n} \qquad \text{(A2.1)}$$

donde m es el número de grupos, n_j es el tamaño del grupo j, n es el número total de objetos agrupados y E_j se calcula según las siguientes expresiones de entropía de un grupo.

Entropía de un grupo

$$E_j = -\sum_i p_{ij} \log\left(p_{ij}\right) \text{ o } E_j = -\frac{1}{\log m} \sum_{i=1}^{m} p_{ij} \log\left(p_{ij}\right) \qquad \text{(A2.2)}$$

Donde p_{ij} es la probabilidad que un miembro del grupo j pertenezca a la clase i.

Overall F-Measure (Steinbach, Karypis et al. 2000)

$$Overall\ F-Measure = \sum_{i=1}^{k} \frac{n_i}{n} \max\left\{F-Measure(i,j)\right\} \qquad \text{(A2.3)}$$

donde k es el número de clases, n_i es el tamaño de la clase i, n es el número total de objetos agrupados y $F\text{-}Measure(i,j)$ se calcula según la expresión siguiente. Si α=1, entonces *Overall F-Measure* se nombra *Purity* (Rosell, Kann et al. 2004b).

F-Measure de la clase *i* respecto al grupo *j*

$$F-Measure(i,j) = \frac{1}{\alpha\left(1/\Pr(i,j)\right) + \left(1-\alpha\right)\left(1/\mathrm{Re}(i,j)\right)} \qquad \text{(A2.4)}$$

Si α=1 entonces $F\text{-}Measure(i,j)$ coincide con *precision*, si α=0 entonces $F\text{-}Measure(i,j)$ coincide con *recall*. α=0.5 significa igual peso para *precision* y *recall*.

Micro-averaged precision y micro-averaged recall (Niu, Ji et al. 2004)

$$\text{MA - Pr} = \frac{\sum_{i=1}^{k} \alpha_i}{\sum_{i=1}^{k} (\alpha_i + \beta_i)} \text{ y } \text{MA - Re} = \frac{\sum_{i=1}^{k} \alpha_i}{\sum_{i=1}^{k} (\alpha_i + \gamma_i)} \quad \text{(A2.5)}$$

Donde α_i es el número de objetos correctamente asignados a la clase i, β_i es el número de objetos incorrectamente asignados a la clase i y γ_i es el número de objetos incorrectamente no asignados a la clase i. MA-Pr=MA-Re si cada objeto pertenece a sólo un grupo y la clasificación de referencia también tiene una clasificación única para cada objeto.

Información mutua (Xu and Gong 2004)

$$MI = \sum_{i,j} p_{ij} \cdot \log_2 \frac{p_{ij}}{p_i \cdot p_j} \quad \text{(A2.6)}$$

donde p_i, p_j denotan las probabilidades que un objeto pertenezca a la clase i y al grupo j, respectivamente, y p_{ij} denota la probabilidad de que ese objeto seleccionado pertenezca a la clase i y al grupo j simultáneamente. Una variante normalizada de la información mutua divide la expresión anterior por la máxima entropía.

Error del agrupamiento (Roussinov and Chen 1999)

$$CE = \frac{E}{P_t}, \text{ donde } P_t = \frac{1}{2} n(n-1) \text{ y } E = E_i + E_a \quad \text{(A2.7)}$$

donde P_t es el número total de pares de objetos posibles, y E es el número total de asociaciones incorrectas y ausentes.

Error del agrupamiento normalizado en el intervalo [0, 1] (Roussinov and Chen 1999)

$$NCE = \frac{E}{A_t}, \text{ donde } A_t = A_m + A_a \quad \text{(A2.8)}$$

Donde A_t es el número total de asociaciones que existen en ambas particiones sin eliminar duplicados, donde A_m es el número total de asociaciones en la partición de referencia y A_a es el número total de asociaciones en la partición resultado del agrupamiento. Sólo se consideran asociaciones en grupos con más de dos objetos.

Cluster Recall y Cluster Precision (Roussinov and Chen 1999)

$$CR = {A_c}/{A_m} \text{ y } CP = {A_c}/{A_a} \quad \text{(A2.9)}$$

Donde A_c=A_a-E_i, representa el número total de asociaciones resultantes del agrupamiento.

