Universidad de Oriente
Facultad Matemática y Computación
Departamento de Computación

Propuestas de Nuevas Medidas de Calidad para Algoritmos de Agrupamiento Solapados

Tesis en opción al título de Licenciado en Ciencia de la Computación

Autor

Henry Rosales Méndez

Tutores

Ms.C. Yunior Ramírez Cruz

Dr.C. Reynaldo Gil García

Santiago de Cuba, junio 2012

Agradecimientos

En el desarrollo de esta tesis me han brindado su apoyo muchas personas a las cuales quiero darles mi eterno agradecimiento. A mis tutores:

- *Reynaldo Gil* por haberme ofrecido un tema de tesis tan importante y por haberme guiado con exactitud y precisión.
- *Yunior Ramírez* por dedicarme un tiempo tan preciado y por su crítica tan constructiva.

Quiero agradecer también al centro CERPAMID - Datys por darme la posibilidad de investigar en un campo tan interesante como el Reconocimiento de Patrones y a sus trabajadores, en especial a: *Lilian Sueiro*, *Javier Sardiñas*, *Reinier Ortega*, *Adrian Fonseca* y *Fernando Artiagas*; por entregarme sus opiniones, sugerencias y conocimientos que fueron decisivos para este trabajo.

A mi familia, ya que esta tesis lleva un poquito de cada uno en particular. De mi madre lleva su tenacidad e inteligencia, de mi padre su austeridad y responsabilidad, de mis hermanos sus experiencias, de mi abuelo su sentido del humor y de *Armando* su confianza en sí mismo.

A mis compañeros de investigación y amigos:

- *Reynaldo Gil* (hijo), por sus magníficos contraejemplos.
- A *Yenisleidi Lora* (Yeni) por ser la coautora anónima de esta tesis.
- A *Yoan Torres* (el más discreto de todos), por su ayuda oportuna.
- A *Jorge L. Toro* por sus opiniones.
- *Erlin Guillén* y *Andro Bermudes*, por formar conmigo un equipo increíble.
- A *Juan C. Pozo* por ser un amigo leal en esta batalla tan importante que es la Universidad.

A todos, muchas gracias.

Dedicatoria

A mi madre, mi mayor tesoro.

A mi padre, constructor de mis sueños.

A la memoria de *Aurora Pons Porrata*.

Resumen

En este trabajo se trata el problema de la evaluación de los algoritmos de agrupamiento solapados. Luego de un exhaustivo estudio del tema, centrado en las medidas de evaluación existentes, así como las condiciones propuestas para determinar la validez de las mismas, se identifica uno de los problemas no resueltos por este tipo de medidas cuando se aplican a algoritmos solapados, sobre la base de lo cual se enuncia una nueva condición para determinar la validez de las medidas de evaluación en este caso.

Dado que la medida que según la literatura cumple todas las condiciones impuestas hasta el momento no cumple la nueva condición en todos los casos, se proponen nuevas medidas de evaluación que resuelven parcialmente el problema detectado. Si bien las nuevas medidas tampoco cumplen la nueva condición en el 100% de los casos, se demuestra que para una de ellas el conjunto de casos en que la cumple es un supraconjunto del conjunto de casos anteriormente resuelto. Adicionalmente, se demuestra que las nuevas medidas también cumplen el resto de las condiciones anteriormente enunciadas.

Palabras clave: algoritmos de agrupamiento solapados, medidas de evaluación, condiciones de validez.

Abstract

In this work, we treat the problem of evaluating overlapping clustering algorithms. After an exhaustive study of the subject, focusing on the existing evaluation measures, as well as the conditions proposed for determining their validity, we identify one of the open problems when this type of measure is applied to overlapping clustering. On that basis, a new condition is enounced to determine the validity of evaluation measures in this case.

Since the measure that according to the literature complies with all previously enounced conditions does not comply with the new condition for all cases, new evaluation measures are proposed, which partially resolve the detected problem. The new measures do not comply with the new condition in 100% of the cases. However, we prove that for one of them the set of cases for which it complies is a superset of the set of previously solved cases. We additionally prove that the new measures also comply with all previously enounced conditions.

Keywords: overlapping clustering algorithms, evaluation measures, validity conditions.

Tabla de Contenido

Introducción

El desarrollo de Internet ha traído aparejado un enorme crecimiento del volumen de información digital a escala mundial, lo que trae como consecuencia la necesidad de herramientas computacionales capaces de procesarla, organizarla, acceder a ella eficientemente, etc., de forma que la información puede ser aprovechada y explotada eficientemente.

La Minería de Datos es la tarea encargada de procesar estas grandes cantidades de información y extraer de ellas el conocimiento útil. Dentro de ésta, una tarea fundamental es el agrupamiento de datos, el cual permite, dada una colección de objetos, organizarlos en un conjunto de grupos de acuerdo con sus semejanzas y diferencias.

En las últimas décadas, los algoritmos de agrupamiento han sido ampliamente aplicadas en esferas como la medicina, para determinar grupos de pacientes que padecen diferentes enfermedades; el procesamiento de noticias, para detectar y seguir los diferentes tópicos y/o sucesos; etc.

Los investigadores han dedicado un importante esfuerzo al desarrollo de medidas confiables para medir la calidad de este tipo de algoritmo. Como resultado de dicho trabajo, se ha propuesto una gran cantidad de estas medidas. Igualmente, se ha trabajado en la propuesta de condiciones que a su vez permitan analizar las medidas de evaluación y determinar su validez, utilidad, etc.

Existe consenso en la comunidad científica en aceptar el conjunto de cuatro condiciones propuestas por Amigó et al. (2008). Estas condiciones son suficientes para las medidas de evaluación cuando los agrupamientos a evaluar son no solapados, o sea, cada objeto pertenece a uno y solo un grupo. Sin embargo, cuando los grupos pueden ser solapados, surgen problemas adicionales que no están cubiertos en su totalidad por dichas condiciones y, por tanto, no son resueltos totalmente por las medidas existentes.

En particular, nos centramos en el problema de que agrupamientos que no son iguales al ideal pueden ser evaluados con el máximo valor por las medidas de evaluación.

Ante esta situación, nos propusimos como objetivo general la formalización en forma de condición del problema existente y la propuesta de nuevas medidas de evaluación que lo manipulen adecuadamente.

Los objetivos específicos son los siguientes:

- Formalizar una nueva condición que exprese el problema analizado en la evaluación de los algoritmos de agrupamiento solapados.
- Proponer nuevas medidas que ataquen el problema.
- Demostrar la validez de las nuevas medidas según las condiciones previamente enunciadas.
- Analizar la mejora que las nuevas medidas introducen con respecto a la nueva condición.

La hipótesis de nuestro trabajo es la siguiente: si se modifican las mejores medidas existentes de modo que se manipulen mejor los casos problemáticos en la evaluación de los algoritmos de agrupamiento solapados, se obtendrán nuevas medidas más eficaces y confiables.

Dado que en las situaciones reales los tipos de agrupamiento que se desea obtener son generalmente solapados, por ejemplo, un paciente puede tener varias enfermedades, disponer de buenas medidas de evaluación para este tipo de algoritmos beneficiará tanto a los investigadores del área como a los desarrolladores de software y empresas que los utilicen y necesiten escoger entre varias opciones.

El presente trabajo está compuesto por dos capítulos. En el Capítulo 1 se describen las medidas de evaluación más usadas actualmente, así como las condiciones enunciadas para determinar su validez. Por su parte, en el Capítulo 2 se presenta el enunciado de la nueva condición propuesta, se describe el proceso de obtención de las nuevas medidas y se demuestra su validez según las condiciones previamente enunciadas, así como las mejoras que introduce según la nueva condición.

Capítulo 1
Evaluación de los algoritmos de agrupamiento

Un algoritmo de agrupamiento divide una colección de objetos en grupos, los cuales se asume que representan su estructura interna. Los algoritmos de agrupamiento pueden clasificarse de las diferentes maneras en dependencia del criterio que se siga (Gil, 2005).

Atendiendo a la forma en que se trata el conjunto de objetos, estos algoritmos se pueden dividir en estáticos, incrementales o dinámicos. Los algoritmos estáticos asumen que la totalidad de los objetos de la colección se conoce antes de su aplicación y que ésta no cambia. Por su parte, los algoritmos incrementales permiten manipular un flujo de objetos, de forma que cada vez que se agrega un objeto a la colección se realizan los cambios necesarios en los grupos sin necesidad de procesar toda la colección nuevamente. Por último, los algoritmos dinámicos permiten tanto la adición de nuevos objetos a la colección como la eliminación de objetos de la misma, igualmente realizando sólo los cambios necesarios en los grupos sin necesidad de procesar toda la colección nuevamente.

Teniendo en cuenta el mecanismo en que se basan para agrupar, los algoritmos de agrupamiento se clasifican en:

- De pasada simple: Los objetos se añaden al grupo a cuyos objetos es más semejante o a un nuevo grupo si no es suficientemente semejante a ninguno de los existentes.
- Basados en grafos: Se construye un grafo cuyos nodos representan objetos y los pesos de sus arcos el nivel de semejanza entre ellos. El proceso de agrupamiento consiste en calcular un cubrimiento del conjunto de nodos de este grafo.

- De optimización: Son algoritmos iterativos que definen un criterio de bondad del agrupamiento para encontrar el cubrimiento óptimo. En esencia, se trata de optimizar una función objetivo que evalúa la calidad de los grupos formados.
- Basados en densidad: Los grupos son los subespacios densos en el espacio de representación de los objetos.
- Basados en árboles: En el proceso de agrupamiento se construye un árbol. Cada vez que se procesa un objeto, se recorre el árbol para determinar el nodo o los nodos en los que éste debe ser colocado. Cada nodo resume las características de los objetos que contiene. Los grupos pueden coincidir con los nodos del árbol o ser una combinación de éstos.

Según la forma en que se organizan los grupos obtenidos, los algoritmos de agrupamiento se pueden clasificar en particionales o jerárquicos. Los algoritmos particionales dividen la colección de objetos en grupos entre los cuales no se asume ninguna relación. Por su parte, los algoritmos jerárquicos producen una secuencia anidada de particiones o cubrimientos del conjunto de objetos donde cada grupo puede verse como la unión de otros grupos. Los algoritmos de agrupamiento jerárquicos se pueden dividir a su vez en aglomerativos o divisivos. Los algoritmos jerárquicos aglomerativos comienzan considerando cada objeto como miembro de un grupo unitario y en cada iteración une al par de grupos más semejante; mientras que los divisivos consideran en un principio a la colección de objetos como un único grupo, y en cada iteración se divide un grupo en dos.

Teniendo en cuenta la pertenencia de los objetos a los grupos, los algoritmos de agrupamiento se dividen en solapados o no solapados (también llamados disjuntos). Los algoritmos no solapados colocan cada objeto en un único grupo, como se muestra en la Figura 1.1; mientras que los solapados permiten que un objeto sea incluido en varios grupos, como puede observarse en la Figura 1.2. En las aplicaciones prácticas, lo más común es que se requiera obtener agrupamientos solapados ya que, por ejemplo, un paciente puede sufrir varias enfermedades, un documento puede tratar sobre varios temas, etc.

