Ediciones pensando el futuro
Editions thinking the future
Informática 2013
XV CONVENCIÓN Y FERIA INTERNACIONAL
15TH INTERNATIONAL CONVENTION AND FAIR
www.informaticahabana.com
www.informaticahabana.cu
ISBN: 978-959-7213-02-4

PROPUESTA DE UNA NUEVA MEDIDA DE CALIDAD PARA ALGORITMOS DE AGRUPAMIENTO SOLAPADOS

PROPOSAL OF A NEW QUALITY MEASURE FOR OVERLAPPING CLUSTERING ALGORITHMS

Henry Rosales-Méndez[1], Yunior Ramírez-Cruz[2], Reynaldo Gil-García[2]

1 Universidad de Oriente, Cuba, henry@csd.uo.edu.cu, Patricio Lumumba S/N Santiago de Cuba.

2 Centro de Estudios de Reconocimientos de Patrones y Minería de Datos, Cuba, {yunior, gil}@cerpamid.co.cu

RESUMEN: *La evaluación de los algoritmos de agrupamiento es un problema abierto en el campo del Reconocimiento de Patrones. La mayoría de las medidas de calidad que se encuentran en la literatura no son aplicables en evaluaciones donde intervengan agrupamientos no solapados, cuando existen problemas reales que se pueden representar en mayor grado por agrupaciones solapadas. Las medidas de calidad, a su vez, son evaluadas por condiciones que miden su robustez y eficacia. En este trabajo se amplía el principal conjunto de condiciones que evalúan a las medidas de calidad externas, con el objetivo de favorecer a las medidas más selectivas. Además proponemos una nueva medida de evaluación para los algoritmos solapados que, aunque no satisface la condición propuesta, se desempeña correctamente para un mayor conjunto de casos que el desempeño obtenido por las medidas existentes.*

Palabras Clave: Medidas de calidad; Evaluación de agrupamientos solapamientos.

ABSTRACT: *The evaluation of clustering algorithms is an open problem in the field of Pattern Recognition. Most evaluation measures that are found in the literature are not applicable where assessments involving non-overlapping clustering, when real problems represent mainly overlapping clustering. The quality measures, in turn, are evaluated by conditions that measuring their robustness and efficiency. This paper extends the main set of conditions that evaluate external quality measures, with the aim of favoring more selective measures. We also propose a new measure for evaluation the overlapped algorithms, and although it does not satisfy the proposed, performs well for a larger set of cases that the performance by the existing measures.*

KeyWords: Quality measure; Evaluation overlapping clustering.

1. INTRODUCCIÓN

Los algoritmos de agrupamiento son los encargados de dividir a una colección de objetos en grupos, los cuales se asume que representan su estructura interna. Evaluar los algoritmos de agrupamiento es una tarea abierta y necesaria. Las medidas que evalúan estos agrupamientos permiten obtener un indicador cuantitativo de la efectividad del agrupamiento.

Actualmente, estas medidas se clasifican en externas o internas [1]. Las medidas internas permiten evaluar los agrupamientos sin ningún conocimiento externo, en cambio, las medidas externas comparan los grupos obtenidos (agrupamiento candidato) contra los grupos conocidos (agrupamiento ideal). Los agrupamientos ideales son construidos por expertos teniendo en cuenta las características de la información que se agrupa, por tal motivo, las mayoría de las investigaciones realizan sus experimentos sobre colecciones que contengan agrupamientos ideales. Por esta razón, este trabajo se centró en la evaluación externa.

Los algoritmos de agrupamientos se pueden dividir

en solapados o no solapados. Los agrupamientos no solapados son los que colocan cada objeto en un solo grupo, en cambio, los agrupamientos solapados permiten que un objeto pueda pertenecer a más de un grupo a la vez. Existen problemas reales que se pueden representar en mayor grado por agrupaciones solapadas. Sin embargo, las mayorías de las medidas de calidad no son aplicables en evaluaciones donde intervengan agrupamientos solapados.

Debido a la gran cantidad de medidas de evaluación, varios autores han trabajado para determinar las características de una buena medida de calidad externa. La literatura registra varios conjuntos de condiciones que deben cumplir las medidas de calidad externas. En este trabajo partimos del conjunto de condiciones propuesto por Amigo et al.,[2] el cual enriquecemos añadiendo una nueva condición, que si bien resulta trivial cuando se evalúan agrupamientos no solapados, en evaluaciones solapadas resulta un reto.

