



## Análisis de estabilidad en clusters solapados

†David N. Campo<sup>1,2</sup>, \*Georgina Stegmayer<sup>1,2</sup> y ‡Diego H. Milone<sup>2</sup>

<sup>1</sup> CIDISI – Facultad Regional Santa Fe – Universidad Tecnológica Nacional – CONICET

<sup>2</sup> sinc(i) – Facultad de Ingeniería y Ciencias Hídricas – Universidad Nacional del Litoral – CONICET

†[dncampo@santafe-conicet.gov.ar](mailto:dncampo@santafe-conicet.gov.ar), \*[gstegmayer@santafe-conicet.gov.ar](mailto:gstegmayer@santafe-conicet.gov.ar), ‡[d.milone@ieee.org](mailto:d.milone@ieee.org).

**Abstract** Analyzing the stability of a clustering solution implies to measure the ability of an algorithm to produce similar results from a given input data source. External validation indexes allow to quantify such similarity between a pair of clustering solutions. Within the most used classical indexes it is possible to validate solutions with non-overlapped clusters, where each pattern belongs only to a unique cluster. However, in practical applications, generally there are situations where a single pattern could have more than one label. In this work an external validation index is analyzed from a probabilistic approach and a new relevant reformulation for overlapped solutions is provided. After the presentation of the new index, a set of experiments over artificial and real datasets are discussed. The results show how the new index can correctly measure the similarity between overlapped clusters, thus allowing to analyze the stability in both cases.

**Resumen** Analizar la estabilidad de una solución de clustering implica medir la capacidad de un algoritmo para producir resultados similares dada una misma fuente de datos de entrada. Los índices de validación externa permiten cuantificar dicha similitud entre un par de soluciones de clustering. Dentro de los índices clásicos más utilizados es posible validar soluciones con clusters no solapados, en donde cada patrón sólo puede pertenecer a un cluster. Sin embargo, en aplicaciones prácticas, generalmente se dan situaciones en las que un patrón podría poseer más de una etiqueta. En este trabajo se analiza un índice de validación externa desde un enfoque probabilístico y se provee una reformulación aplicable a soluciones con clusters solapados. Luego de presentar el nuevo índice, se muestran y discuten resultados de experimentos realizados sobre ejemplos artificiales y bases de datos reales. Los resultados muestran cómo el nuevo índice puede medir adecuadamente la similitud entre clusters solapados, permitiendo así analizar la estabilidad en ambos casos.

**Keywords:** stability analysis, overlapped clusters, validation measures.

**Palabras Clave:** análisis de estabilidad, clusters solapados, medida de validación.

### 1. Introducción

Los algoritmos de clustering reciben un conjunto de datos como entrada y mediante un proceso no-supervisado lo particionan en cierto número de clusters o grupos. Se puede definir un cluster como un grupo de objetos homogéneos que poseen alguna medida de similitud entre ellos y que se muestran diferentes a los objetos agrupados en otros clusters [21, 19]. Dado que la aplicación de estos algoritmos sobre una base de datos siempre devuelve algún resultado, incluso cuando no exista estructura en la misma, ha surgido el análisis de estabilidad de soluciones de clustering. Para realizar este análisis es importante medir la capacidad de un algoritmo de agrupamiento para producir grupos similares de forma repetida [4, 18, 5]. Aunque no hay un total acuerdo en cuanto a su definición, existen autores que relacionan el concepto de estabilidad con soluciones de agrupamiento que no se ven alteradas bajo

alguna perturbación de los datos de entrada [5, 3]. Cuando se hace análisis de estabilidad en clustering, también se suele hablar de que se mantengan las estructuras naturales “subyacentes en los datos” y no que aparezcan nuevas estructuras como producto artificial de un algoritmo concreto.