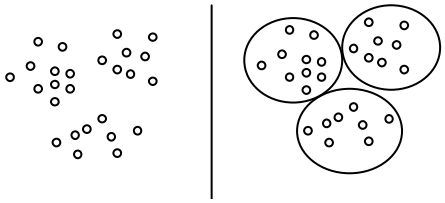

Fig 1.1. Agrupamiento no solapado

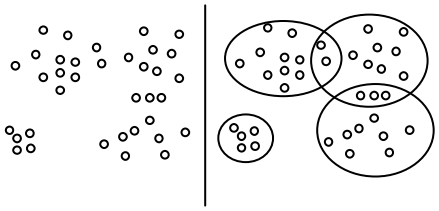

Fig 1.2. Agrupamiento solapado

En la comunidad de investigadores se han desarrollado considerables esfuerzos por desarrollar metodologías adecuadas para medir la calidad de los agrupamientos, producto de lo cual se han propuesto un número importante de medidas de evaluación. Estas medidas se dividen en dos grandes grupos: las intrínsecas, o internas, y las extrínsecas, o externas (Steinbach et al., 2000). Las medidas intrínsecas evalúan los agrupamientos teniendo en cuenta solamente características del conjunto de grupos, tales como la semejanza entre los objetos de un mismo grupo, la diferencia entre los objetos de grupos diferentes, etc. Por su parte, las medidas extrínsecas comparan los agrupamientos obtenidos por los algoritmos con agrupamientos ideales proporcionados por especialistas de las diferentes áreas de aplicación.

Generalmente se ha preferido la evaluación extrínseca, ya que ésta permite determinar hasta qué punto diferentes algoritmos son adecuados para los problemas de interés que surgen en situaciones reales. En este trabajo nos centraremos en este tipo de medidas. En lo que sigue, nos referiremos como *grupos* a los obtenidos por algún algoritmo de agrupamiento, mientras que denominaremos *clases*[1] a los grupos de un agrupamiento ideal.

1.1. Condiciones que deben cumplir las medidas de evaluación

Al elegir una medida de calidad para la evaluación se deben tener en cuenta ciertos requisitos que demuestren su validez. En la literatura se han reportado varias propuestas de conjuntos de condiciones.

Dom (2001) propone cinco condiciones que descomponen el conjunto de grupos G en dos subconjuntos G_{useful} y G_{noise}, donde se incluyen los grupos útiles y ruidosos, respectivamente, y una variable de error asociada a cada uno de estos dos conjuntos. El

[1] Otros autores los denominan *categorías* o *tópicos* en el caso de las colecciones de documentos.

proceso de decisión para incluir los grupos en uno de estos conjuntos y estimar las variables de error utiliza un criterio probabilístico. Rosenberg et al. (2007) proponen extender las condiciones establecidas por Dom agregándole dos nuevas restricciones.

Por su parte, Meila (2003) propone otro conjunto de 12 condiciones las cuales no están basadas en aplicaciones específicas sobre agrupamientos, sino que están enfocadas a aspectos intrínsecos de las medidas.

Amigó et al. (2008) proponen cuatro condiciones, las cuales demuestran que cubren las propuestas por otros autores. Actualmente hay consenso acerca de la utilización de estas cuatro condiciones para analizar las medidas de evaluación, por lo cual las tomaremos como base en nuestro trabajo.

Cada condición se expresa en forma de una preferencia, donde dado un par de agrupamientos (D_1, D_2), en el cual se considera que D_2 es mejor que D_1, una medida de evaluación Q que satisfaga la condición debe cumplir que $Q(D_1) < Q(D_2)$.

A continuación se analizan las condiciones propuestas por Amigó et al.

1.1.1. Homogeneidad

Según esta condición, la calidad de un agrupamiento cuyos grupos mezclen objetos de diferentes clases debe ser menor que la calidad de un agrupamiento cuyos grupos no los mezclen, como se muestra en la Figura 1.3. Formalmente, sea S un conjunto de objetos pertenecientes a las clases $C_1,\ C_2 \ldots, C_n$. Sea D_1 un agrupamiento, uno de cuyos grupos G contiene objetos de dos clases C_i y C_j. Sea D_2 un agrupamiento idéntico a D_1, excepto que el grupo G es dividido en dos grupos, uno que contiene sólo objetos de C_i y otro sólo objetos de C_j. Una medida de evaluación Q que satisfaga la condición de homogeneidad debe cumplir que $Q(D_1) < Q(D_2)$.

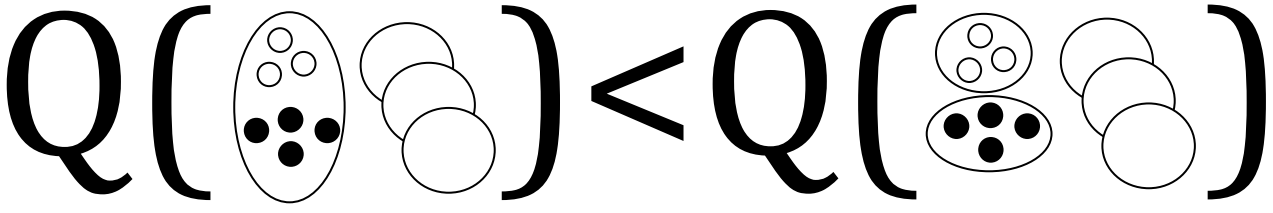

Fig 1.3. Homogenenidad

1.1.2. Completitud

Según esta condición, la calidad de un agrupamiento en el cual objetos de una clase son colocados en diferentes grupos es inferior a la de uno en el cual estos son colocados en el mismo grupo, como se muestra en la Figura 1.4. Formalmente, sea D_1 un agrupamiento, dos de cuyos grupos G_1 y G_2 sólo contienen objetos pertenecientes a una clase C. Sea D_2 un agrupamiento idéntico a D_1, excepto que G_1 y G_2 son mezclados en un mismo grupo. Una medida de evaluación Q que satisfaga la condición de completitud debe cumplir que $Q(D_1) < Q(D_2)$.

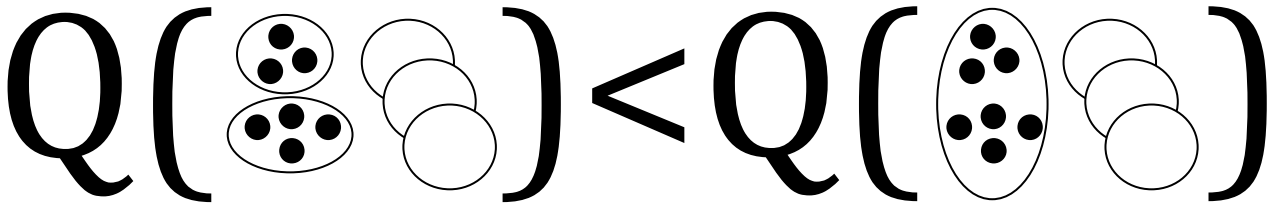

Fig 1.4. Completitud

1.1.3. Saco de Ruido

Según esta condición, la calidad de un agrupamiento en el cual se introduce un objeto incorrectamente en un grupo que sería homogéneo sin ese objeto es inferior a la de uno que introduce un objeto incorrectamente en un grupo ruidoso, como se muestra en la Figura 1.5. Formalmente, sea G_{clean} un grupo con n objetos pertenecientes a la misma clase C. Sea G_{noise} un grupo con n objetos de clases unarias (debe existir un objeto por cada clase). Sea o_{noise} un objeto perteneciente a una clase unitaria. Sea D_1 un agrupamiento con los grupos $G_{clean} \cup \{o_{noise}\}$ y G_{noise} y D_2 un agrupamiento con los grupos G_{clean} y $G_{noise} \cup \{o_{noise}\}$. Una medida de evaluación Q que satisfaga la condición de saco de ruido debe cumplir que $Q(D_1) < Q(D_2)$.

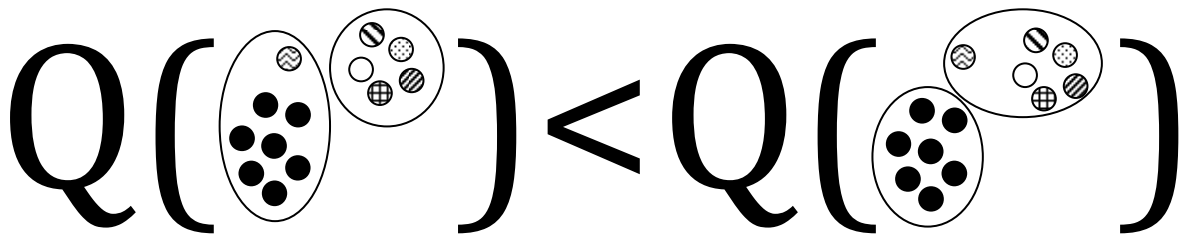

Fig 1.5. Saco de ruido

1.1.4. Tamaño contra calidad

Según esta condición, un error pequeño en un grupo grande es preferible que varios errores en grupos pequeños, como se muestra en la Figura 1.6. Formalmente, sea D un agrupamiento que contiene un grupo G_l con $n+1$ objetos pertenecientes a la misma clase C_1 y n grupos adicionales $G_1, G_2, \dots, G_n$ donde cada uno de ellos contiene dos objetos de la misma clase $C_1, C_2, \dots, C_n$. Sean D_1 un nuevo agrupamiento idéntico a D excepto en que cada grupo G_i es dividido en dos grupos unitarios y D_2 un agrupamiento idéntico a D excepto en que el grupo G_l es dividido en un grupo de tamaño n y un grupo de tamaño1. Una medida de evaluación Q que satisfaga la condición de tamaño contra calidad debe cumplir que $Q(D_1) < Q(D_2)$.

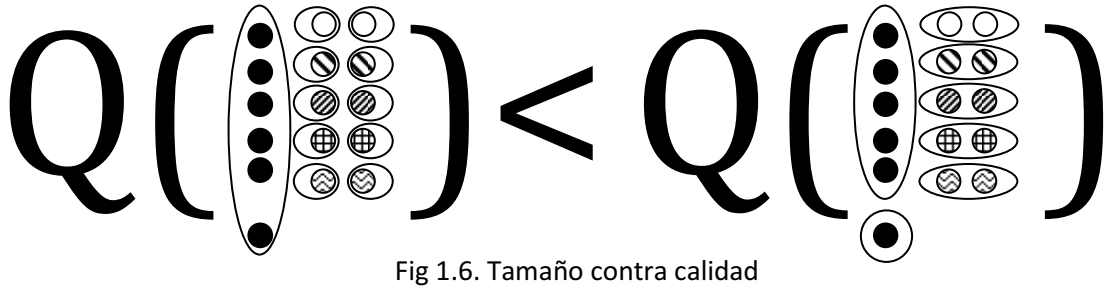

Fig 1.6. Tamaño contra calidad

1.2. Medidas de evaluación existentes

Como se mencionó anteriormente, los algoritmos de agrupamiento obtienen cubrimientos de un conjunto de objetos. En particular, los algoritmos no solapados obtienen una partición.