Amigo et al., proponen además, en su trabajo, una familia de medidas llamada *BCubed Extendo*, y plantean que estas medidas resultan las más adecuadas para evaluar agrupamientos solapados. Sin embargo, la familia *BCubed Extendido* puede obtener una puntuación máxima para un agrupamiento candidato aún cuando éste no sea idéntico al agrupamiento ideal. Atacando este problema en este trabajo se propone una nueva familia de medidas de calidad aplicables en evaluaciones que intervengan agrupamientos solapados, donde una de ellas se desempeña correctamente para una mayor cantidad de casos que el conjunto de medidas *BCubed Extendido*.

El resto del artículo se estructura como sigue: en la Sección 2 mencionamos las principales medidas registradas en la literatura y presentamos nuestras propuestas. En la Sección 3 ampliamos el principal conjunto de condiciones para evaluar las medidas de calidad externas. Y en la Sección 4 exponemos nuestras conclusiones.

2. EVALUCIÓN DE LOS ALGORITMOS DE AGRUPAMIENTO

Con el desarrollo de los algoritmos de agrupamiento han surgido nuevas medidas de calidad y desde diferentes enfoques. Estas medidas se pueden agrupar teniendo en cuenta el análisis que realizan.

2.1 Medidas basadas en emparejamiento

Las medidas basadas en emparejamiento tienen como principal característica que evalúan a los agrupamientos candidatos comparando par a par los grupos y las clases[1]. Aquí podemos encontrar a las medidas H, L y D propuestas por Meila [3], donde $match(k)$ es la función que devuelve para cada grupo G_k su mejor emparejamiento entre las clases, y además, $n_{kt} = |G_k \cap C_t|$.

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

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

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

Otra medida que pertenece a esta familia, proveniente de la Recuperación de Información es la medida F_1 y 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.

$$P = \frac{|C_t \cap G_k|}{G_k}$$

$$R = \frac{|C_t \cap G_k|}{C_t}$$

$$F_1 = \frac{2 \cdot R(C_t, G_k) \cdot P(C_t, G_k)}{R(C_t, G_k) + P(C_t, G_k)}$$

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 $\sigma(C_t)$ el cual obtiene el máximo valor por la medida F_1, o sea

$$\sigma(C_t) = \arg\max_{k}\ F(C_t, G_k)$$

La medida F_1 *Macro-promediada* [4] 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.

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

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

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

[1] Trataremos como clases a los grupos del agrupamiento ideal.

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.

2.2 Medidas basadas en conteo de pares

Las medidas que 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.

Teniendo como base esta tabla se han propuesto varias medidas de evaluación [5]-[6], 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.

2.3 Medidas basadas en la entropía

La entropía [7] es utilizada para medir el grado de desorden de un grupo contra una clase. Siendo $p(t,k)$ la probabilidad de encontrar un objeto de la clase C_t en el grupo G_k, 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

$$H(D_I|D) = -\sum_{t}\sum_{k} p(t,k)\, log_2 p(k|t)$$

La medida *Información Mutua* [8] estima la cantidad de información en común entre los agrupamientos a comparar y se presenta a continuación

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

La medida *Variación de Información* [2] se basa en la cantidad de información ganada y perdida al cambiar una clase C por un grupo G y se expresa a través de

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

Las medidas expuestas anteriores de esta sección no tienen en cuenta la completitud de los grupos. La *Medida V* [9] 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

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

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

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

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 [10] propone la medida Q_0, la cual incluye un modelo de costo de términos que es sumado al valor de la entropía. Siendo $h(k) = \sum_c h(k,c)$ y $h(c) = \sum_k h(k,c)$ la tabla marginal correspondiente a la tabla de contingencia, entonces

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

Dom también presenta una versión normalizada

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

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 [11] propone los siguientes pasos para la transformación:

- Crear un conjunto vacío para cada clase.
- Agregar los objetos de un grupo al conjunto con cuya clase asociada tenga más objetos en común.
- Mover losobjetos ubicados incorrectamente alos 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.

2.5 Medidas Híbridas

Como su nombre lo indica, estas medidas combinan ideas de varias de las familias mencionadas anteriormente. La familia de medidas *BCubed* [12] 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.

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

La medida *F-BCubed* es similar a F_1 Medida, excepto que utiliza otra forma de calcula la precisión y la relevancia.