Rand Statistic

$$R = (a+b)/m \quad \text{(A2.10)}$$

Coeficiente de Jaccard

$$J = a/(a+b+c) \quad \text{(A2.11)}$$

Índice de Folkes y Mallows

$$FM = \left(\frac{a}{a+b} \cdot \frac{a}{a+c} \right)^{1/2} \quad \text{(A2.12)}$$

Donde a es el número de pares de objetos que pertenecen al mismo grupo y a la misma clase, b es el número de aquellos pares que pertenecen al mismo grupo y a clases diferentes, c es el total de pares que pertenecen a grupos diferentes y a la misma clase, d es el número de pares de objetos que pertenecen a grupos y clases diferentes y m=a+b+c+d es el número máximo de todos los pares de objetos (es decir, m=n(n-1)/2 donde n es el número total de objetos).

Anexo 3. Medidas internas para la evaluación del agrupamiento

Overall Similarity (Steinbach, Karypis et al. 2000)

$$OverallSimilarity(Grupo) = \frac{1}{|Grupo|^2} \sum_{O_i, O_j \in Grupo} \text{distancia}(O_i, O_j) \tag{A3.1}$$

Índices Dunn

$$I(C) = \frac{\min_{i \neq j} \{\delta(C_i, C_j)\}}{\max_{1 \leq l \leq k} \{\Delta(C_l)\}} \tag{A3.2}$$

Donde $C=\{C_1, \ldots, C_k\}$es el agrupamiento de un conjunto de objetos O, $\delta: C \times C \rightarrow R$ es una medida de distancia de grupo a grupo, y $\triangle: C \rightarrow R$ es una medida de diámetro del grupo.

$$\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \text{ y } \Delta(C_i) = \max_{x, y \in C_i} d(x, y) \tag{A3.3}$$

donde $d: C \times C \rightarrow R$ es una función que mide la distancia entre los objetos de O.

Propuesta de Bezdek para el cálculo de $\delta(C_i, C_j)$ y $\Delta(C_i)$

$$\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} d(x, y) \text{ y } \Delta(C_i) = 2\left(\frac{\sum_{x \in C_i} d(x, c_i)}{|C_i|} \right) \tag{A3.4}$$

Donde c_i es el centro del grupo C_i.

Índice de separación (Höppner, Klawonn et al. 1999)

$$F_{Indice\ de\ separación} = \min_{i \in K} \left\{ \min_{j \in K - \{i\}} \left\{ \frac{d(A_i, A_j)}{\max_{k \in K} diam(A_k)} \right\} \right\} \tag{A3.5}$$

Donde $diam(A_k)=\max\{d(x_i, x_j) \mid x_i, x_j \in A_k\}$ y la función de distancia d es extendida a los conjuntos por $d(A_i, A_j)=\min\{d(x_i, x_j) \mid x_i \in A_k, x_j \in A_k\}$ (para $i, j, k \in K$).

Índice Davies – Bouldin (Davies and Bouldin 1979)

$$DB(C) = \frac{1}{k} \cdot \sum_{i=1}^{k} R_i \qquad \text{(A3.6)}$$

$$R_i = \max_{\substack{j=1,\ldots,n, \\ i \neq j}} R_{ij}, \text{ donde } R_{ij} = \frac{(s(C_i) + s(C_j))}{\delta(C_i, C_j)} \text{ y } s(C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} \|x - c_i\| \qquad \text{(A3.7)}$$

Donde $C = \{C_1, \ldots, C_k\}$ es un agrupamiento de un conjunto de objetos, c_i es el cemtroide del grupo C_i, $s{:}C \rightarrow \mathrm{R}$ mide la dispersión dentro de un grupo y $\delta{:}C \times C \rightarrow \mathrm{R}$ es una medida de distancia de grupo a grupo.

Las medidas $\mathbf{\Lambda}$ y ρ consideran la colección de objetos como un grafo pesado $G=(V,E,w)$ con el conjunto de nodos V, aristas E y la función de peso $w{:}E \rightarrow [0,1]$ donde V representa los objetos y w define la similitud entre dos objetos adyacentes. Considérese $C=\{C_1,\ldots,C_k\}$ un agrupamiento de un grafo pesado $G=(V,E,w)$.