Formalmente, siendo U el conjunto de objetos, un agrupamiento D lo divide en K grupos $G_1, G_2, \dots, G_K$, tales que $\bigcup_k^K G_k = U$. En el caso particular de los algoritmos no solapados, $G_k \cup G_l = \emptyset$ para todo $k, l \leq K, k \neq l$. La cantidad de objetos en U y G_k se denotará en lo adelante mediante n y n_k, respectivamente. Sea $D_l = \{C_1, C_2, \dots, C_T\}$ el agrupamiento ideal de U contra el cual se desea comparar a D. La cantidad de objetos en C_t se denotará en lo adelante mediante n_t. Se asume que las clases y los grupos son no vacíos, o sea $n_k > 0$ y $n_t > 0$. En todos los casos, $1 \leq k \leq K$, $1 \leq t \leq T$.

Todo criterio de comparación de agrupamientos puede ser descrito a través de una tabla de contingencia[2] entre el par D y D_I. La tabla de contingencia es una matriz de T filas y K columnas donde el tk-ésimo valor es la cantidad de objetos que pertenecen a la intersección de la clase C_t el grupo G_k (Meila, 2003), o sea,

$$n_{tk} = |C_t \cap G_k|$$

1.2.1. Medidas basadas en el emparejamiento de conjuntos

Las medidas pertenecientes a esta familia comparan par a par los grupos y las clases para medir la calidad de los agrupamientos.

Meila (2003) define la medida de evaluación H, donde cada grupo G_k se corresponde con la clase C_t con la que tiene el mejor emparejamiento. Si definimos $match(k)$ como la función que devuelve para cada grupo G_k este mejor emparejamiento, entonces

$$H = \frac{1}{n} \sum_{t=match(k)} n_{kt}$$

Meila también propone las medidas de evaluación L y D, las cuales se basan en el mismo principio y se definen como sigue:

$$L = \frac{1}{K} \sum_{k} \max_{t} \frac{2n_{tk}}{n_k + n_t}$$

$$D = 2n - \sum_{k} \max_{t} n_{tk} - \sum_{t} \max_{k} n_{tk}$$

Las medidas más populares, provenientes de la Recuperación de Información, son la precisión, la relevancia y la medida F_1, las cuales se definen como sigue:

$$Precisión(C_t, G_k) = \frac{|C_t \cap G_k|}{G_k}$$

$$Relevancia(C_t, G_k) = \frac{|C_t \cap G_k|}{C_t}$$

$$F_1(C_t, G_k) = \frac{2 \cdot Relevancia(C_t, G_k) \cdot Precisión(C_t, G_k)}{Relevancia(C_t, G_k) + Precisión(C_t, G_k)}$$

[2] También conocida como *matriz de confusión* o *matriz de asociación*

La medida F_1 se corresponde con la media armónica entre los valores de precisión y relevancia, por lo que favorece a los algoritmos que no sacrifican uno de estos valores a favor del otro. Utilizando estas medidas se compara un grupo C_t con una clase G_k. Para evaluar el agrupamiento en general se parte de la asociación de cada clase C_t con el grupo $\sigma(C_t)$ con el cual tiene el máximo valor de la medida F_1, o sea

$$\sigma(C_t) = \arg\max_k F(C_t, G_k)$$

La medida F_1 Macro-promediada se calcula como la media de la medida F_1 sobre todas las clases, asociándolas con su grupo mejor emparejado.

$$macroF_1 = \frac{1}{T}\sum_{t=1}^{T} F_1(C_t, \sigma(C_t))$$

Por otra parte, la F_1 Micro-promediada le da el mismo peso a cada objeto y, por tanto, se considera un promedio por objeto (Pons, 2004).

$$microF = \frac{2 \cdot microP(C,G) \cdot microR(C,G)}{microP(C,G) + microR(C,G)}$$

donde

$$microP = \frac{1}{T}\sum_{t=1}^{T} \frac{Relevancia(C_t, G_k) \cdot Precisión(C_t, G_k)}{Precisión(C_t, \sigma(C_t))}$$

$$microR = \frac{1}{T}\sum_{t=1}^{T} \frac{Relevancia(C_t, G_k) \cdot Precisión(C_t, G_k)}{Relevancia(C_t, G_k)}$$

Por último, la medida F_1 global se calcula como la media ponderada de la medida F_1 donde cada clase se asocia también con su grupo mejor emparejado.

$$globalF_1 = \sum_{t=1}^{T} \frac{|C_t|}{n} F_1(C_t, \sigma(C_t))$$

Las medidas pureza y pureza inversa también están relacionadas con la precisión y la relevancia, respectivamente, de la siguiente manera:

$$Pureza(C_t, G_k) = \sum_k \frac{|G_k|}{K} \max_t \ Precisión(C_t, G_k)$$

$$Pureza\ Inversa(C_t, G_k) = \sum_t \frac{|C_t|}{T} \max_k \ Relevancia(C_t, G_k)$$

De forma análoga a la precisión y la relevancia, la pureza y la pureza inversa pueden combinarse mediante la medida F_1.

Según Meila (2003), el principal problema de esta tipo de medidas es que, si bien encuentran el mejor emparejamiento para cada grupo, ignoran completamente las partes de cada grupo que no son emparejadas, aún cuando intuitivamente éstas deberían influir en la puntuación final.

1.2.2. Medidas de evaluación basadas en conteo de pares

Las medidas de esta familia se basan en el número de pares de objetos (x_u, y_v) que pertenecen tanto a los grupos evaluados como a las clases. En este caso se puede tomar en cuenta la tabla de contingencia como la matriz $A = \{a_{ij} \,|i, j \in \{0,1\}\}$ donde cada a_{ij} se corresponde con el número de pares comunes entre los conjuntos comparados. Así, el valor de a_{00} es la cantidad de pares de objetos que pertenecen tanto a C como a G, a_{11} la cantidad de pares de objetos que no pertenecen ni a C ni a G, etc.

	G	$\neg G$
C	a_{00}	a_{01}
$\neg C$	a_{10}	a_{11}

Teniendo como base esta tabla se han propuesto varias medidas de evaluación (Halkidi, 2001; Fowlkes et al., 1983), las cuales se muestran a continuación:

$$Rand = \frac{a_{00} + a_{11}}{a_{00} + a_{10} + a_{01} + a_{11}}$$

$$Jaccard = \frac{a_{00}}{a_{00} + a_{10} + a_{01}}$$

$$Fowlkes\ and\ Mallows = \sqrt{\frac{a_{00}}{a_{00} + a_{01}} \frac{a_{00}}{a_{00} + a_{10}}}$$

El problema fundamental de las medidas basadas en pares se debe a que existe una dependencia cuadrática entre el tamaño de los grupos y la cantidad de pares, debido a lo cual los valores cambian mucho cuando grupos grandes son fragmentados o unidos (Amigó et al., 2008).

1.2.3. Medidas basadas en la Entropía

La entropía (Shannon, 1948) es una medida de la cantidad media de información que contiene una variable aleatoria. La información que aporta un determinado valor x_i de una variable aleatoria discreta X se define como

$$I(x_i) = -log_2\, p(x_i)$$

Así, la entropía se define como

$$H(X) = -\sum_i p(x_i)\ log_2\, p(x_i)$$

Siendo $p(t,k)$ la probabilidad de encontrar un objeto de la clase C_t en el grupo G_k. Así, la entropía para un grupo se calcula como

$$H(G_k) = -\sum_t p(t,k)\ \ log_2\, p(t,k)$$

La entropía condicional se ha utilizado como medida de comparación entre agrupamientos. Ésta se define como sigue:

$$H(D_I|D) = -\sum_t \sum_k p(t,k)\ log_2\, p(k|t)$$

La medida Información Mutua (Xu et al., 2003) estima la cantidad de información en común entre los agrupamientos a comparar y se define como

$$I = \sum_{t=1}^{T}\sum_{k=1}^{K} P(t,k)\ log\frac{P(t,k)}{P(t)P(k)}$$

La medida Variación de Información (Meila, 2003) se basa en la cantidad de información ganada y perdida al cambiar una clase C por un grupo G y se define como

$$VI(C,G) = H(C|G) + H(G|C)$$

El término $H(C|G)$ se corresponde con la cantidad de información perdida al cambiar C por G, mientras que el término $H(G|C)$ se corresponde con la información ganada, según se muestra en la Figura 1.7.

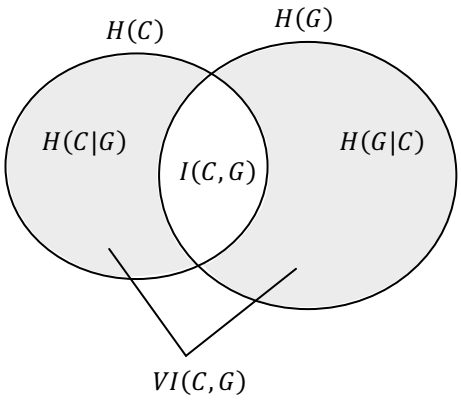

Fig 1. 7. Variación de la información representada con el área sombreada

Las medidas de esta familia expuestas anteriormente no tienen en cuenta la completitud de los grupos. La medida V (Rosenberg et al., 2007) combina a su vez dos medidas, una que mide la homogeneidad y otra que mide la completitud, de una forma similar a como lo hace la medida F_1. Ésta se define como sigue:

$$h = \begin{cases} 1 & si\ H(C',C) = 0 \\ 1 - \dfrac{H(C|C')}{H(C)} & e.o.c. \end{cases} \qquad c = \begin{cases} 1 & si\ H(C',C) = 0 \\ 1 - \dfrac{H(C|C')}{H(C')} & e.o.c. \end{cases}$$

$$V_\beta = \frac{(1+\beta) * h * c}{(\beta * h) + c}$$

Si $\beta > 1$, la completitud tiene más peso en el cálculo, mientras que la homogeneidad tiene más peso si $\beta > 1$.

Dado que la entropía mide fundamentalmente la homogeneidad, prestando menos atención a otras propiedades, Dom (2001) propone la medida Q_0, la cual incluye un modelo de costo de términos que es sumado al valor de la entropía.

$$Q_0(C,G) = H(C|G) + \frac{1}{n}\sum_{k}^{K} log\binom{h(k) + |C| - 1}{|C| - 1}$$

Dom también presenta una versión normalizada

$$Q_2(C,G) = \frac{\frac{1}{n}\sum_{t=1}^{T} log\binom{h(t)+|C|-1}{|C|-1}}{Q_0(C,G)}$$

1.2.4. Medidas basadas en distancia de edición

Estas medidas basan su análisis en la cantidad de transformaciones a realizar para transformar un grupo G en la clase C. Pantel et al. (2002) proponen los siguientes pasos para la transformación:

- Crear un conjunto vacío asociado a cada clase.
- Agregar los objetos de un grupo al conjunto con cuya clase asociada tenga más objetos en común.
- Mover los objetos ubicados incorrectamente a los conjuntos que le correspondan.

Cada acción de unir dos grupos o mover un objeto de un grupo a otro se considera como una transformación a contar en la distancia de edición.

Esta medida no satisface la condición Saco de Ruido, pues, por ejemplo, considera como iguales las acciones de mover un objeto a un grupo ruidoso o a uno homogéneo.

1.2.5. Medidas de evaluación híbridas

Como su nombre lo indica, estas medidas combinan ideas de varias de las familias mencionadas anteriormente.