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

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

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

La medida *F-BCubed* es una medida de calidad efectiva sólo en evaluaciones de agrupamientos no solapados, en tareas solapadas esta medida no es aplicable. Amigó et al., proponen la familia de medidas *BCubed Extendido*, una extensión a las medidas *BCubed*, destinada a tareas solapadas. De forma tal que en la evaluación de agrupamientos no solapados se comporta idénticamente a las medidas *BCubed*.

$$P = \frac{1}{|D|}\sum_{o\in D}\frac{1}{|\cup_{g\in G(o)} g|}\sum_{o'\in E(o,G)}\frac{\min(|G(o,o')|, |G(o,o')|)}{|G(o,o')|}$$

$$R = \frac{1}{|D|}\sum_{o\in D}\frac{1}{|\cup_{c\in G(o)} c|}\sum_{o'\in E(o,C)}\frac{\min(|G(o,o')|, |G(o,o')|)}{|C(o,o')|}$$

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

donde

G: Agrupamiento a candidato.

C: Agrupamiento ideal.

$G(o)$: Conjunto de grupos que contienen a o.

$C(o)$: Conjunto de clases que contienen a o.

$G(o_1, o_2)$: Conjunto de grupos que contienen al par de objetos o_1 y o_2.

$G(o_1, o_2)$: Conjunto de grupos que contienen al par de objetos o_1 y o_2.

Mientras que en *BCubed* sólo tienen en cuenta si los pares de objetos pertenecen o no a un grupo o clase, en *BCubed Extendido* se analizan la cantidad de estos grupos y clases.

3. PROPUESTA DE UNA NUEVA MEDIDA DE EVALUACIÓN

Como mencionamos anteriormente, varios autores han propuesto conjuntos de condiciones para las medidas de calidad. Amigó et al., proponen cuatro condiciones que evalúan agrupamientos no solapados. Estas condiciones suponen dos agrupamientos D_1 y D_2 de tal forma que D_1 es un peor agrupamiento que D_2. Por tanto, las condiciones plantean que todas las medidas de evaluación deben penalizar más al agrupamiento D_1 que al agrupamiento D_2. Estas condiciones son mostradas a continuación:

- *Homogeneidad:* Siendo S un conjunto de objetos pertenecientes a las clases $C_1, C_2, \dots, C_n$. Siendo D_1 un agrupamiento con un grupo G_k de n objetos de dos clases: C_i con tamaño N_1 y C_j con tamaño N_2. Sea D_2 un agrupamiento idéntico a D_1, excepto que el grupo G_k es dividido en dos grupos, uno que contiene los n_1 objetos de G_k que pertenecen C_i y otros que contiene los n_2 objetos de G_k que pertenecen a C_j (Figura 1). Entonces toda medida de calidad debe dar una peor puntuación al agrupamiento D_1 sobre el agrupamiento D_2.

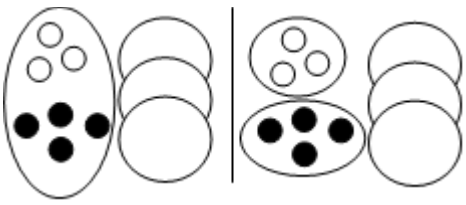

Figura. 1: Par de agrupamientos donde se ejemplifica la condición de homogeneidad.

- *Completitud:* Siendo D_1 un agrupamiento con dos grupos G_1 y G_2 de tamaños n_1 y n_2 respectivamente, que sólo contienen elementos pertenecientes a una misma clase C_k de tamaño N. Siendo D_2 un agrupamiento idéntico a D_1, excepto que G_1 y G_2 son mezclados en un mismo grupo G_k de tamaño n (Figura 2). Entonces toda medida de calidad debe dar una peor puntuación al agrupamiento D_1 sobre el agrupamiento D_2.

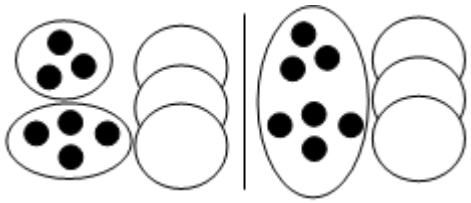

Figura. 2: Par de agrupamientos donde se ejemplifica la condición de completitud

- *Saco de ruido*: Siendo 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 D_1 un agrupamiento con los grupos G_{clean}y G_{noise} donde al grupo G_{clean} se le es mezclado un nuevo objeto de una nueva clase. Sea D_2 un agrupamiento con los grupos G_{clean}y G_{noise}, donde al grupo G_{noise} se le es mezclado un nuevo objeto de una nueva clase. Entonces toda medida de calidad debe dar una peor puntuación al agrupamiento D_1 sobre el agrupamiento D_2.