Medida de conectividad parcial pesada $\mathbf{\Lambda}$

$$\Lambda(C) = \sum_{i=1}^{k} |C_i| \cdot \lambda_i \qquad \text{(A3.8)}$$

donde λ_i designa la conectividad de las aristas pesadas de $G(C_i)$. λ de un grafo $G=(V, E, w)$ es definida como $\min \sum_{\{u,v\} \in E'} w(u,v)$, donde $E' \subset E$ y $G'=(V, E \backslash E')$ es no conexo. λ es también designada como la capacidad de un corte mínimo de G.

Medida de densidad esperada ρ

$$\rho(C) = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^{\theta}}, \text{ donde } |V|^{\theta} = w(G) \text{ y } w(G) = |V| + \sum_{e \in E} w(e) \qquad \text{(A3.9)}$$

θ se calcula para grafos ponderados según la expresión siguiente.

$$w(G) = |V|^{\theta} \quad \Leftrightarrow \quad \theta = \frac{\ln(w(G))}{\ln(|V|)} \tag{A3.10}$$

Índice de dispersión-distancia (Scatter-Distance; SD) (Halkidi, Vazirgiannis et al. 2000)

$$SD(nc) = a \cdot Scat(nc) + Dis(nc) \tag{A3.11}$$

$$Sact(nc) = \frac{1}{nc}\sum_{i=1}^{nc} \|\sigma(v_i)\| / \|\sigma(X)\| \text{ y } Dis(nc) = \frac{D_{\max}}{D_{\min}} \sum_{k=1}^{nc} \left(\sum_{z=1}^{nc} \|v_k - v_z\| \right)^{-1} \tag{A3.12}$$

donde $D_{\max} = \max\left(\|v_i - v_j\|\right)$ es la distancia máxima entre los centros de grupos y $D_{\min} = \min\left(\|v_i - v_j\|\right)$ es la distancia entre centros de grupos, $\forall i, j \in \{1,2,...,nc\}$.

Modularity (Newman and Girvan 2004)

$$Q = \sum_i \left(e_{ii} - a_i^2\right) = \mathbf{Tr}\,\mathbf{e} - \|\mathbf{e}^2\| \tag{A3.13}$$

Donde **e** es una matriz simétrica de orden *k* cuyo elemento e_{ij} es la fracción de todas las aristas en el grafo que conectan nodos del grupo *i* con nodos del grupo *j*, $\|\mathbf{e}\|$ indica la suma de los elementos de la matriz **e** y $\mathbf{Tr}\,\mathbf{e} = \sum_i e_{ii}$ es la traza de la matriz que da la fracción de aristas en en el grafo que conectan nodos en el mismo grupo.

Anexo 4. Variantes para el cálculo del umbral de similitud entre objetos

Algunas variantes para el cálculo inicial del umbral β_0 se describen a continuación (García 1999):

a) La primera variante calcula el umbral β_0 hallando la media de las distancias entre todos los pares de objetos posibles. Así se expresa en la fórmula (A4.1):

$$\beta_0 = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d(O_i, O_j) \qquad \text{(A4.1)}$$

b) La segunda variante halla la media de los valores máximos de las distancias entre cualquier par de objetos, según la expresión (A4.2):

$$\beta_0 = \frac{1}{n} \sum_{i=1}^{n} \max_{\substack{j=1..n \\ i \neq j}} \{d(O_i, O_j)\} \qquad \text{(A4.2)}$$

c) La tercera variante toma la mínima de todas las distancias posibles entre pares de objetos, sin tener en cuenta las distancias que sean cero, así se muestra en la fórmula (A4.3):

$$\beta_0 = \min_{\substack{i=1..n-1 \\ i \neq j}} \left\{ \min_{j=i+1..n} \{d(O_i, O_j)\} \right\} \qquad \text{(A4.3)}$$

La descripción de la notación utilizada es la siguiente: n es la cantidad de objetos de la colección y $d(O_i, O_j)$ es el valor de la distancia entre los vectores documento O_i y O_j.