La medida BCubed (Bagga et al., 1998) parte de una relación entre objetos denominada correctitud, la cual se cumple para pares de objetos que pertenecen tanto al mismo grupo como a la misma clase.

$$Correctitud(\mathrm{e},\mathrm{e}') = \begin{cases} 1 & si\ C(e) = C(o') \leftrightarrow G(o) = G(o') \\ 0 & ecc \end{cases}$$

Como tal BCubed no es en sí una sola medida, sino una forma diferente de calcular precisión y relevancia. La precisión BCubed de un objeto es la proporción de objetos que comparten grupo con él (incluido él mismo) que tienen su misma clase. La precisión BCubed del agrupamiento es el promedio de estos valores para todos los objetos. Por su

parte, la relevancia BCubed de un objeto es la proporción de objetos que comparten clase con él (incluido él mismo) que son ubicados en el mismo grupo. La relevancia BCubed del agrupamiento es el promedio de estos valores para todos los objetos. Ambas medidas se definen como sigue:

$$Precisión\ BCubed\ = Avg_o\left\{Avg_{o'.C(o)=C(o')}\{Correctitud(o,o')\}\right\}$$

$$Relevancia\ BCubed\ = Avg_o\left\{Avg_{o'.L(o)=L(o')}\{Correctitud(o,o')\}\right\}$$

donde o y o' son objetos. De la misma forma que precisión y relevancia se combinan normalmente con F_1, precisión BCubed y relevancia BCubed se combinan con la medida F_α, la cual se define como

$$F_\alpha\ BCubed = \frac{1}{\propto\left(\frac{1}{Precisión\ BCubed}\right)+(1-\propto)\left(\frac{1}{Relevancia\ BCubed}\right)}$$

Luego de analizar la mayoría de las medidas de evaluación más utilizadas, Amigó et al. (2008) demuestran que F_α BCubed con $\propto = 0.5$ es la única que cumple con las cuatro condiciones descritas en la Sección 1.1, por lo que se considera como la más adecuada.

En la definición de BCubed se asume que el agrupamiento es no solapado. BCubed Extendido (Amigó et al., 2008) es una generalización de esta medida a cualquier tipo de agrupamiento, ya sea solapado o no. BCubed Extendido (en lo adelante EBCubed) sigue la misma idea de BCubed, pero adicionalmente tiene en cuenta que si dos objetos comparten n grupos entonces deben compartir n clases y viceversa. EBCubed se basa en las funciones auxiliares siguientes:

$$Precisión\ Multiplicidad\ \ (o,o') = \frac{\min\left(|G(o)\cap G(o')|,|C(o)\cap C(o')|\right)}{|G(o)\cap G(o')|}$$

$$Relevancia\ Multiplicidad(o,o') = \frac{\min\left(|G(o)\cap G(o')|,|C(o)\cap C(o')|\right)}{|C(o)\cap C(o')|}$$

a partir de las cuales se calcula precisión EBCubed y relevancia EBCubed como sigue:

$$Precisión\ EBCubed\ = Avg_o\left\{Avg_{o'.C(o)\cap C(o')\neq\emptyset}\{Precisión\ Multiplicidad\ \ (o,o')\}\right\}$$

$$Relevancia\ EBCubed\ = Avg_o\left\{Avg_{o'.L(o)\cap L(o')\neq\emptyset}\{Relevancia\ Multiplicidad\ (o,o')\}\right\}$$

Para agrupamientos no solapados, los valores obtenidos por precisión BCubed y relevancia BCubed son iguales a los obtenidos por precisión EBCubed y relevancia EBCubed, respectivamente.

Capítulo 2
Medidas de calidad propuestas

Los algoritmos de agrupamiento solapados encuentran mayor aplicabilidad en problemas reales que los algoritmos de agrupamiento no solapados. Por ejemplo, en la medicina, es usual que un paciente padezca varias enfermedades. Debido a esto, es especialmente importante contar con algoritmos solapados de calidad, lo cual hace necesario contar con medidas de evaluación confiables para estos tipos de algoritmo.

Un problema común señalado a los algoritmos de agrupamiento solapados actuales es que el nivel de solapamiento que obtienen es excesivo. La principal implicación de este problema es que mediante este excesivo solapamiento es común que se enmascaren errores o deficiencias inherentes al algoritmo. Por ejemplo, al colocar un objeto en varios grupos se aumenta la probabilidad de que alguna o algunas de de esas decisiones sea correcta, pero igualmente aumenta la probabilidad de que algunas de las decisiones sea incorrecta. El efecto de esto durante la evaluación en ocasiones puede interpretarse como una especie de “engaño” a la medida de evaluación.

En nuestro trabajo abordamos uno de los problemas que se presenta al evaluar agrupamientos solapados, el cual consiste en que algunos agrupamientos diferentes del ideal pueden aún obtener el valor óptimo de la medida de evaluación utilizada. Este comportamiento deseado es tratado en nuestro trabajo como una nueva condición, la cual denominamos **Obtención de óptimo sólo para emparejamiento perfecto** y enunciamos como sigue: una medida debe calificar a un agrupamiento con puntuación máxima si y sólo si este agrupamiento es perfecto.

A continuación, mostramos un ejemplo de un agrupamiento en el cual F_α EBCubed, que como se vio en el Capítulo 1 puede considerarse como la medida intuitivamente más adecuada por cumplir todas las condiciones de Amigó et al. y manipular tanto agrupamientos solapados como no solapados, no satisface esta nueva condición.

En la Figura 2.1 se muestra un agrupamiento el cual, comparado con el agrupamiento ideal mostrado según F_α EBCubed, obtendrá un valor de 1 aún cuando son diferentes.

$C_1 = 1,2,4$ $\quad$ $G_1 = 1,3,4$

$C_2 = 1,3$ $\quad$ $G_2 = 1,2$

$C_3 = 4,3$ $\quad$ $G_3 = 4,2$

$C_4 = 2,5$ $\quad$ $G_4 = 3,5$

$C_5 = 3,5,6$ $\quad$ $G_5 = 2,5,6$

$C_6 = 2,6$ $\quad$ $G_6 = 3,6$

Fig 2.1. Ejemplo en el cual F- EBCubed no cumple la nueva condición

Cuando F_α EBCubed compara los agrupamientos de la Figura 2.1, donde cada par de objetos co-ocurre en la misma cantidad de grupos que de clases, da como resultado la máxima puntuación. Sin embargo, como puede observarse, estos agrupamientos son distintos, por lo cual F_α EBCubed no cumple con la nueva condición propuesta en este trabajo.

Para resolver parcialmente ese problema, proponemos un conjunto de tres nuevas medidas de evaluación, una de las cuales, si bien no cumple la nueva condición en todos los casos, sí lo hace en un supraconjunto de los casos en que la cumple F_α EBCubed.

2.1. Medidas de evaluación SBCubed

Las nuevas propuestas de medidas de evaluación, denominadas en su conjunto SBCubed, constituyen una extensión de EBCubed. Mientras que en EBCubed se tiene en cuenta la cantidad de pares de objetos que coinciden en cuanto a grupo y clase, nuestra propuesta, además de mantener esta característica, analiza la coincidencia de los objetos con los cuales cada par comparte grupo y clase.

Para realizar la comparación se realizan tres puntuaciones: a los pares de objetos, a cada objeto de forma individual y al agrupamiento. Un par es evaluado para la precisión

SBCubed (en lo adelante P-SBCubed) y la relevancia SBCubed (en lo adelante R-SBCubed), respectivamente, como se muestra a continuación:

$$P_{prec}(o_1, o_2) = \frac{\min(|G(o_1) \cap G(o)|, |C(o_1) \cap C(o_2)|) + |E(o_1, o_2, D) \cap E(o_1, o_2, D_I)|}{|G(o_1) \cap G(o_2)| + |E(o_1, o_2, D) \cup E(o_1, o_2, D_I)|}$$

$$P_{rel}(o_1, o_2) = \frac{\min(|G(o_1) \cap G(o_2)|, |C(o) \cap C(o_2)|) + |E(o_1, o_2, D) \cap E(o_1, o_2, D_I)|}{|C(o_1) \cap C(o_2)| + |E(o_1, o_2, D) \cup E(o_1, o_2, D_I)|}$$

Con el objetivo de tratar la deficiencia señalada a EBCubed, se agregaron dos sumandos en la fracción que evalúa los pares, de forma que los objetos que comparten un grupo o una clase con cada par sean tomados en cuenta. En el numerador se añadió $|E(o, o', D) \cap E(o, o', D_I)|$, que se corresponde con la cantidad de objetos que comparten al menos una clase, y además, al menos un grupo con el par. Por otro lado, en el denominador se añadió $|E(o, o', D) \cup E(o, o, D_I)|$, que se corresponde con la cantidad de objetos que comparten al menos una clase o al menos un grupo con el par.

La evaluación de un objeto se realiza a partir del promedio de la puntuación de los pares que se pueden formar con el objeto evaluado y los objetos que comparten un grupo o una clase con él.

$$Avg_p(o) = \frac{1}{|\bigcup_{g \in G(o)} g|} \sum_{o' \in E(o,G)} P_{prec}(o_1, o_2)$$

$$Avg_r(o) = \frac{1}{|\bigcup_{g \in G(o)} g|} \sum_{o' \in E(o,C)} P_{rel}(o_1, o_2)$$

La evaluación del agrupamiento se obtiene a través del promedio de todos los pares de objetos que se encuentran en los grupos o clases, según el criterio que se aplique. Las medidas P-SBCubed y la relevancia R-SBCubed se definen como sigue:

$$P - SBCubed = \frac{1}{|D|} \sum_{o \in D} Avg_p(o)$$

$$R - SBCubed = \frac{1}{|D|} \sum_{o \in D} Avg_r(o)$$

Donde

- $G(o)$: Conjunto de grupos que contienen al objeto o
- $C(o)$: Conjunto de clases que contienen al objeto o
- G: Agrupamiento a evaluar
- C: Agrupamiento ideal
- $E(o_1, o_2, A)$: Unión de todos los grupos que contienen a los objetos o y o_2 en A.
- $E(o, A)$: Unión de todos los grupos de A que contienen a o_1

Las medidas P-SBCubed y R-SBCubed se combinan igualmente mediante la medida F_α SBCubed, la cual se define como

$$F_\alpha\, SBCubed = \frac{1}{\propto \left(\frac{1}{P-\ SBCubed}\right) + (1-\propto)\left(\frac{1}{R-\ SBCubed}\right)}$$

Siguiendo la idea de Amigó et al., en este trabajo tomamos $\alpha = 0.5$.

2.2. Demostración del cumplimiento de las condiciones de Amigó et al. por las medidas SBCubed

A diferencia de EBCubed, para el cual sólo F_α cumple las cuatro condiciones (según Amigó et al. para $\alpha = 0.5$), tanto P-SBCubed como R-SBCubed cumplen cada una por separado las cuatro condiciones.

A continuación se demuestra que SBCubed cumple con las condiciones de Amigó et al. Como se explicó en el Capítulo 1, cada condición se expresa en forma de una preferencia, donde dado un par de agrupamientos (D_1, D_2), en el cual se considera que D_2 es mejor que D_1, una medida de evaluación Q que satisfaga la condición debe cumplir que $Q(D_1) < Q(D_2)$.

En las demostraciones se analizarán a los criterios de precisión y relevancia de forma individual concluyendo que ambos cumplen con todas las condiciones.