Luego de realizar el agrupamiento en distintas condiciones, se desea medir o cuantificar la similitud entre las soluciones para poder compararlas [21]. Ultimamente, con el crecimiento en el estudio de análisis de agrupamientos se han estado utilizando nuevos algoritmos y medidas de comparación de clusters [16]. Para medir la similitud entre soluciones de clustering se pueden utilizar dos tipos de medidas de validación: internas y externas. Las primeras miden atributos de homogeneidad y separación de los datos en los clusters. Las medidas externas hacen una comparación entre distintas soluciones de clustering, tomando una como referencia y comparándola con otras [10, 11]. Con respecto a las medidas externas, se dispone principalmente de tres tipos para realizar la comparación entre soluciones: las basadas en el conteo de pares de patrones, en las que las soluciones coinciden o no; las basadas en el análisis de correspondencia de conjuntos y las basadas en la estadística y teoría de la información. Con respecto a las primeras, la más utilizada y difundida es la de Fowlkes-Mallows (FM) [8] que trabaja con las frecuencias de pares de patrones que se detecten agrupados juntos en ambas soluciones. En [6] los autores evalúan las distintas soluciones obtenidas cuantificando la similaridad mediante este índice. De las métricas basadas en conjuntos, una muy utilizada es la de máxima coincidencia [14] que se basa en comparar los clusters de ambas soluciones analizadas, haciendo coincidir aquellos que tengan mayor cantidad de elementos en común. Por último, de las medidas basadas en estadística, la información mutua normalizada es una de las más representativas y se basa en cuantificar la cantidad de información compartida entre ambas soluciones, a través del concepto de entropía [16, 13]. Sin embargo, ninguna de estas medidas es capaz de representar correctamente las similitudes al trabajar con soluciones que posean clusters solapados.

Ultimamente se ha suscitado interés por el análisis de soluciones con clusters solapados. En [9] se compara y estudia la evolución de grupos de personas en redes sociales y se propone un algoritmo para computar nuevas distancias entre colecciones de grupos potencialmente solapados. En [20] se propone una nueva medida de validación para clusters solapados en el contexto de recuperación de documentos, en particular tratando el tema de solapamiento entre soluciones con diferente número de clusters. En [1] se propone un conjunto de restricciones sobre métricas para validación externa y se extiende el análisis para tratar soluciones solapadas. En este trabajo, en cambio, se hace un aporte al área ya que presentamos un análisis detallado del índice FM. Con un enfoque probabilístico se analiza cada factor del mismo considerando el solapamiento de clusters y se muestra cómo falla este índice cuando se lo intenta aplicar a soluciones que posean clusters solapados. Luego, a partir del análisis anterior, se deriva una nueva propuesta de índice teniendo especial atención en la situación mencionada. Aplicando ambos índices a diferentes casos de estudio y una base de datos reales, se verifican experimentalmente las mejoras expresadas cuando existen clusters solapados.

La organización de este trabajo es la siguiente. En la Sección 2 se realiza un análisis detallado de los factores del índice FM. En la Sección 3 se deriva el nuevo índice propuesto para poder tratar con las soluciones solapadas. Luego, en la Sección 4, se presentan resultados de aplicar ambos índices sobre datos artificiales y reales, y se discuten dichos experimentos. Por último se presentan las conclusiones del trabajo en la Sección 5.

## 2. Análisis conceptual del índice Fowlkes-Mallows

El índice de Fowlkes-Mallows es una medida de similaridad entre soluciones de clustering, generalmente utilizado para validación externa. Como se mencionó en la Sección 1, en el contexto de análisis de estabilidad se utiliza para poder comparar soluciones y verificar si las mismas son o no estables. FM recibe el etiquetado de ambas soluciones, sobre el mismo conjunto de datos, y devuelve un valor  $B_k \in [0, 1]$  que cuantifica la similitud entre las mismas. El valor 1 representa que ambas son exactamente iguales, mientras que 0 denota completa desigualdad. FM está definido en [8] como

$$B_k = \frac{T_k}{\sqrt{P_k Q_k}}, \quad (1)$$

donde

$$T_k = \sum_{i=1}^k \sum_{j=1}^k m_{ij}^2 - N, \quad (2)$$

$$P_k = \sum_{i=1}^k m_{i*}^2 - N, \quad (3)$$

$$Q_k = \sum_{j=1}^k m_{*j}^2 - N, \quad (4)$$

$$m_{i*} = \sum_{j=1}^k m_{ij}, \quad (5)$$

$$m_{*j} = \sum_{i=1}^k m_{ij}, \quad (6)$$

siendo  $M = \{m_{ij}\}$  la matriz de contingencia obtenida entre 2 soluciones de  $k$  clusters cada una con  $N$  datos en el conjunto a particionar. Esta matriz posee tantas filas como cantidad de clusters haya en la primer solución, que llamaremos  $C$ , y tantas columnas como clusters tenga la segunda,  $C'$ . En cada posición  $ij$  de la matriz se coloca la cantidad de patrones en comun que se pueden contar entre los elementos del cluster  $i$  de la solución  $C$  y los patrones del cluster  $j$  de la solución  $C'$ . Es decir,  $m_{ij} = |c_i \cap c'_j|$ .