Anexo 5. Comparación de las medidas aplicadas sobre bases de casos con y sin ruido

Algoritmo	%	Umbral	PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
DBScan	5	.274	.215	.278	.339	.192	.192	.260	.217	.217	.217	.205	.227	.244
	10	.664	.502	.502	.526	.550	.550	.478	.590	.550	.391	.351	.433	.498
	15	.455	.653	.619	.821	.614	.639	.550	.821	.794	.455	.476	.351	.794
	20	.498	.679	.744	.821	.715	.848	.681	.794	.958	.986	.903	.478	.821
	25	.498	.903	.931	.903	.876	.958	.986	.958	.931	.614	.614	.741	.732
EM	5	.715	.212	.267	.260	.314	.295	.243	.212	.295	.244	.244	.306	.314
	10	.958	.858	.943	.691	.470	.764	.102	.644	.284	.102	.132	.647	.520
	15	.741	.717	.868	.881	.627	.654	.778	.654	.881	.476	.520	.872	.765
	20	.639	.601	.557	.794	.709	.737	.778	.601	.627	.035	.058	.968	.717
	25	.958	.658	.687	.639	.768	.741	.546	.664	.768	.394	.289	.314	.351
FarthestFirst	5	.375	.925	.730	1.000	1.000	.881	.826	.654	.936	.778	.737	.778	.296
	10	.434	.438	.326	.295	.469	.376	.326	.455	.296	.421	.478	.408	.332
	15	.639	.975	.875	.823	.911	.940	.959	.550	.601	.394	.394	.959	.313
	20	.931	.925	.975	.794	.601	.737	.877	.575	.823	.391	.391	.679	.601
	25	.852	1.000	1.000	.940	.970	.970	.970	.940	.970	.940	.940	1.000	.852
SimpleKMeans	5	.259	.638	.683	.502	.823	.681	.557	.958	.852	.709	.970	.557	.986
	10	.259	.586	.554	.852	.687	.765	.777	.590	.478	.355	.204	.711	.566
	15	.455	.968	.936	.689	.852	.876	.687	.931	.881	.911	.848	.658	.903
	20	.498	.546	.560	.768	.664	.848	.687	.931	.715	.689	.794	.737	.639
	25	.455	.654	.681	.498	.768	.664	.550	.664	.566	.394	.339	.478	.741
Xmeans	5	.244	.463	.352	.520	.184	.147	.557	.433	.398	.010	.015	.777	.520
	10	.498	.460	.394	.687	.658	.629	.723	.455	.334	.044	.062	.723	.394
	15	.322	.796	.877	.601	.654	.823	.523	.681	.687	.741	.614	.356	.715
	20	.614	.215	.179	.523	.159	.212	.605	.145	.266	.351	.351	.679	.159
	25	.768	.744	.528	.709	.681	.737	.744	.654	.502	.232	.247	.777	.520

Anexo 6. Correlaciones entre medidas basadas en RST e internas referenciadas

Se considera al interpretar los resultados que todas las medidas basadas en RST, las medidas overall similarity, los índices Dunn y la variante con las expresiones de Bezdek, conectividad parcial pesada y densidad esperada es deseable obtener valores tan altos como sea posible, mientras que en la medida Davies-Bouldin se desean resultados opuestos.

Tabla A6.1 Correlaciones entre valores de medidas basadas en RST e internas, aplicadas a resultados de agrupamientos con el algoritmo DensityBasedClusterer

		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
OS	Correl.	.188(*)	.173	.196(*)	.188(*)	.195(*)	.187(*)	.188(*)	.186(*)	.231(*)	.227(*)	.175	.142
	Sig.	.046	.067	.038	.047	.038	.048	.046	.048	.014	.016	.064	.133
DD	Correl.	.434(**)	.398(**)	.434(**)	.426(**)	.428(**)	.426(**)	.432(**)	.408(**)	.689(**)	.693(**)	.304(**)	.364(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.000
DB	Correl.	.446(**)	.422(**)	.431(**)	.454(**)	.444(**)	.437(**)	.449(**)	.421(**)	.538(**)	.539(**)	.402(**)	.464(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
IDB	Correl.	-.410(**)	-.383(**)	-.388(**)	-.424(**)	-.407(**)	-.394(**)	-.414(**)	-.383(**)	-.409(**)	-.409(**)	-.464(**)	-.478(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
CPP	Correl.	.169	.119	.169	.175	.174	.163	.169	.146	-.049	-.041	.202(*)	.134(*)
	Sig.	.074	.208	.073	.064	.065	.085	.073	.124	.605	.663	.032	.015
DE	Correl.	.580(**)	.559(**)	.588(**)	.565(**)	.577(**)	.585(**)	.574(**)	.576(**)	.750(**)	.753(**)	.490(**)	.517(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Tabla A6.2 Correlaciones entre valores de medidas basadas en RST e internas, aplicadas a resultados de agrupamientos con el algoritmo EM