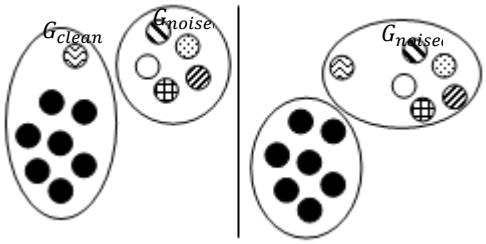

Figura. 3: Par de agrupamientos donde se ejemplifica la condición de Saco de Ruido

- *Tamaño contra calidad:* Si consideremos un agrupamiento D 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 $C_1, C_2, \dots, C_n$. Si D_1 es un nuevo agrupamiento similar a D donde cada G_i es dividido en dos grupos unitarios y D_2 es un agrupamiento similar a D donde G_l es dividido en un grupo de tamaño n y un grupo de tamaño 1 (Figura 4). Entonces toda medida de calidad debe dar una peor puntuación al agrupamiento D_1 sobre el agrupamiento D_2.

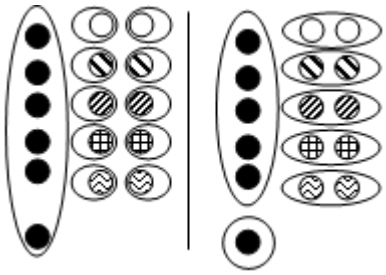

Figura. 4: Par de agrupamientos donde se ejemplifica la condición de tamaño contra calidad

Meila propone doce condiciones destinadas a evaluar la medida Variation Information. Dom propone cinco condiciones que basan su análisis en un criterio probabilístico a partir del cual se decide cuáles son los grupos acertados. Un trabajo posterior, de Rosenberg et al., extiende las condiciones propuestas por Dom a siete.

Amigó et al., analizaron las medidas de calidad principales existentes en la literatura y obtuvieron como resultado que estas medidas no cumple con todas las condiciones que propusieron. Esto evidencia que las condiciones propuestas por Amigó et al., cubren las presentadas en trabajos anteriores, y por esta razón son las que utilizamos en esta investigación. En [12] se evaluaron las principales medidas de calidad, y como resultado se obtuvo que las únicas medidas de calidad que cumplen con las cuatro condiciones son las familias de medias *BCubed* y *BCubed Extendido*.

3.1 Ampliación del conjunto de condiciones propuesto por Amigó et al.

Las medidas que agrupa *BCubed Extendido* basan su análisis en el conteo de pares. Por esta razón todos los agrupamientos candidatos donde ocurran los mismos pares que el agrupamiento ideal, aunque sea en diferentes grupos, son evaluados con máxima puntuación. Es decir, la familia *BCubed Extendido* puede evaluar con la máxima puntuación a un agrupamiento aún cuando éste no sea idéntico al ideal.

$$\mathrm{P-SBCubed} = \frac{1}{|D|}\sum_{o\in D}\frac{1}{|\cup_{g\in G(o)} g|}\sum_{o'\in E(o,G)}\frac{\min(|G(o,o')|,|C(o,o')|) + |E(o,o',G)\cap E(o,o',C)|}{|G(o,o')| + |E(o,o',G)\cup E(o,o',C)|} \quad (1)$$

$$\mathrm{R-SBCubed} = \frac{1}{|D|}\sum_{o\in D}\frac{1}{|\cup_{c\in G(o)} c|}\sum_{o'\in E(o,C)}\frac{\min(|G(o,o')|,|C(o,o')|) + |E(o,o',G)\cap E(o,o',C)|}{|C(o,o')| + |E(o,o',G)\cup E(o,o',C)|} \quad (2)$$

$$\mathrm{F-SBCubed} = \frac{1}{\propto\left(\frac{1}{P-\mathrm{SBCubed}}\right)+(1-\propto)\left(\frac{1}{R-\mathrm{SBCubed}}\right)} \quad (3)$$

Consideramos que este comportamiento no es adecuado y que, por tanto, las condiciones propuestas por Amigó et al., no son suficientes para medir el comportamiento de las medidas de calidad que evalúan agrupamientos solapados.