A continuación se hará un análisis conceptual de cada factor del índice FM. Consideremos  $T_k$ , y reemplazando  $\sum_{i=1}^k \sum_{j=1}^k m_{ij} = N$  en (2) se puede escribir

$$\begin{aligned} T_k &= \sum_{i=1}^k \sum_{j=1}^k m_{ij}^2 - N = \sum_{i=1}^k \sum_{j=1}^k (m_{ij}m_{ij}) - \sum_{i=1}^k \sum_{j=1}^k m_{ij} \\ &= \sum_{i=1}^k \sum_{j=1}^k m_{ij}(m_{ij} - 1) = 2 \sum_{i=1}^k \sum_{j=1}^k \binom{m_{ij}}{2}, \end{aligned} \quad (7)$$

que representa el doble de la sumatoria de cada uno de los elementos de la matriz de contingencia tomados de a 2. Es decir, los elementos en común entre cada par de clusters de ambas soluciones tomados de a 2. Esto representa la cantidad de formas de elegir dos elementos en dicha intersección.

Análogamente, a partir de (3) se puede obtener

$$\begin{aligned} P_k &= \sum_{i=1}^k m_{i*}^2 - N = \sum_{i=1}^k (m_{i*}m_{i*}) - \sum_{i=1}^k \sum_{j=1}^k m_{ij} \\ &= \sum_{i=1}^k (m_{i*}m_{i*}) - \sum_{i=1}^k m_{i*} = \sum_{i=1}^k (m_{i*}(m_{i*}) - 1) = 2 \sum_{i=1}^k \binom{m_{i*}}{2}, \end{aligned} \quad (8)$$

siendo, de forma similar a como ocurre con  $T_k$ , la cantidad de formas que hay de tomar 2 elementos de cada cluster de la solución analizada. Idéntico razonamiento se aplica para la interpretación de  $Q_k$ .

Consideremos  $C$  y  $C'$ , dos soluciones de clustering con  $k$  clusters. Sean  $x$  e  $y$  dos patrones cualesquiera. Podemos interpretar al primer factor,  $P_k$ , como la probabilidad de que los dos patrones se encuentren en un mismo cluster  $c_i$  de la solución  $C$ , considerando un muestreo uniforme de los datos. Dado que los patrones tienen la misma probabilidad de ser seleccionados, al tomar un segundo patrón bajo la misma hipótesis, la probabilidad de que los dos se agrupen en el cluster  $c_i$  es:

$$Pr(x \in c_i \wedge y \in c_i) = \frac{\binom{|c_i|}{2}}{\binom{N}{2}} = \frac{|c_i|(|c_i| - 1)}{N(N - 1)}, \quad (9)$$

en donde el numerador representa la cantidad de formas de tomar de a 2 los patrones agrupados bajo el cluster  $c_i$  y el denominador considera todas las formas posibles de tomar 2 elementos del conjunto completo de datos. Esto último surge de considerar un caso extremo en el cual todos los patrones queden en un único cluster. Si se consideran los  $k$  clusters de  $C$ , ahora tenemos

$$\tilde{p}_k = \sum_{i=1}^k \frac{\binom{|c_i|}{2}}{\binom{N}{2}} = \frac{1}{N(N-1)} \sum_{i=1}^k |c_i| (|c_i| - 1) = \frac{1}{N(N-1)} \sum_{i=1}^k |c_i|^2 - N, \quad (10)$$

que guarda relación directa con (3), dado que  $|c_i| = m_{i*}$ . La interpretación de  $Q_k$  es la misma que la de  $P_k$ , salvo que se analizan los clusters de la solución  $C'$ .