		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
OS	Correl.	-.138	-.162	-.132	-.135	-.128	-.143	-.136	-.152	.028	.022	-.141	-.142
	Sig.	.146	.086	.163	.153	.176	.132	.150	.108	.768	.821	.137	.133
DD	Correl.	.798(**)	.731(**)	.789(**)	.797(**)	.788(**)	.795(**)	.799(**)	.761(**)	.829(**)	.845(**)	.781(**)	.734(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
DB	Correl.	.350(**)	.387(**)	.318(**)	.341(**)	.314(**)	.351(**)	.348(**)	.354(**)	.372(**)	.342(**)	.413(**)	.439(**)

	Sig.	.000	.000	.001	.001	.001	.000	.000	.000	.000	.000	.000	.000
IDB	Correl.	-.580(**)	-.584(**)	-.570(**)	-.574(**)	-.565(**)	-.571(**)	-.578(**)	-.576(**)	-.547(**)	-.539(**)	-.602(**)	-.606(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
CPP	Correl.	-.120	-.136	-.127	-.129	-.137	-.126	-.122	-.133	-.128	-.123	-.110	-.103
	Sig.	.207	.150	.179	.172	.148	.183	.197	.160	.176	.196	.247	.279
DE	Correl.	.412(**)	.468(**)	.431(**)	.398(**)	.419(**)	.418(**)	.407(**)	.459(**)	.489(**)	.481(**)	.344(**)	.324(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Tabla A6.3 Correlaciones entre valores de medidas basadas en RST e internas, aplicadas a resultados de agrupamientos con el algoritmo FarthestFirst

		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
OS	Correl.	.298(**)	.238(*)	.297(**)	.298(**)	.302(**)	.268(**)	.296(**)	.260(**)	.454(**)	.447(**)	.300(**)	.220(*)
	Sig.	.001	.011	.001	.001	.001	.004	.001	.005	.000	.000	.001	.019
DD	Correl.	.443(**)	.393(**)	.436(**)	.444(**)	.440(**)	.409(**)	.442(**)	.407(**)	.485(**)	.489(**)	.352(**)	.349(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
DB	Correl.	.438(**)	.395(**)	.431(**)	.438(**)	.433(**)	.411(**)	.437(**)	.408(**)	.405(**)	.410(**)	.453(**)	.369(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
IDB	Correl.	-.630(**)	-.608(**)	-.628(**)	-.623(**)	-.623(**)	-.615(**)	-.626(**)	-.618(**)	-.633(**)	-.635(**)	-.594(**)	-.547(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
CPP	Correl.	.085	.072	.089	.110	.112	.096	.094	.085	.037	.043	.163	.112
	Sig.	.371	.446	.348	.244	.236	.314	.322	.373	.699	.651	.084	.239
DE	Correl.	.585(**)	.597(**)	.592(**)	.573(**)	.577(**)	.592(**)	.578(**)	.605(**)	.719(**)	.722(**)	.439(**)	.526(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Tabla A6.4 Correlaciones entre valores de medidas basadas en RST e internas, aplicadas a resultados de agrupamientos con el algoritmo SimpleKMeans

		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
OS	Correl.	.131	.122	.139	.136	.143	.140	.133	.133	.217(*)	.214(*)	.163	.098
	Sig.	.166	.199	.141	.151	.130	.139	.160	.159	.021	.023	.085	.302
DD	Correl.	.477(**)	.439(**)	.477(**)	.470(**)	.471(**)	.470(**)	.476(**)	.449(**)	.668(**)	.673(**)	.343(**)	.412(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

DB	Correl.	.502(**)	.487(**)	.492(**)	.509(**)	.502(**)	.496(**)	.504(**)	.488(**)	.570(**)	.572(**)	.467(**)	.495(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
IDB	Correl.	-.584(**)	-.618(**)	-.590(**)	-.577(**)	-.582(**)	-.589(**)	-.582(**)	-.614(**)	-.647(**)	-.645(**)	-.573(**)	-.554(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
CPP	Correl.	.172	.129	.173	.176	.176	.166	.172	.152	-.057	-.050	.211(*)	.132
	Sig.	.068	.174	.067	.063	.063	.079	.068	.107	.548	.599	.025	.165
DE	Correl.	.600(**)	.583(**)	.611(**)	.585(**)	.597(**)	.609(**)	.596(**)	.600(**)	.761(**)	.766(**)	.511(**)	.536(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Tabla A 6.5 Correlaciones entre valores de medidas basadas en RST e internas, aplicadas a resultados de agrupamientos con el algoritmo XMeans