2.2.1. Homogeneidad

Cómo se discutió en el Capítulo 1, según esta condición, la calidad de un agrupamiento cuyos grupos mezclen objetos de diferentes clases debe ser menor que la calidad de un agrupamiento cuyos grupos no los mezclen.

Según la condición, el agrupamiento D_1 contiene un grupo G_k con n objetos pertenecientes a dos clases C' y C'', mientras que el agrupamiento D_2 es idéntico a D_1 excepto en que el grupo G_k queda dividido en dos grupos G' y G'' de tamaños n_1 y n_2, respectivamente, tales que $n_1 + n_2 = n$, uno que contiene los objetos de la clase C' y el otro los de C''. A su vez, $|C'| = N_1$ y $|C''| = N_2$.

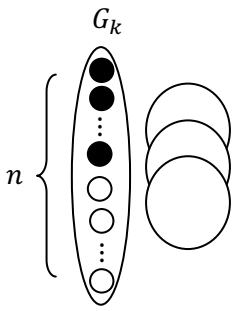

Fig 2.2. Agrupamiento D_1

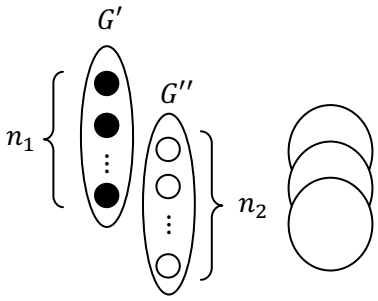

Fig 2.3. Agrupamiento D_2

A continuación se demuestra que $P - SBCubed(D_1) < R - SBCubed(D_2)$. La evaluación de los agrupamientos D_1 y D_2 se obtiene promediando la puntuación de todos sus objetos. Para estos agrupamientos podemos afirmar que sólo los objetos pertenecientes a G_k son evaluados de forma distinta en ambos agrupamientos, y por tanto también los de G' y G''. Nuestro análisis se centra en demostrar que los objetos de G_k son penalizados más que los objetos de G' y G''.

En D_1, todos los pares (o_i, o') que pertenezcan al grupo G_k y a la clase C' cumplen que $|E(o_i, o', D_1) \cap E(o_i, o', D_I)| = n_1$ y además que $|E(o_i, o', D_1) \cup E(o_i, o', D_I)| = N_1 + n_2$. Los pares con objetos de clases distintas son medidos con la mínima puntuación[3] por no compartir ninguna clase en común.

El promedio individual para cada objeto o_i tal que $o_i \in G_k \wedge o_i \in C'$ sería,

$$Avg_{o_i} = \frac{1}{n}\left(\sum_{1}^{n_1} \frac{n_1 + 1}{N_1 + n_2 + 1}\right) = \frac{n_1(n_1 + 1)}{n(N_1 + n_2 + 1)}$$

[3] La puntuación mínima que puede obtener un par de objetos, un objeto y un agrupamiento, es 0.

De forma similar, para cada $o_i \in G_k \wedge o_i \in C''$ sería,

$$\overline{Avg_{o_i}} = \frac{1}{n}\left(\sum_{1}^{n_2} \frac{n_2+1}{N_2+n_1+1}\right) = \frac{n_2(n_2+1)}{n(N_2+n_1+1)}$$

Para el agrupamiento D_2, todos los pares de G' pertenecen a una misma clase, por tanto, $|E(o_i, o', D_2) \cap E(o_i, o', D_I)| = n_1$ y $|E(o_i, o', D_2) \cup E(o_i, o', D_I)| = N_1$. El promedio individual de cada objeto $o_i \in G'$ sería

$$Avg'_{o_i} = \frac{1}{n1}\left(\sum_{1}^{n_1} \frac{n_1+1}{N_1+1}\right) = \frac{1}{n1}\left(n_1 \cdot \frac{n_1+1}{N_1+1}\right) = \frac{n_1+1}{N_1+1}$$

Y para cada $o_i \in G''$ sería

$$\overline{Avg'_{o_i}} = \frac{1}{n2}\left(\sum_{1}^{n_2} \frac{n_2+1}{N_2+1}\right) = \frac{1}{n2}\left(n_2 \cdot \frac{n_2+1}{N_2+1}\right) = \frac{n_2+1}{N_2+1}$$

Comparando las expresiones mediante el producto cruzado, obtenemos

$n_1(n_1+1)(N_1+1) \quad n(n_1+1)(N_1+n_2+1)$	$n_2(n_2+1)(N_2+1) \quad n(n_2+1)(N_1+n_1+1)$
$= n_1(N_1+1) \quad = n(N_1+n_2+1)$	$= n_2(N_2+1) \quad = n(N_2+n_1+1)$
como $n_1 \leq n$, entonces	como $n_2 \leq n$ entonces
$Avg_{o_i} \leq Avg'_{o_i}$	$\overline{Avg_{o_i}} \leq \overline{\overline{Avg'_{o_i}}}$

Por tanto, $P - SBCubed(D_1) < P - SBCubed(D_2)$.

Para demostrar que $R - SBCubed(D_1) < R - SBCubed(D_2)$, debe tenerse en cuenta que la condición de homogeneidad no exige que G_k contenga todos los objetos de las clases G' y G'', sino que otros grupos también pueden contener objetos de estas dos clases. La medida R- SBCubed, a diferencia de P-SBCubed, realiza el análisis tomando los pares de objetos que comparten una clase, de forma que todos los pares de G' y G'' no necesariamente tienen que estar contenidos en G_k. Por tanto, no se puede tomar a la puntuación de los objetos como punto de referencia en la demostración del cumplimiento de la condición de homogeneidad para R-SBCubed.

Un par que se encuentre en una misma clase pero en diferentes grupos es evaluado por R-SBCubed con su mínima puntuación, ya que la intersección del grupo y la clase a la que pertenece el par es vacía. La evaluación de cada objeto se realiza a través del promedio de los pares formados por el objeto a evaluar y los objetos que comparten una clase con él. Así, la demostración se basa en justificar que estos pares de G' y G'', que pertenecen a G_k son más penalizados en D_1 que en D_2.

El análisis para estos pares se realiza de la misma forma al calcular P-SBCubed y R-SBCubed, por tanto

Todo (o_i, o') de la clase C' que se encuentran en G_k es medido para D_1 como

$$P_1(o_i, o') = \frac{n_1 + 1}{N_1 + n_2 + 1}$$

Todo (o_i, o') de la clase C'' que se encuentran en G_k es medido para D_1 como

$$\bar{P}_1(o_i, o') = \frac{n_2 + 1}{N_2 + n_1 + 1}$$

Todo (o_i, o') de la clase C' que se encuentran en G_k es medido para D_2 como

$$P_2(o_i, o') = \frac{n_1 + 1}{N_1 + 1}$$

Todo (o_i, o') de la clase C'' que se encuentran en G_k es medido para D_2 como

$$\bar{P}_2(o_i, o') = \frac{n_2 + 1}{N_2 + 1}$$

Como n_1,$n_2 > 0$ entonces $P_1(o_i, o') < P_2(o_i, o')$ y $\bar{P}_1(o_i, o') < \bar{P}_2(o_i, o')$. Por tanto, $R - SBCubed(D_1) < R - SBCubed(D_2)$.

2.2.2. Completitud

Como se explicó en el Capítulo 1, según esta condición, la calidad de un agrupamiento en el cual objetos de una clase son colocados en diferentes grupos es inferior a la de uno en el cual estos son colocados en el mismo grupo.

Según la condición, se tiene un agrupamiento D_1 donde dos grupos G' y G'', de tamaño n_1 y n_2 respectivamente, contienen objetos de una misma clase C_k, de tamaño N. Se tiene además un agrupamiento D_2 idéntico a D_1 excepto en que los grupos G' y G'' son unidos en un solo grupo de tamaño n, tal que $n_1 + n_2 = n$. A continuación se demuestra que tanto la P-SBCubed como R-SBCubed privilegian a D_2 sobre D_1.

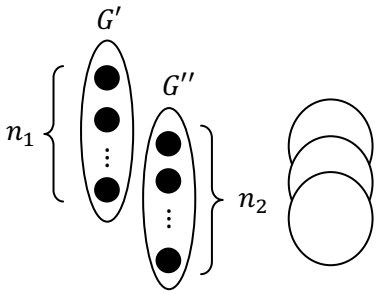

Fig 2.4. Agrupamiento D_1

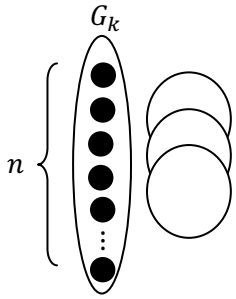

Fig 2.5. Agrupamiento D_2

En el cálculo de R-SBCubed para D_1, se tomarán en cuenta los pares que se encuentran en los grupos G' y G'', donde cada objeto del par pertenece a una misma clase C_k, por tanto $|E(o_i, o', D_1) \cap E(o_i, o', D_I)|$ toma como valor el tamaño del grupo a que pertenecen y $|E(o_i, o', D_1) \cup E(o_i, o', D_I)|$ el tamaño de C_k.

Así, cada $o_1 \in G'$ y $\bar{o}_1 \in G''$ toma como promedio en D_1

$$Avg_{o_1} = \frac{1}{n_1}\left(\sum_{1}^{n_1} \frac{n_1 + 1}{N + 1}\right) = \frac{1}{n_1}\left(n_1 \cdot \frac{n_1 + 1}{N + 1}\right) = \frac{n_1 + 1}{N + 1}$$

$$Avg_{\bar{o}_1} = \frac{1}{n_2}\left(\sum_{1}^{n_1} \frac{n_2 + 1}{N + 1}\right) = \frac{1}{n_2}\left(n_2 \cdot \frac{n_2 + 1}{N + 1}\right) = \frac{n_2 + 1}{N + 1}$$

En el agrupamiento D_2, estos dos grupos son unidos en C_k. Para los pares de objetos (o_i, o') tales que ambos objetos pertenecen a C_k, $|E(o_i, o', D_2) \cap E(o_i, o', D_I)|$ toma como valor el tamaño del grupo C_k y $|E(o_i, o', D_2) \cup E(o_i, o', D_I)|$ el tamaño de la clase que comparten. Para todo $o_k \in G_k$, el valor del promedio individual será

$$Avg_{o_k} = \frac{1}{n}\left(\sum_{1}^{n} \frac{n + 1}{N + 1}\right) = \frac{1}{n}\left(n \cdot \frac{n + 1}{N + 1}\right) = \frac{n + 1}{N + 1}$$

Como $n_1 + n_2 = n$, entonces

$$\frac{n_1 + 1}{N + 1} < \frac{n + 1}{N + 1} \qquad \frac{n_2 + 1}{N + 1} < \frac{n + 1}{N + 1}$$

$$Avg_{o_1} < Avg_{o_k} \qquad Avg_{\bar{o}_1} < Avg_{o_k}$$

Por tanto, $P - SBCubed(D_1) < P - SBCubed(D_2)$.