Con respecto a  $T_k$ , se puede considerar que representa la probabilidad de muestrear un par de patrones al azar y que éstos pertenezcan a un mismo cluster en  $C$  y  $C'$ . El razonamiento es similar al seguido para  $P_k$ , sólo que en vez de considerarse todos los patrones por cluster, se van considerando los pares de patrones comunes al emparejamiento entre cada par de clusters de ambas soluciones

$$Pr((x, y) \in c_i \wedge (x, y) \in c'_j) = \frac{\binom{|c_i \cap c'_j|}{2}}{\binom{N}{2}} = \frac{|c_i \cap c'_j| (|c_i \cap c'_j| - 1)}{N(N-1)}. \quad (11)$$

De la misma forma que en el desarrollo de  $P_k$ , el denominador representa la cantidad de formas de tomar 2 elementos de todo el conjunto de datos. Este sería el caso extremo en el que en ambas soluciones, todos los patrones hayan quedado en un único cluster. Si tomamos entonces todas las posibles comparaciones entre cada par de clusters de ambas soluciones, obtenemos la probabilidad

$$\begin{aligned} \tilde{t}_k &= \sum_{i=1}^k \sum_{j=1}^k \frac{\binom{|c_i \cap c'_j|}{2}}{\binom{N}{2}} = \frac{1}{N(N-1)} \sum_{i=1}^k \sum_{j=1}^k |c_i \cap c'_j| (|c_i \cap c'_j| - 1) = \\ &= \frac{1}{N(N-1)} \sum_{i=1}^k \sum_{j=1}^k |c_i \cap c'_j|^2 - N, \end{aligned} \quad (12)$$

que guarda relación directa con (2) dado que, como se observó antes,  $m_{ij} = |c_i \cap c'_j|$ .

Luego, el índice FM puede entenderse como la razón entre la probabilidad de obtener dos patrones juntos en las soluciones conjuntas sobre la media geométrica de dicha probabilidad considerando las soluciones por separado. Esto es:

$$B_k = \frac{\frac{1}{\binom{N}{2}} \sum_{i=1}^k \sum_{j=1}^k \binom{|c_i \cap c'_j|}{2}}{\sqrt{\frac{1}{\binom{N}{2}} \sum_{i=1}^k \binom{|c_i|}{2} \frac{1}{\binom{N}{2}} \sum_{j=1}^k \binom{|c'_j|}{2}}} = \frac{\sum_{i=1}^k \sum_{j=1}^k (m_{ij}^2 - N)}{\sqrt{\sum_{i=1}^k (m_{i*}^2 - N) \sum_{j=1}^k (m_{*j}^2 - N)}} = \frac{T_k}{\sqrt{P_k Q_k}}. \quad (13)$$

De esta manera se arriba a la formulación original de FM a través del enfoque probabilístico alternativo.

Ahora bien, cuando este índice se aplica a clusters solapados, los resultados que se obtienen no son los intuitivamente esperados. Por ejemplo, en el gráfico de la Figura 1 se observan dos soluciones de clustering a)  $C$  con  $k = 1$  y b)  $C'$  con  $k = 2$ . En c) se encuentra la matriz de contingencia para ambas soluciones. En la solución a), se puede observar cómo todos los patrones se agruparon juntos en un único cluster. En la solución b) se puede observar cómo ambos clusters comparten todos los patrones. En este caso hay un solapamiento completo de ambos grupos.

Para calcular el índice FM sobre este ejemplo se procederá a calcular cada uno de sus factores. El valor del factor  $P_k$  se corresponde con la solución  $C$  y el de  $Q_k$  con la  $C'$ . Así  $P_k = \sum_{i=1}^1 (m_{i*}^2 - N) = 12^2 - 6 = 138$ . En cambio  $Q_k = \sum_{i=1}^2 (m_{*j}^2 - N) = 6^2 + 6^2 - 6 = 66$ . Para el cálculo de  $T_k$  debemos obtener las intersecciones de los objetos del cluster de la solución a) con los de la b). Así, según la matriz de contingencia de la Figura 1.c) obtenemos  $T_k = \sum_{i=1}^1 \sum_{j=1}^2 (m_{ij}^2 - N) = 6^2 + 6^2 - 6 = 66$ . De esta forma  $B_k = 0,69$ , valor que está por debajo de 1, aunque intuitivamente se esperaría que el índice pueda reflejar la similitud entre ambas soluciones. Esto es debido a que la probabilidad de que dos patrones cualesquiera se agrupen juntos en una de las soluciones es mucho menor que el valor obtenido.