		PA	**CA**	**PGRM1**	**PGRM2**	**PGRM3**	**PGC**	**PGOS**	**CGRM1**	**CGRM2**	**CGRM3**	**CGC**	**CGOS**
OS	Correl.	.131	.122	.139	.136	.143	.140	.133	.133	.217(*)	.214(*)	.163	.098
	Sig.	.166	.199	.141	.151	.130	.139	.160	.159	.021	.023	.085	.302
DD	Correl.	.477(**)	.439(**)	.477(**)	.470(**)	.471(**)	.470(**)	.476(**)	.449(**)	.668(**)	.673(**)	.343(**)	.412(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
DB	Correl.	.502(**)	.487(**)	.492(**)	.509(**)	.502(**)	.496(**)	.504(**)	.488(**)	.570(**)	.572(**)	.467(**)	.495(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
IDB	Correl.	-.584(**)	-.618(**)	-.590(**)	-.577(**)	-.582(**)	-.589(**)	-.582(**)	-.614(**)	-.647(**)	-.645(**)	-.573(**)	-.554(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
CPP	Correl.	.172	.129	.173	.176	.176	.166	.172	.152	-.057	-.050	.211(*)	.132
	Sig.	.068	.174	.067	.063	.063	.079	.068	.107	.548	.599	.025	.165
DE	Correl.	.600(**)	.583(**)	.611(**)	.585(**)	.597(**)	.609(**)	.596(**)	.600(**)	.761(**)	.766(**)	.511(**)	.536(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Anexo 7. Correlaciones entre medidas basadas en RST y externas referenciadas

Se considera al interpretar los resultados que todas las medidas basadas en RST, las medidas overall F-measure, precision y recall desean obtener valores tan altos como sea posible, mientras que la entropía requiere resultados opuestos.

Tabla A7.1 Correlaciones entre valores de medidas basadas en RST y externas, aplicadas a resultados de agrupamientos con el algoritmo DensityBasedClusterer

		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
E	Correl.	-.498(**)	-.447(*)	-.479(**)	-.516(**)	-.509(**)	-.514(**)	-.501(**)	-.454(**)	-.483(**)	-.485(**)	-.530(**)	-.532(**)
	Sig.	.004	.010	.006	.002	.003	.003	.003	.009	.005	.005	.002	.002
P	Correl.	.373(*)	.361(*)	.374(*)	.388(*)	.391(*)	.387(*)	.373(*)	.362(*)	.493(**)	.491(**)	.350(*)	.327
	Sig.	.035	.042	.035	.028	.027	.029	.036	.041	.004	.004	.049	.068
R	Correl.	.389(*)	.412(*)	.404(*)	.389(*)	.406(*)	.423(*)	.386(*)	.409(*)	.570(**)	.567(**)	.412(*)	.345
	Sig.	.028	.019	.022	.028	.021	.016	.029	.020	.001	.001	.019	.053
OFM	Correl.	.397(*)	.393(*)	.392(*)	.415(*)	.414(*)	.416(*)	.397(*)	.388(*)	.516(**)	.514(**)	.424(*)	.401(*)
	Sig.	.025	.026	.027	.018	.018	.018	.024	.028	.003	.003	.016	.023

Tabla A 7.2 Correlaciones entre valores de medidas basadas en RST y externas, aplicadas a resultados de agrupamientos con el algoritmo EM

		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
E	Correl.	-.343(**)	-.302(*)	-.306(*)	-.286(*)	-.250(*)	-.343(**)	-.339(**)	-.302(*)	-.214	-.226	-.419(**)	-.415(**)
	Sig.	.006	.015	.014	.021	.044	.006	.006	.015	.086	.069	.001	.001
P	Correl.	.157	.149	.137	.101	.081	.165	.145	.125	.044	.056	.226	.206
	Sig.	.206	.230	.270	.417	.517	.184	.243	.315	.721	.650	.069	.098
R	Correl.	.310(*)	.278(*)	.290(*)	.262(*)	.258(*)	.302(*)	.315(*)	.278(*)	.262(*)	.266(*)	.331(**)	.327(**)
	Sig.	.013	.025	.020	.035	.038	.015	.011	.025	.035	.032	.008	.009
OFM	Correl.	.315(*)	.298(*)	.294(*)	.274(*)	.246(*)	.339(**)	.319(*)	.282(*)	.242	.246(*)	.399(**)	.379(**)
	Sig.	.011	.016	.018	.027	.048	.006	.010	.023	.052	.048	.001	.002