De forma análoga a lo planteado al analizar el cumplimiento de la condición de homogeneidad por R-SBCubed, la condición de completitud no exige que todos los objetos de la clase C_k queden incluidos en G' y G'' o en G_k. Debido a esto, no es posible obtener el promedio individual para cada objeto de C_k. Sin embargo, los únicos pares que son clasificados de forma distinta que R-SBCubed para D_1 y D_2 son los que se encuentran en G' y G'', por lo cual nos centraremos en demostrar que estos pares son evaluados con una calificación más baja para D_1 que para D_2.

Haciendo un análisis análogo al hecho para P-SBCubed, para cada par (e_i, e'), donde o_i y o' pertenecen a C_k y a los grupos G' y G'' respectivamente

$$P_1(o_i, o') = \frac{n_1 + 1}{N + 1} \qquad \bar{P}_1(o_i, o') = \frac{n_1 + 1}{N + 1}$$

Al evaluar D_2, para los pares (o_i, o'), donde ambos objetos pertenecen a G_k, se cumple que $|E(o_i, o', D_2) \cap E(o_i, o', D_I)| = n$ y $|E(o_i, o', D_2) \cup E(o_i, o', D_I)| = N$. De esta forma, para todo par (o_i, o'), donde $o_i \in C_k \wedge o' \in C_k$

$$P_2(o_i, o') = \frac{n + 1}{N + 1}$$

Como $n_1 < n$, entonces $P_1(o_i, o') < P_2(o_i, o')$ y $\bar{P}_1(o_i, o') < P_2(o_i, o')$.

Por tanto, $R - SBCubed(D_1) < R - SBCubed(D_2)$.

2.2.3. Saco de Ruido

Como se explicó en el Capítulo 1, según esta condición, la calidad de un agrupamiento en el cual se introduce un objeto incorrectamente en un grupo que sería homogéneo sin ese objeto es inferior a la de uno que introduce un objeto incorrectamente en un grupo ruidoso.

Según la condición, se tiene un agrupamiento inicial con dos grupos G_1 y G_2, en el cual G_1 está compuesto por n objetos pertenecientes a una misma clase y G_2 por n objetos de clases unitarias, como se muestra en la Figura 2.6. A continuación demostraremos que si se incluye un objeto de una nueva clase unitaria en el grupo homogéneo (Figura 2.7), el agrupamiento D_1 obtenido es más penalizado por P-SBCubed y R-SBCubed que el

agrupamiento D_2 que resulta si el nuevo objeto se hubiera incluido en el grupo ruidoso (Figura 2.8).

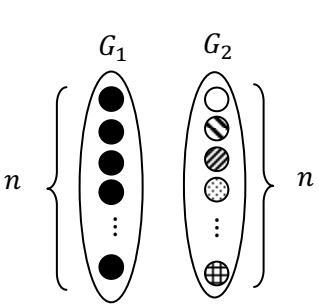

Fig 2.6. Condiciones iniciales

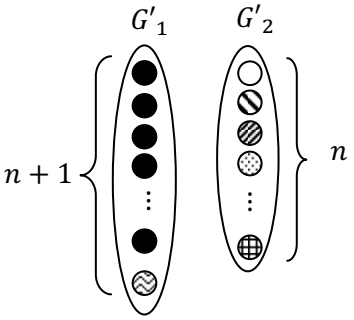

Fig 2.7. Agrupamiento $D1$

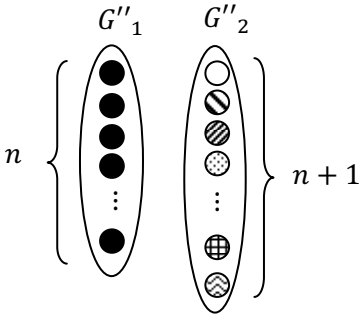

Fig 2.8. Agrupamiento $D2$

Para el cálculo de P-SBCubed, cuando se inserta el nuevo objeto en el grupo homogéneo sólo es afectado el promedio individual de los objetos que pertenecen a ese grupo. Como todos los objetos contenidos en G'_1 pertenecen a una misma clase C_1, entonces todo par $(o_i, o') \in G_1$ cumple también que $(o_i, o') \in C_1$. Por tanto, $|E(o_i, o', D_1) \cap E(o_i, o', D_I)| = n$ y $|E(o_i, o', D_1) \cup E(o_i, o', D_I)| = n + 1$. Así, todo o_i tal que $o_i \in G'_1 \wedge o_i \neq o_{n+1}$ cumple que

$$Avg_{o_i} = \frac{1}{n+1}\left(\sum_{1}^{n}\frac{n+1}{n+2}\right) = \frac{1}{n+1}\cdot\frac{n\,(n+1)}{n+2} = \frac{n}{n+2}$$

Para $o_{n+1} \in G'_1$, todo par (o_{n+1}, o_i) tal que $o_i \in G'_1 \wedge o_i \neq o_{n+1}$ contribuirá negativamente a la evaluación del agrupamiento ya que (o_{n+1}, o_i) no pertenece a ninguna clase. Sólo se tendrá en cuenta el par $(o_{n+1},\ o_{n+1})$. Para este caso

$$Avg_{o_{n+1}} = \frac{2}{(n+1)(n+2)}$$

En G'_2 todos los objetos pertenecen a clases unitarias. Por tanto, ningún par que integre alguno de estos objetos pertenece a alguna clase, excepto el par (o_i, o_i). Así, se tiene que $|E(o_i, o_i, D_1) \cap E(o_i, o_i, D_I)| = 1$, valor que se corresponde con el objeto o_i y $|E(o_i, o', D_1) \cup E(o_i, o', D_I)| = n + 1$.

De esta forma, todo $o_i \in G'_2$ cumple que

$$Avg_{o_i} = \frac{1}{n}\left(\frac{2}{n+1}\right) = \frac{2}{n(n+1)}$$

El promedio global parar el agrupamiento D_1 es

$$Avg_{D_1} = \frac{1}{2n+1}\left(n \cdot \frac{n}{n+2} + \frac{2}{(n+1)(n+2)} + n \cdot \frac{2}{n(n+1)}\right)$$

$$Avg_{D_1} = \frac{1}{2n+1}\left(\frac{n^2}{n+2} + \frac{2}{(n+1)(n+2)} + \frac{2}{(n+1)}\right)$$

$$Avg_{D_1} = \frac{1}{2n+1}\left(\frac{n^2(n+1) + 2(n+2) + 2}{(n+1)(n+2)}\right)$$

$$Avg_{D_1} = \frac{n^3 + n^2 + 2n + 6}{(n+1)(n+2)(2n+1)}$$

En el agrupamiento D_2, el grupo homogéneo G''_1 es idéntico a la clase C_1. Por tanto, para todo $o_i \in G''_1$, $Avg_{o_i} = 1$. Los objetos del grupo G''_1 pertenecen todos a clases unitarias y ningún par, excepto (o_i, o_i), aporta a la evaluación del agrupamiento.

Para este caso, donde $o_j \in G''_2$,

$$Avg_{o_j} = \frac{2}{(n+1)(n+2)}$$

Según lo cual el promedio global sería

$$Avg_{D_2} = \frac{1}{2n+1}\left(n + \frac{2(n+1)}{(n+1)(n+2)}\right)$$

$$Avg_{D_2} = \frac{n^2 + 2n + 2}{(n+2)(2n+1)}$$

Comparando las fracciones mediante el producto cruzado se verifica que

$$2n^5 + 7n^4 + 11n^3 + 24n^2 + 34n + 12 \; < 2n^5 + 11n^4 + 25n^3 + 30n^2 + 18n + 4$$

Es decir, $Avg_{D_1} < Avg_{D_2}$

Por tanto, $P - SBCubed(D_1) < P - SBCubed(D_2)$.

En el cálculo de R-SBCubed, para el agrupamiento D_1 tenemos que todo $o_i \in C_1$ cumple que

$$Avg_{o_i} = \frac{1}{n}\left(\sum_{1}^{n}\frac{n+1}{n+2}\right) = \frac{1}{n}\cdot\frac{n\,(n+1)}{n+2} = \frac{n+1}{n+2}$$

En este caso, como $o_{n+1} \notin C_1$, no se tiene en cuenta el par $(o_i,\ o_{n+1})$.

Las clases $C_i,\ 1 < i \leq n+1$, son unitarias. Por tanto, en ellas sólo se considera el par $(o_k,\ o_k)$, $o_k \in C_i$, calculado anteriormente. Finalmente se obtiene

$$Avg'_{D_1} = \frac{1}{2n+1}\left(n\cdot\frac{(n+1)}{n+2} + \frac{2}{n+2} + n\cdot\frac{2}{n+1}\right)$$

$$Avg'_{D_1} = \frac{n(n+1)^2 + 2(n+1) + 2n(n+2)}{(2n+1)(n+1)(n+2)}$$

$$Avg'_{D_1} = \frac{n^3 + 4n^2 + 7n + 2}{(2n+1)(n+1)(n+2)}$$

Al calcular R-SBCubed sobre el agrupamiento D_2, los objetos del grupo G''_1 se evalúan con su valor máximo por ser éste idéntico a C_1. Por tanto, $Avg_{o_i} = 1$ cuando $o_i \in C_1$. De igual forma que en los casos anteriores, para los objetos que pertenecen a las clases unitarias,

$$Avg_{o_i} = \frac{2}{(n+2)}$$

A raíz de lo cual

$$Avg'_{D_2} = \frac{1}{2n+1}\left(n + \frac{2(n+1)}{(n+2)}\right)$$

$$Avg'_{D_2} = \frac{1}{2n+1}\left(\frac{n(n+2) + 2(n+1)}{(n+2)}\right)$$

$$Avg'_{D_2} = \frac{n^2 + 4n + 2}{(n+2)(2n+1)}$$

Comparando las fracciones mediante el producto cruzado se verifica que

$$2n^5 + 13n^4 + 36n^3 + 47n^2 + 24n + 4 \ < 2n^5 + 15n^4 + 39n^3 + 44n^2 + 22n + 4$$

Es decir,

$$Avg'_{D_1} < Avg'_{D_2}$$

Por tanto, $R - SBCubed(D_1) < R - SBCubed(D_2)$.

2.2.4. Tamaño contra calidad

Como se explicó en el Capítulo 1, según esta condición, un error pequeño en un grupo grande es preferible que varios errores en grupos pequeños. Sea D un agrupamiento con un grupo G_k de tamaño $n+1$, donde todos sus objetos pertenecen a la clase C_k, del mismo tamaño, y además n grupos donde cada uno de ellos contiene dos objetos de la misma clase. Sea D_1 un agrupamiento idéntico a D, excepto en que los grupos de tamaño 2 son divididos en grupos unitarios y D_2 un agrupamiento idéntico a D excepto en que el grupo G_k es dividido en dos grupos, G'_k de tamaño n y G''_k de tamaño 1.

Se demostrará que, bajo estas condiciones, tanto P-SBCubed como R-SBCubed penalizan más a D_1 que a D_2.