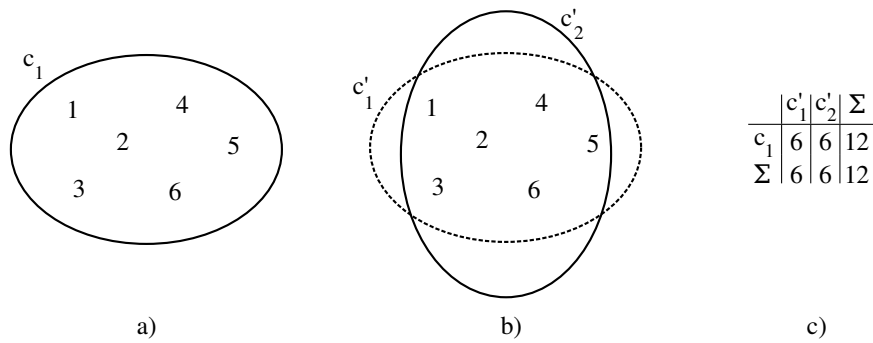


Figura 1: Ejemplo ilustrativo con dos soluciones, a)  $C$  con  $k = 1$ , b)  $C'$  con solapamiento y  $k = 2$  y c) matriz de contingencia correspondiente a las soluciones  $C$  y  $C'$ .

### 3. Índice de estabilidad para clusters solapados

Como se ha observado a través del ejemplo anterior, cuando se trabaja con soluciones que poseen clusters solapados el índice FM es incapaz de producir resultados correctos. Ante la necesidad de contar con una herramienta que permita cuantificar la similitud de soluciones de clustering, ya sean estas solapadas o no, en esta sección se reinterpreta dicho índice y se propone uno nuevo denominado *overlapped FM* (oFM). La definición del nuevo índice parte de una forma más cuidadosa al calcular las probabilidades de que un patrón pueda agruparse junto con otro, ya sea en una solución, en la otra o en ambas a la vez. Particularmente, se debe tener especial cuidado en la normalización de las frecuencias, ya que se quiere evitar la contabilización repetida de los patrones sin considerar la cantidad de veces que aparezcan en otros clusters. Para que el nuevo índice pueda considerar la situación propuesta, se redefinirán cada uno de sus factores. Además, para ganar generalidad, ahora se considerarán explícitamente soluciones con distintos tamaños. Así se tomará  $k_1$  para denotar la cantidad de clusters de  $C$  y  $k_2$  para la de  $C'$ .

Con respecto a  $\tilde{p}_k$ , la nueva situación a ser contemplada admite escenarios en donde los patrones podrían llegar a estar en más de un cluster. Se debe tener especial consideración de no contabilizar dichos objetos más de una vez, o bien, normalizarlos para que su probabilidad de ocurrencia esté bien acotada. Para lograrlo, se debe tener en cuenta un caso extremo en el que todos los patrones estén en todos los clusters. A diferencia de la forma anterior de calcular este factor, ahora se lo debería normalizar por este caso extremo, dividiendo por algún valor que tenga relación directa con la cantidad de veces que podría llegar a contarse a todos los pares de patrones. Se redefine luego la nueva forma de calcular la probabilidad de que dos patrones se agrupen juntos en una misma solución como

$$\tilde{p}_k = \frac{\sum_{i=1}^{k_1} \binom{|c_i|}{2}}{k_1 \binom{N}{2}}, \tag{14}$$

donde en el denominador se están considerando  $k_1$  clusters con todos los patrones agrupados juntos,  $k_1$  veces. En el numerador se cuentan los pares de patrones directamente tantas veces como clusters haya. Luego, para  $\tilde{q}_k$ , la situación es similar. Así se define

$$\tilde{q}_k = \frac{\sum_{j=1}^{k_2} \binom{|c'_j|}{2}}{k_2 \binom{N}{2}}. \tag{15}$$

El último factor a considerar es  $\tilde{t}_k$ , donde interviene la interacción de ambas soluciones. La cuenta de los pares de patrones agrupados es similar a la realizada en los casos anteriores. En cambio, la normalización de la misma involucra ciertos valores que surgen de ambas soluciones. Sean  $n_1$  y  $n_2$  la cantidad de elementos dentro de cada solución, considerando repeticiones por solapamiento. Luego, en la influencia recíproca de ambas soluciones, observamos intuitivamente que la cantidad de pares de elementos que

podrían contarse está limitada por el valor más pequeño. Esto es, si  $n_1 < n_2$ , la cantidad de datos que puedan llegar a contarse en el emparejamiento entre ambas nunca podrá superar  $n_1$ , puesto que hay a lo sumo esa cantidad de objetos para contar en una de las soluciones. Así se tiene una noción de cómo se puede normalizar esta frecuencia, en donde la idea sigue el mismo camino que en las probabilidades de los factores anteriores. A su vez esta cantidad se la limita por la cantidad real de patrones  $N$ .