Tabla A 7.3 Correlaciones entre valores de medidas basadas en RST y externas, aplicadas a resultados de agrupamientos con el algoritmo FarthestFirst

		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
E	Correl.	-.067	.000	-.069	-.068	-.069	-.034	-.066	-.016	-.229	-.226	-.112	-.550(**)
	Sig.	.716	.998	.706	.710	.708	.853	.718	.931	.208	.213	.542	.001
P	Correl.	.166	.113	.171	.185	.191	.118	.168	.124	.408(*)	.409(*)	.074	.379(*)
	Sig.	.363	.540	.350	.310	.296	.521	.358	.498	.021	.020	.686	.032
R	Correl.	.529(**)	.454(**)	.532(**)	.501(**)	.505(**)	.547(**)	.525(**)	.492(**)	.177	.183	.699(**)	.444(*)
	Sig.	.002	.009	.002	.003	.003	.001	.002	.004	.332	.317	.000	.011
OFM	Correl.	.183	.115	.185	.200	.203	.142	.183	.134	.358(*)	.358(*)	.187	.542(**)
	Sig.	.316	.530	.310	.272	.266	.438	.315	.463	.044	.044	.306	.001

Tabla A 7.4 Correlaciones entre valores de medidas basadas en RST y externas, aplicadas a resultados de agrupamientos con el algoritmo SimpleKMeans

		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
E	Correl.	-.509(**)	-.501(**)	-.501(**)	-.511(**)	-.514(**)	-.530(**)	-.510(**)	-.502(**)	-.536(**)	-.541(**)	-.519(**)	-.513(**)
	Sig.	.003	.003	.003	.003	.003	.002	.003	.003	.002	.001	.002	.003
P	Correl.	.373(*)	.409(*)	.382(*)	.373(*)	.384(*)	.390(*)	.372(*)	.402(*)	.515(**)	.510(**)	.352(*)	.327
	Sig.	.035	.020	.031	.036	.030	.027	.036	.023	.003	.003	.048	.068
R	Correl.	.369(*)	.437(*)	.399(*)	.369(*)	.397(*)	.403(*)	.368(*)	.428(*)	.584(**)	.576(**)	.404(*)	.345
	Sig.	.037	.012	.024	.038	.024	.022	.039	.015	.000	.001	.022	.053
OFM	Correl.	.395(*)	.428(*)	.404(*)	.398(*)	.409(*)	.417(*)	.395(*)	.421(*)	.543(**)	.539(**)	.413(*)	.387(*)
	Sig.	.025	.015	.022	.024	.020	.018	.025	.016	.001	.001	.019	.029

Tabla A 7.5 Correlaciones entre valores de medidas basadas en RST y externas, aplicadas a resultados de agrupamientos con el algoritmo XMeans

		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
E	Correl.	-.509(**)	-.501(**)	-.501(**)	-.511(**)	-.514(**)	-.530(**)	-.510(**)	-.502(**)	-.536(**)	-.541(**)	-.519(**)	-.513(**)
	Sig.	.002	.004	.003	.002	.001	.002	.002	.003	.001	.001	.001	.001
P	Correl.	.390(*)	.405(*)	.398(*)	.394(*)	.407(*)	.390(*)	.391(*)	.401(*)	.517(**)	.516(**)	.376(*)	.377(*)
	Sig.	.030	.024	.027	.028	.023	.030	.030	.025	.003	.003	.037	.037
R	Correl.	.349	.401(*)	.370(*)	.357(*)	.379(*)	.377(*)	.350	.393(*)	.536(**)	.526(**)	.424(*)	.380(*)
	Sig.	.054	.025	.041	.049	.035	.036	.053	.029	.002	.002	.038	.035

OFM	Correl.	.452(*)	.470(**)	.457(**)	.466(**)	.477(**)	.466(**)	.456(**)	.466(**)	.596(**)	.594(**)	.497(**)	.482(**)
	Sig.	.011	.008	.010	.008	.007	.008	.010	.008	.000	.000	.004	E