A continuación proponemos la condición de *Emparejamiento Perfecto*, que pretende precisamente requerir esta condición en las medidas.

Condición de Emparejamiento Perfecto: una medida de calidad debe calificar a un agrupamiento con puntuación máxima si y sólo si este agrupamiento es perfecto.

Esta condición, que en tareas no solapadas resulta trivial, para tareas solapadas se convierte en un reto. Teniendo en cuenta que la familia *BCubed Extendido* hasta la actualidad es el conjunto de medidas más completo registrado en la literatura que evalúa agrupamientos solapados, nos planteamos como objetivo proponer una nueva medida de calidad que manipule más adecuadamente un conjunto de situaciones de las medidas *BCubed Extendido*.

3.1.1 SBCubed

La familia de medidas que proponemos (1), (2) y (3) es una extensión de las medidas *BCubed Extendido*, que además de tener en cuenta la cantidad de grupos y clases en que se encuentran los pares, analiza también a los objetos que comparten un grupo con el par analizado. Es recomendable usar el parámetro $\propto = 0.5$ para que *F-SBCubed* se comporte como el promedio armónico entre la Precisión *SBCubed* y *Rrelevancia SBCubed*, en lo adelante *P-SBCubed* y *R-SBCubed* respectivamente.

Para un agrupamiento A, el término $E(o_1, o_2, A)$ representa el conjunto de objetos que comparten un grupo con el par o_1, o_2 en el agrupamiento A.

3.1.2 Validación de SBCubed

El conjunto de agrupamientos para los cuales el conjunto de medidas *SBCubed* cumple con la condición de *Emparejamiento Perfecto* es un supraconjunto del conjunto formado por los agrupamientos que cumplen la condición por las medidas *BCubed Extendido*. Si suponemos un agrupamiento $\hat{G}$ donde todos los pares (o_1, o_2) que pertenezcan a un mismo grupo o una misma clase cumplen que $|G(o_1)\cap G(o)|, |C(o_1)\cap C(o_2)|$ y además, existe al menos un par $(\overline{o_1}, \overline{o_2})$ tal que $E(\overline{o_1}, \overline{o_2}, G) \neq E(\overline{o_1}, \overline{o_2}, C)$, como es el caso de los agrupamientos representados en la figura 5 , entonces $\hat{G}$ es evaluado por *BCubed Extendido* con máxima puntuación, por el contrario, nuestra propuesta penaliza a $\hat{G}$ como es debido y no lo califica con máxima puntuación.

$C_1 = 1,2,4$ $\quad$ $G_1 = 1,3,4$
$C_2 = 1,3$ $\quad$ $G_2 = 1,2$
$C_3 = 4,3$ $\quad$ $G_3 = 4,2$
$C_4 = 2,5$ $\quad$ $G_4 = 3,5$
$C_5 = 3,5,6$ $\quad$ $G_5 = 2,5,6$
$C_6 = 2,6$ $\quad$ $G_6 = 3,6$

Figura. 5: Par de agrupamientos distintos para los cuales la medida *F-BCubed* Extendido obtiene puntuación máxima

A continuación se analiza cómo se comporta *F-SBCubed* cuando ocurre un incremento $E > 0$ en *P-SBCubed* o *R-SBCubed*. Sustituyendo $\propto = \frac{1}{E}$ en *F-SBCubed* obtenemos

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

Si *P-SBCubed* aumenta en $\varepsilon_p > 0$ o *R-SBCubed* aumenta en $\varepsilon_r > 0$, para los valores de $F_\alpha SBCubed(R, P+\varepsilon_p)$ y $F_\alpha SBCubed(R+\varepsilon_r, P)$ se cumple que

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

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

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

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

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

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

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

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

Según lo anterior, si al comparar un agrupamiento D_2 con un agrupamiento D_1 y la medida *P-SBCubed* obtiene una mayor puntuación para el agrupamiento D_2, mientras que *R-SBCubed* los evalúa con igual puntuación, entonces la medida *F-SBCubed* también evaluará con mayor puntuación al agrupamiento D_2. De igual forma ocurre cuando la medida *P-SBCubed* obtiene la misma puntuación para ambos agrupamientos y la media *R-SBCubed* obtiene una mayor puntuación para el agrupamiento D_2, en este caso, la medida *F-SBCubed* también obtendrá mayor puntuación para D_2. En caso de que ambas medidas obtengan mayor puntuación para el agrupamiento D_2 entonces *F-SBCubed* también obtendrá mayor puntuación para a D_2.