Por otro lado, hay que considerar también que la cantidad de clusters en cada solución puede ser diferente. Por ello, en el denominador se debe considerar también el caso límite en el que todos los patrones estén juntos en ambas soluciones, tantas veces como el máximo de solapamientos que haya entre las mismas. Este valor estará limitado por la proporción existente entre la cantidad máxima de clusters que existe en ambas soluciones multiplicado por la cantidad de formas de tomar todos los patrones de a pares. De esta forma se define  $\tilde{t}_k$  como

$$\tilde{t}_k = \frac{\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \binom{|c_i \cap c'_j|}{2}}{\binom{N}{2} \max(k_1, k_2) \frac{\min(n_1, n_2)}{N}}, \quad (16)$$

donde en el numerador se cuentan como antes las intersecciones de los patrones de cada uno de los clusters de una solución con los de la otra, tomados de a dos. El denominador, como se dijo anteriormente, representa el caso extremo en el que todos los patrones se agrupan juntos tantas veces como clusters haya en la solución.

Retomando el ejemplo de la Sección 2, el cálculo del nuevo índice arroja los siguientes valores. Para  $\tilde{p}_k = \sum_{i=1}^1 \binom{|c_i|}{2} / \binom{6}{2} = 15/15 = 1$ . Luego  $\tilde{q}_k = \sum_{j=1}^2 \binom{|c'_j|}{2} / 2 \binom{6}{2} = 30/30 = 1$ . Y considerando  $\max(1, 2) = 2$  y  $\min(6, 12)/6 = 1$ , se tiene  $\tilde{t}_k = \sum_{i=1}^1 \sum_{j=1}^2 \binom{|c_i \cap c'_j|}{2} / 2 \binom{6}{2} = 30/30 = 1$ . Ahora sí se observa cómo el valor del nuevo índice oFM =  $\tilde{t}_k / \sqrt{\tilde{p}_k \tilde{q}_k} = 1$  refleja la similitud esperada entre ambas soluciones, puesto que la probabilidad de encontrar dos patrones agrupados juntos en cualquiera de ellas es efectivamente 1.

## 4. Resultados y Discusión

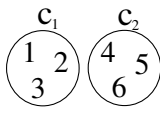
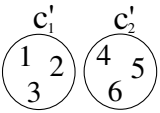
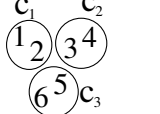
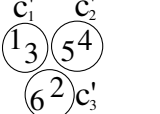
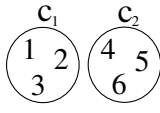
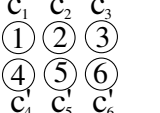
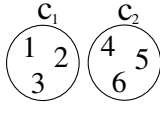
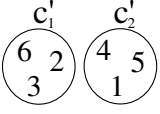
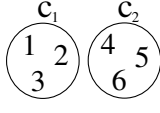
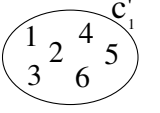
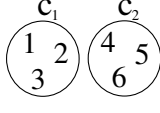
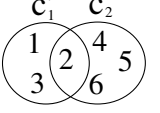
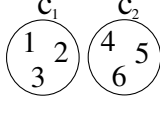
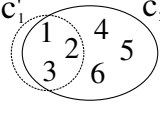
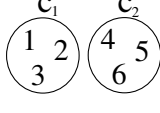
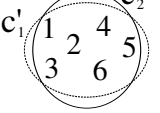
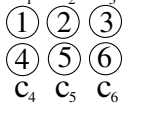
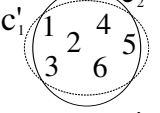
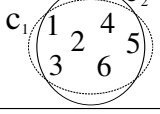
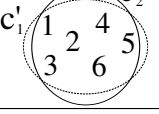
En esta sección se presentan los resultados obtenidos en la evaluación de ambos índices. Primero se muestra y comenta una serie de ejemplos sencillos generados de forma artificial para observar el comportamiento de los índices en casos extremos y particulares. Luego se describen los conjuntos de datos reales utilizados en los experimentos y se presenta el algoritmo de clustering utilizado para agrupar los datos. A continuación se presentan los resultados obtenidos con ambos índices sobre dichos conjuntos de datos y finalmente se contrastan los resultados de ambos.