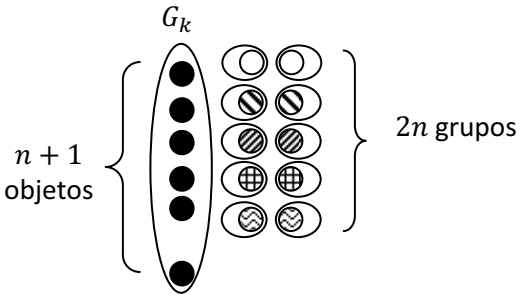

Fig 2.9. Agrupamiento D_1

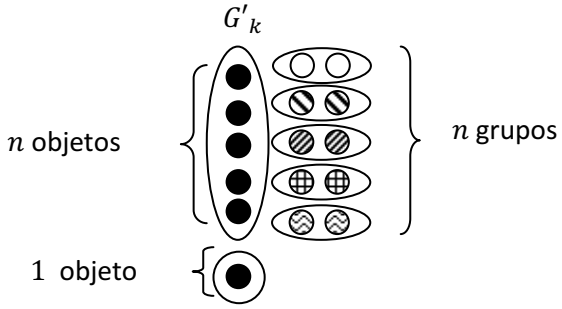

Fig 2.10. Agrupamiento D_2

El grupo G_k de D_1 es idéntico a la clase C_k. Por este motivo, para el cálculo de la precisión P-SBCubed cada objeto $o_1 \in G_k$ obtiene una puntuación máxima. Por otro lado, para cada objeto o_i que pertenece a las clases unitarias, $|E(o_i, o_i, D_1) \cap E(o_i, o_i, D_I)| = 1$ y $|E(o_i, o_i, D_1) \cup E(o_i, o_i, D_I)| = 2$.

Así, para D_1

$$Avg_{D_1} = \frac{1}{3n+1}\left((n+1) + 2n \cdot \left(\frac{2}{3}\right)\right)$$

$$Avg_{D_1} = \frac{7n+3}{9n+3}$$

En el agrupamiento D_2, para la evaluación de los objetos $o_j \in G'_k$ se tendrá en cuenta el objeto o_u que fue separado hacia un grupo unitario. Tenemos entonces que

$$Avg_{o_j} = \frac{1}{n}\left(\sum_{1}^{n} \frac{1+n}{1+(n+1)}\right) = \frac{1}{n}\left(n \cdot \frac{n+1}{n+2}\right) = \frac{n+1}{n+2}$$

$$Avg_{o_u} = \frac{2}{n+2}$$

Los otros grupos contienen dos objetos y, a su vez, para cada uno de estos grupos existe una clase idéntica. Por ello, el promedio interno para estos objetos será el máximo. De esta forma, tenemos que

$$Avg_{D_2} = \frac{1}{3n+1}\left(n \cdot \frac{(n+1)}{n+2} + \frac{2}{n+2} + 2n\right)$$

$$Avg_{D_2} = \frac{1}{3n+1}\left(\frac{n(n+1)+2+2n(n+2)}{n+2}\right)$$

$$Avg_{D_2} = \frac{3n^2+5n+2}{(3n+1)(n+2)}$$

Aplicando el producto cruzado entre los valores finales para ambos agrupamientos obtenemos

$$21n^3 + 44n^2 + 29n + 6 \ < \ 27n^3 + 72n^2 + 39n + 6$$

Es decir,

$$Avg_{D_1} < Avg_{D_2}$$

Por tanto, $P - SBCubed(D_1) < P - SBCubed(D_2)$.

Para el cálculo de R-SBCubed, como el grupo G_k del agrupamiento D_1 es idéntico a la clase C_k, entonces a todos los pares medidos sobre esta clase se les asignará el máximo valor. En el caso de las clases que contienen dos objetos, el promedio individual de cada objeto será favorecido al medir contra él mismo y desfavorecido al medir contra el otro objeto ya que ambos habrán sido ubicados en diferentes grupos unitarios.

Para cada objeto $o_i \in C_i$, con $C_i \neq C_k$

$$Avg_{o_i} = \frac{1}{2}\left(\frac{2}{3}\right) = \frac{2}{6}$$

De esta forma, para D_1

$$Avg_{D_1} = \frac{1}{3n+1}\left((n+1) + 2n \cdot \left(\frac{2}{6}\right)\right)$$

$$Avg_{D_1} = \frac{1}{3n+1}\left(\frac{10n+6}{6}\right)$$

$$Avg_{D_1} = \frac{10n+6}{18n+6}$$

Al evaluar los objetos que pertenecen a C_k, todos los pares que incluyan al objeto aislado o_u que se ubicó en el grupo unitario influirán negativamente. En los otros casos, para cada par (o_k, o_j), donde $o_k \in C_k$ y $o_j \in C_k$, $\left|E(o_k, o_j, D_2) \cap E(o_k, o_j, D_I)\right| = n$ y $\left|E(o_k, o_j, D_2) \cup E(o_k, o_j, D_I)\right| = n+1$, correspondiendo al tamaño de la clase.

De esta forma, para cada $\bar{o}_k \in C_k$, donde $\bar{o}_k \neq o_u$

$$Avg_{\bar{o}_1} = \frac{1}{n+1}\left(n \cdot \frac{n+1}{n+2}\right) = \frac{n}{n+2}$$

Todos los pares que contengan al objeto aislado $o_u \in C_k$, excepto el formado con él mismo, recibirán la mínima puntuación. Por otro lado, el par (o_u, o_u) cumple que $|E(o_u, o_u, G) \cap E(o_u, o_u, C)| = 1$ y $\left|E(o_k, o_j, G) \cup E(o_k, o_j, C)\right| = n+2$. Luego,

$$Avg_{o_u} = \frac{1}{n+1}\left(\frac{2}{n+2}\right) = \frac{2}{(n+1)(n+2)}$$

Dado que en D_2 las clases formadas por dos objetos y los grupos de igual tamaño son idénticos, entonces para los objetos pertenecientes a estas clases se obtiene la máxima puntuación.

Finalmente, para D_2

$$Avg_{D_2} = \frac{1}{3n+1}\left(\frac{n^2}{n+2} + \frac{2}{(n+1)(n+2)} + 2n\right)$$

$$Avg_{D_2} = \frac{3n^3 + 7n^2 + 4n + 2}{(3n+1)(n+1)(n+2)}$$

Aplicando el producto cruzado entre los valores finales para cada agrupamiento obtenemos que

$$30n^4 + 118n^3 + 150n^2 + 74n + 12 \; < 54n^4 + 144n^3 + 144n^2 + 60n + 12$$

Es decir,

$$Avg_{D_1} < Avg_{D_2}$$

Por tanto, $R - SBCubed(D_1) < R - SBCubed(D_2)$.

A continuación se analiza cómo se comporta $F_\alpha\, SBCubed$ cuando ocurre un incremento en la P-SBCubed o R-SBCubed.

Sustituyendo $\propto = \frac{1}{E}$ con $E > 0$, en $F_\alpha\, SBCubed(R, P)$ obtenemos

$$F_\alpha\, SBCubed(R, P) \; = \frac{1}{\frac{1}{E \cdot P} + \frac{E-1}{E \cdot R}}$$

Si P-SBCubed aumenta en $\varepsilon_p > 0$ o R-SBCubed aumenta en $\varepsilon_r > 0$, para los valores de $F_\alpha\, SBCubed\big(R, P + \varepsilon_p\big)$ o $F_\alpha\, SBCubed(R + \varepsilon_r, P)$ se cumpliría, respectivamente, lo siguiente:

$$F\big(R, P + \varepsilon_p\big) = \frac{1}{\frac{1}{E \cdot \big(P + \varepsilon_p\big)} + \frac{E-1}{E \cdot R}}$$

Como $\frac{1}{E \cdot (P + \varepsilon_p)} < \frac{1}{E \cdot (P)}$ entonces

$$F\big(R, P + \varepsilon_p\big) > F(R, P)$$

$$F(R + \varepsilon_r, P) = \frac{1}{\frac{1}{E \cdot P} + \frac{E-1}{E \cdot (R + \varepsilon_r)}}$$

Como $\frac{E-1}{E \cdot (R + \varepsilon_r)} < \frac{E-1}{E \cdot R}$ entonces

$$F(R + \varepsilon_r, P) > F(R, P)$$

De igual forma, si aumentan simultáneamente ambos valores, entonces

$$F\big(R + \varepsilon_r, P + \varepsilon_p\big) = \frac{1}{\frac{1}{E \cdot \big(P + \varepsilon_p\big)} + \frac{E-1}{E \cdot (R + \varepsilon_r)}}$$

Como $\frac{1}{E \cdot (P + \varepsilon_p)} < \frac{1}{E \cdot (P)}$ y $\frac{E-1}{E \cdot (R + \varepsilon_r)} < \frac{E-1}{E \cdot R}$ entonces

$$F\left(R+\varepsilon_r, P+\varepsilon_p\right) > F(R,P)$$

Según lo anterior, si al comparar un agrupamiento D_2 con un agrupamiento D_1 aumenta la medida P-SBCubed manteniéndose R-SBCubed, aumenta R-SBCubed manteniéndose P-SBCubed, o ambas aumentan simultáneamente, también aumentará F_α SBCubed.

Al evaluar el cumplimiento de las cuatro condiciones de Amigó et al. para P-SBCubed y R-SBCubed, se observó que bajo las suposiciones en las cuales se basan dichas condiciones, en ambos casos la medida analizada prefirió a D_2 sobre D_1. Siguiendo el razonamiento anterior, en cada uno de esos casos F_α SBCubed también preferirá a D_2 sobre D_1, por lo cual puede afirmarse que F_α SBCubed también cumple con las cuatro condiciones.

2.3. Análisis del cumplimiento de la condición Obtención de óptimo sólo para emparejamiento perfecto por las medidas SBCubed

Se considera que una medida de evaluación Q satisface esta condición si

$$D_I = D \iff Q(D_I, D) = 1$$

Anteriormente se vio que F_α EBCubed no cumple con esta condición para todos los casos. Primeramente, debemos señalar que las medidas SBCubed tampoco cumplen con la condición en todos los casos, como puede observarse en el contraejemplo de la Figura 2.11.

$C_1 = 1,2,3,4$	$G_1 = 1,2,3$
$C_2 = 1,2$	$G_2 = 1,2,4$
$C_3 = 2,3$	$G_3 = 1,3,4$
$C_4 = 1,3$	$G_4 = 2,3,4$
$C_5 = 1,4$	$G_5 = 1$
$C_6 = 2,4$	$G_6 = 2$
$C_7 = 3,4$	$G_7 = 3$
	$G_8 = 4$
	$G_6 = 2$

Fig 2.11. Ejemplo en el cual F_α SBCubed no cumple la nueva condición

En este caso, para todos los pares se cumple que $E(o_k, o_j, D) = E(o_k, o_j, D_I)$ y $|G(o) \cap G(o')| = |C(o) \cap C(o')|$, por lo que F_α SBCubed da como resultado la máxima puntuación al comparar ambos agrupamientos. Sin embargo, como puede observarse, éstos no lo son.

A pesar de esto, sí puede comprobarse que el conjunto de casos en que F_α SBCubed cumplen con el postulado de la condición es un supraconjunto del conjunto de casos en que F_α SBCubed lo hace, por lo cual ésta representa una mejor solución parcial al problema.

A continuación demostraremos que en los casos que F_α EBCubed cumple el postulado de la condición F_α SBCubed también lo cumple y que existen casos en que F_α EBCubed no cumple el postulado y F_α SBCubed sí. En lo adelante utilizaremos la notación Q_1 para referirnos a F_α EBCubed y con Q_2 para referirnos a F_α SBCubed.