Las condiciones propuestas por Amigó et al., también son cumplidas por *SBCubed* y para demostrarlo analizaremos las medias *P-SBCubed* y *R-SBCubed* de forma separada.

Para analizar el cumplimiento de la Homogeneidad por parte de la medida *P-SBCubed*, se puede observar que todas las puntuaciones de los objetos en los agrupamientos D_1 y D_2 son iguales excepto para los objetos que están contenidos en el grupos G_k. Los objetos de G_k que pertenecen a la clase C_i obtienen una puntuación de $\frac{n_1(n_1+1)}{n(N_1+n_2+1)}$, un valor menor que la puntuación $\frac{n_1+1}{N_1+1}$ que obtienen los mismos elementos pero en el agrupamiento D_2.

De igual forma los objetos de la clase C_j que pertenecen a G_k son evaluados con $\frac{n_2(n_2+1)}{n(N_2+n_1+1)}$, una menor puntuación que $\frac{n_2+1}{N_2+1}$ obtenida por los mismos objetos pero pertenecientes al agrupamiento D_2. Como la puntuación de un agrupamiento es el promedio de la puntuación de todos los objetos, y todos los objetos de D_1 son evaluados con menor o igual puntuación entonces podemos afirmar que la precisión de *SBCubed* cumple con esta condición.

La condición *Homogeneidad* no se puede demostrar en la relevancia a través de la puntuación de los objetos porque, como los pares son tomados de las clases, no necesariamente tienen que pertenecer al grupo G_k. Nos enfocaremos en la puntuación de los pares. Los pares de C_i que pertenecen a G_k, en el agrupamiento D_1 son evaluados como $\frac{n_1+1}{N_1+n_2+1}$, valor que es superado por la puntuación $\frac{n_1+1}{N_1+1}$ que reciben estos mismos pares en D_2. Los pares de objetos del agrupamiento D_1 que se encuentran en C_j son evaluados con valor $\frac{n_2+1}{N_2+n_1+1}$, superado por la puntuación que reciben los pares en el agrupamiento D_2 que obtienen $\frac{n_2+1}{N_2+1}$. Como la puntuación de un agrupamiento es un doble promedio de la puntuación de los pares de objetos, y todos los objetos de D_1 son evaluado con menor o igual puntuación entonces podemos afirmar que la relevancia de *SBCubed* cumple con esta condición.

Demostraremos el cumplimiento de condición de *Completitud* de una forma similar a la demostración de la Homogeneidad. La precisión evalúa a los objetos de G_1 y G_2 con puntuaciones $\frac{n_1+1}{N+1}$ y $\frac{n_2+1}{N+1}$ respectivamente en el agrupamiento D_1, pero prefiere a los objetos que pertenecen al grupo G_k otorgándole la puntuación más favorable de $\frac{n+1}{N+1}$ en al agrupamiento D_2. La relevancia también evalúa a los pares de la clase C_k que se encuentran en los grupos G_1 y G_2 con puntuación $\frac{n_1+1}{N+1}$ y $\frac{n_2+1}{N+1}$ respectivamente en el agrupamiento D_1 y por encima de ellos son preferidos los pares de esta clase que pertenece al grupo G_k en el agrupamiento D_2 que son evaluados con $\frac{n+1}{N+1}$.

Las condiciones restantes son más cómodas para demostrar, ya que especifican todos los grupos de los agrupamientos. La precisión de *SBCubed* cumple con la condición Sanco de Ruido puesto que califica al agrupamiento D_1 con puntuación $\frac{n^3+n^2+2n+6}{(n+1)(n+2)(2n+1)}$ y al agrupamiento D_2 con puntuación $\frac{n^2+2n+2}{(n+2)(2n+1)}$, prefiriendo al agrupamiento D_2. La relevancia obtiene una puntuación de $\frac{n^3+4n^2+7n+2}{(2n+1)(n+1)(n+2)}$ y $\frac{n^2+4n+2}{(n+2)(2n+1)}$ respectivamente para los agrupamientos D_1 y D_2 y de igual forma prefiriendo al agrupamiento D_2.