2.3.1. Cumplimiento del postulado de la condición por la medida $\mathbf{F_\alpha}$ SBCubed para los casos en que $\mathbf{F_\alpha}$ EBCubed lo cumple

Estos casos acontecen cuando EBCubed evalúa a D con máxima puntuación y además se cumple que $D_I = D$. A continuación se demuestra que para $D_I = D$ la medida F_α SBCubed evalúa a D con máxima puntuación.

En el cálculo de P-SBCubed y R-SBCubed, para todo par (o_1, o_2) donde o_1 y o_2 son objetos pertenecientes a un mismo grupo o clase, respectivamente

$$P(o_1, o_2) = \frac{\min(|G(o_1) \cap G(o_2)|, |C(o_1) \cap C(o_2)|) + |E(o_1, o_2, D) \cap E(o_1, o_2, D_I)|}{|G(o_1) \cap G(o_2)| + |E(o_1, o_2, D) \cup E(o_1, o_2, D_I)|}$$

para el caso de P-SBCubed y

$$P(o_1, o_2) = \frac{\min(|G(o_1) \cap G(o_2)|, |C(o_1) \cap C(o_2)|) + |E(e_1, e_2, D) \cap E(e_1, e_2, D_I)|}{|C(o_1) \cap C(o_2)| + |E(o_1, o_2, D) \cup E(o_1, o_2, D_I)|}$$

para el caso de R-SBCubed.

Dado que $D_I = D$, $|G(o_1) \cap G(o_2)| = |C(o_1) \cap C(o_2)$ se puede afirmar tanto para la medida P-SBCubed como para R-SBCubed que

$$P(o_1, o_2) = \frac{\min(|G(o_1) \cap G(o_2)|, |G(o_1) \cap G(o_2)|) + |E(o_1, o_2, D) \cap E(o_1, o_2, D)|}{|G(o_1) \cap G(o_2)| + |E(o_1, o_2, D) \cup E(o_1, o_2, D)|}$$

$$P(o_1, o_2) = \frac{|G(o_1) \cap G(o_2)| + |E(o_1, o_2, D)|}{|G(o_1) \cap G(o_2)| + |E(o_1, o_2, D)|}$$

$$P(o_1, o_2) = 1$$

Así, todos los pares son evaluados por las medidas P-SBCubed y R-SBCubed con puntuación máxima. Como estas medidas evalúan a los agrupamientos a través de promedios sobre la puntuación de estos pares entonces estos agrupamientos, a su vez, obtendrán máxima puntuación. Por tanto,

$$D_I = D \implies Q_2(D_I, D) = 1$$

Por otro lado, cuando $D_I \neq D$ y $F_\alpha\,\text{SBCubed} < 1$, existe al menos un par de objetos (o_i, o_j) que pertenecen a una misma clase o a un mismo grupo y cumplen que $|G(o_i) \cap G(o_j)| \neq |C(o_i) \cap C(o_j)|$, por lo cual P-SBCubed < 1, R-SBCubed < 1 y, consecuentemente, F_α SBCubed < 1.

Por tanto,

$$[D_I \neq D \wedge F_\alpha\,\text{SBCubed} < 1] \implies Q_2(D_I, D) < 1$$

2.3.2. Tipificación de los casos en que F_α EBCubed no cumple el postulado de la nueva condición y F_α SBCubed lo cumple

Sean un agrupamiento ideal $\widehat{D_I}$ y un agrupamiento a evaluar $\widehat{D}$, tales que cualquier par de objetos que se tome de un grupo de G o una clase de C pertenezca a la misma cantidad de grupos que de clases. Supongamos que existe al menos un par (o_1, o_2) tal que la unión de todos los grupos que lo contienen es distinta de la unión de todas las clases que lo contienen. Formalmente

$$(1) \quad \forall (o_1, o_2), |C(o_1) \cap C(o_2)| = |G(o_1) \cap G(o_2)|$$

y

$$(2) \quad \exists (o_1, o_2)\ E(o_1, o_2, \widehat{D_I}) \neq E(o_1, o_2, \widehat{D})$$

Para evaluar $\widehat{D}$ las medidas EBCubed tienen en cuenta la cantidad de clases y grupos a las que pertenecen los pares. Por tanto, tanto para precisión EBCubed como para relevancia EBCubed se obtienen máximos, o sea

$$P_{prec}(e_1, e_2) = \frac{\min(|G(e_1) \cap G(e_2)|, |C(e_1) \cap C(e_2)|)}{|G(e_1) \cap G(e_2)|} = \frac{|G(e_1) \cap G(e_2)|}{|G(e_1) \cap G(e_2)|} = 1$$

$$P_{rel}(e_1, e_2) = \frac{\min(|G(e_1) \cap G(e_2)|, |C(e_1) \cap C(e_2)|)}{|C(e_1) \cap C(e_2)|} = \frac{|C(e_1) \cap C(e_2)|}{|C(e_1) \cap C(e_2)|} = 1$$

Al promediar, tanto precisión EBCubed como relevancia EBCubed tomarán valor 1, aún cuando el agrupamiento no es perfecto. Sin embargo, precisión SBCubed y relevancia SBCubed, además de tener en cuenta la cantidad de grupos y clases las que pertenece un par, también analiza la unión de estos grupos y clases y, por tanto, penaliza a $\widehat{D}$ en estos casos.

$$P_{prec}(o_1, o_2) = \frac{\min(|G(o_1) \cap G(o_2)|, |C(o_1) \cap C(o_2)|) + |E(o_1, o_2, \widehat{D_I}) \cap E(o_1, o_2, \widehat{D})|}{|G(o_1) \cap G(o_2)| + |E(o_1, o_2, \widehat{D_I}) \cup E(o_1, o_2, \widehat{D})|}$$

$$= \frac{|G(o_1) \cap G(o_2)| + |E(o_1, o_2, \widehat{D_I}) \cap E(o_1, o_2, \widehat{D})|}{|G(o_1) \cap G(o_2)| + |E(o_1, o_2, \widehat{D_I}) \cup E(o_1, o_2, \widehat{D})|}$$

$$P_{rel}(o_1, o_2) = \frac{\min(|G(o_1) \cap G(o_2)|, |C(o_1) \cap C(o_2)|) + |E(o_1, o_2, \widehat{D_I}) \cap E(o_1, o_2, \widehat{D})|}{|C(o_1) \cap C(o_2)| + |E(o_1, o_2, \widehat{D_I}) \cup E(o_1, o_2, \widehat{D})|}$$

$$= \frac{|C(o_1) \cap C(o_2)| + |E(o_1, o_2, \widehat{D_I}) \cap E(o_1, o_2, \widehat{D})|}{|C(o_1) \cap C(o_2)| + |E(o_1, o_2, \widehat{D_I}) \cup E(o_1, o_2, \widehat{D})|}$$

en ambos casos, como $E(o_1, o_2, \widehat{D_I}) \neq E(o_1, o_2, \widehat{D})$, entonces

$$|E(o_1, o_2, \widehat{D_I}) \cap E(o_1, o_2, \widehat{D})| \leq |E(o_1, o_2, \widehat{D_I}) \cup E(o_1, o_2, \widehat{D})|$$

De esta forma, P-SBCubed < 1, R-SBCubed < 1 y, consecuentemente, F_α SBCubed < 1, por lo cual puede afirmarse que F_α SBCubed cumple el postulado de la condición.

Según lo analizado, el conjunto de casos en que las medidas SBCubed cumplen el enunciado de la condición Obtención de óptimo sólo para emparejamiento perfecto es un supraconjunto del conjunto de casos en que lo hace F_α EBCubed, compuesto por la unión de todos los casos para los cuales F_α EBCubed cumple el postulado y los agrupamientos que satisfacen las restricciones (1) y (2) anteriormente enunciadas.

Conclusiones

En este trabajo se ha abordado el problema de la evaluación de los algoritmos de agrupamiento. Primeramente, se realizó un exhaustivo estudio del tema, centrado en las medidas de evaluación existentes, así como las condiciones propuestas para determinar la validez de dichas medidas de evaluación.

Como resultado de este análisis, se identificó uno de los problemas no resueltos en este tipo de medidas cuando se aplican a algoritmos solapados. La identificación de este problema permitió enunciar una nueva condición para determinar la validez de las medidas de evaluación.

La medida F_α EBCubed, la cual, según se reporta en la literatura, cumple todas las condiciones impuestas hasta el momento, no cumple la nueva condición en todos los casos. A raíz de esto, se proponen las nuevas medidas precisión SBCubed, relevancia SBCubed y F_α SBCubed, basadas en F_α EBCubed, que resuelven parcialmente el problema detectado. Si bien la nueva medida F_α SBCubed tampoco cumple la nueva condición en el 100% de los casos, se demuestra que el conjunto de casos en que la cumple es un supraconjunto del conjunto de casos en que la cumple F_α EBCubed. Adicionalmente, se demuestra que las nuevas medidas también cumplen el resto de las condiciones anteriormente enunciadas. La obtención de estos resultados confirma la hipótesis inicial de la que se partió en nuestro trabajo.

Los resultados obtenidos en este trabajo permitirán aumentar el rigor con que se evalúan los algoritmos de agrupamiento, con lo cual se beneficiarán tanto los investigadores que trabajan en este campo como los desarrolladores de software y empresas que necesitan evaluar diferentes algoritmos para resolver problemas reales.

Como trabajo futuro, nos proponemos modificar las medidas propuestas a fin de resolver los problemas que hacen que aún no cumpla la nueva condición en todos los casos.

Referencias Bibliográficas

Amigó, Enrique; Gonzalo, Julio; Javier, Artiles; and Verdejo, Felisa (2008). *"A comparison of extrinsic clustering evaluation metrics based on formal constraints".* Journal of Information Retrieval.

Bagga, A.; Baldwin, B. (1998). *"Entity-based cross-document coreferencing using the vector space model"*. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98).

Dom, B. (2001). *"An information-theoretic external cluster-validity measure".* IBM Research Report.

Fowlkes, E.; Mallows, C (1983). *"A method for comparing two hierarchical clustering".* Journal of the American Statistical Association.

Gil, Reynaldo (2005). *"Algoritmos de agrupamiento sobre grafos y su paralelización"*. Tesis Doctoral. Universidad Jaume I.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). *"On clustering validation techniques".* Journal of Intelligent Information Systems.

Meila, M. (2003). *"Comparing clusterings by the Variation Information".* In Proceedings of COLT 03. Washington, DC.

Pantel, P.; Lin, D. (2002). *"Efficiently clustering documents with committees"*. In Proceedings of the PRICAI 2002 7th Pacific Rim International Conference on Artificial Intelligence.

Pons, A (2004), Desarrollo de algoritmos para la estructuración dinámica de información y su aplicación a la detección de sucesos. Tesis Doctoral. Universidad Jaume I.

Rosenberg, A.; Hirschberg, J. (2007). *"V-measure: A conditional entropy-based external cluster evaluation measure".* In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).

Shannon. C. (1948) *"A Mathematical Theory of Communication".* Bell System Technical Journal.

Steinbach, Michael; Karypis, George; Kumar, Vipin (2000). "*A Comparison of Document Clustering Techniques*". University of Minnesota.

Xu, W.; Liu, X.; Gong, Y. (2003). *"Document clustering based on non-negative matrix factorization".* Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.