El criterio de precisión en la condición *Tamaño contra calidad* obtiene una puntuación para el agrupamiento D_1 de $\frac{7n+3}{9n+3}$, superado por la puntuación $\frac{3n^2+5n+2}{(3n+1)(n+2)}$ del agrupamiento D_2. Por otro lado, el criterio de relevancia obtiene una puntuación de $\frac{10n+6}{18n+6}$ y $\frac{3n^3+7n^2+4n+2}{(3n+1)(n+1)(n+2)}$ para los agrupamientos D_1 y D_2 prefiriendo así al agrupamiento D_2.

Como *P-SBCubed* y *R-SBCubed* cumplen con las condiciones propuestas por Amigó et al., entonces *F-SBCubed* también la cumples. Además, el conjunto de casos para los cuales cumple con la nueva

condición es un supraconjunto de casos para los cuales los cumple la medida *BCubed Extendido.* Por tanto, la consideramos más adecuada para evaluar los algoritmos de agrupamientos solapados.

4. CONCLUSIONES

En este trabajo se trató el tema concerniente a las medidas de evaluación de los algoritmos de agrupamiento, con énfasis en los que obtienen agrupamientos solapados. Se realiza un extenso análisis de las medidas existentes, así como de los conjuntos de condiciones utilizadas para valorar la efectividad de las medidas.

A raíz de problemas detectados en la medida actualmente reportada en la literatura como más adecuada, enunciamos una nueva condición.

Una vez enunciada dicha nueva condición, se propuso una nueva familia de medidas de calidad, *SBCubed*. Se demuestra que, respecto a las condiciones previamente existentes, las nuevas medidas son tan adecuadas como las de *BCubed Extendido*, las mejores reportadas hasta el momento. En cuanto a la nueva condición enunciada, si bien las medidas propuestas aún no la cumplen, demostramos que el conjunto de los casos en que nuestras propuestas cumplen el enunciado de la nueva condición es un supraconjunto de los casos en que la familia de medidas *BCubed Exrendido* lo hacen, por lo cual consideramos que constituyen una mejora sobre las mismas.

Obviamente, la principal dirección de trabajo futuro que seguiremos a partir de este punto es la mejora de las nuevas medidas propuestas con el fin de obtener una reformulación de las mismas que cumpla con todas las condiciones.

Nos proponemos igualmente trabajar en el análisis de nuevas situaciones que pudieran implicar la necesidad de enunciar nuevas condiciones a cumplir por los algoritmos de agrupamiento para algunos casos particulares.

5. REFERENCIAS BIBLIOGRÁFICAS

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

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

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

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

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

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

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

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

9.Rosenberg, A.; Hirschberg, J. "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). 2007.

10. Dom, B. An information-theoretic external cluster-validity measure.IBM Research Report. 2001.

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

12. Bagga, A.; Baldwin, B. "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. 1998.

6. SÍNTESISCURRICULARES DE LOS AUTORES

Lic. Henry Rosales Méndez, recibió el título de Licenciado en Ciencia de la Computación (2012, Título de Oro) por la Universidad de Oriente donde fue reconocido como graduado más integral. Actualmente se desempeña como profesor de la Facultad de Matemática y Computación en esta universidad. Su línea de investigación es la Minería de Datos, área donde realizó su Tesis de Diploma.

Yunior Ramírez Cruz, graduado como Licenciado en Ciencia de la Computación (2006, Título de Oro) y Máster en Ciencia de la Computación (2008) por la Universidad de Oriente (UO), se desempeña como Especialista en Informática en la empresa DATYS, Investigador Asociado del Centro de Estudios de Reconocimiento de Patrones y Minería de Datos (CERPAMID) y Profesor Adjunto de la Facultad de Matemática y Computación de la UO. Sus intereses de investigación se centran en el Procesamiento del Lenguaje Natural, específicamente el Reconocimiento de Nombres de Entidades; así como la Minería de Textos, en específico la Construcción Automática de Resúmenes. Es autor o coautor de 12 artículos científicos y tutor de tres trabajos de diploma y una tesis de maestría. Ha participado en seis proyectos de investigación/desarrollo, desempeñando tareas de dirección en dos de ellos.

Recibió el Premio al Mejor Artículo de la Comisión de Lingüística Computacional del X Simposio Internacional de Comunicación Social (2007) y el *Best Student Paper Award* de la conferencia internacional MICAI 2008. Es miembro de la Sociedad Cubana de Matemática y Computación, la Asociación Cubana de Reconocimiento de Patrones y la Sociedad Mexicana de Inteligencia Artificial.