### 4.1. Casos de estudio

Como se puede observar en el Cuadro 1, se han creado 10 ejemplos artificiales que muestran situaciones de agrupamiento interesantes. En dicho cuadro se enumera cada caso del  $I$  al  $X$ . Luego, se presentan dos columnas que representan las soluciones a ser comparadas:  $C$  y  $C'$ . Finalmente, las dos últimas columnas representan los valores de los índices FM y oFM, respectivamente. En el ejemplo  $I$  se observa claramente que ambas soluciones son exactamente las mismas. No existen clusters solapados y ambos índices logran mostrar correctamente las similitudes de  $C$  y  $C'$ . En  $II$  se puede apreciar como cada solución posee 3 clusters, pero en ningún caso existen pares de patrones que se puedan encontrar simultáneamente en ambas. Así los dos índices logran, nuevamente, reflejar correctamente la situación observada. Algo similar ocurre con el siguiente ejemplo, el  $III$ . En este escenario la solución  $C'$  presenta la particularidad de que no es posible agrupar patrones de a pares puesto que está cada uno en un cluster distinto, y por ello esta solución no se asemeja en nada a la  $C$ .

El ejemplo  $IV$  se plantea un caso similar a  $I$  pero con un par de patrones intercambiados en los clusters de la solución  $C'$ . Esto provoca que la cantidad total de pares de patrones que se pueden contabilizar en ambas soluciones a la vez decaiga en una proporción de 3 con respecto al ejemplo  $I$  anteriormente mencionado, y esto se refleja perfectamente tanto en el valor de FM como en el de oFM. En  $V$  se observa

cómo la solución  $C'$  posee sólo un cluster y cómo solamente la mitad de los patrones de la misma se podrían considerar similarmente agrupados en  $C$ , sean estos los que están en  $c_1$  o  $c_2$ . De hecho, el valor de oFM es más cercano a 0,50, que es lo esperado.

Ejemplos sobre datos artificiales				
	Soluciones		Índices	
	$C$	$C'$	FM	oFM
I	$c_1$ $c_2$ 	$c'_1$ $c'_2$ 	1,00	1,00
II	$c_1$ $c_2$ 	$c'_1$ $c'_2$ 	0,00	0,00
III	$c_1$ $c_2$ 	$c'_1$ $c'_2$ $c'_3$ 	0,00	0,00
IV	$c_1$ $c_2$ 	$c'_1$ $c'_2$ 	0,33	0,33
V	$c_1$ $c_2$ 	$c'_1$ 	0,63	0,45
VI	$c_1$ $c_2$ 	$c'_1$ $c'_2$ 	0,68	0,82
VII	$c_1$ $c_2$ 	$c'_1$ $c'_2$ 	0,54	0,87
VIII	$c_1$ $c_2$ 	$c'_1$ $c'_2$ 	0,45	0,89
IX	$c_1$ $c_2$ $c_3$ 	$c'_1$ $c'_2$ 	0,17	0,00
X	$c_1$ $c_2$ 	$c'_1$ $c'_2$ 	0,49	1,00

Cuadro 1: Resultados de los índices FM y oFM para algunos ejemplos de prueba artificiales.

	clusters en $C'$	FM		oFM	
		$Vn = 0$	$Vn = 1$	$Vn = 0$	$Vn = 1$
Iris	$k_1 = 4$ vs $k_2 = 25$	0,38	0,30	0,15	0,46
	$k_1 = 4$ vs $k_2 = 100$	0,17	0,16	0,03	0,13
	$k_1 = 25$ vs $k_2 = 100$	0,33	0,23	0,16	0,45
Wine	$k_1 = 4$ vs $k_2 = 25$	0,40	0,31	0,16	0,48
	$k_1 = 4$ vs $k_2 = 100$	0,19	0,18	0,04	0,15
	$k_1 = 25$ vs $k_2 = 100$	0,34	0,23	0,17	0,43
Yeast	$k_1 = 4$ vs $k_2 = 25$	0,32	0,23	0,13	0,39
	$k_1 = 4$ vs $k_2 = 100$	0,16	0,14	0,03	0,12
	$k_1 = 25$ vs $k_2 = 100$	0,29	0,18	0,14	0,38
Glass	$k_1 = 4$ vs $k_2 = 25$	0,33	0,27	0,13	0,41
	$k_1 = 4$ vs $k_2 = 100$	0,15	0,14	0,03	0,12
	$k_1 = 25$ vs $k_2 = 100$	0,38	0,24	0,19	0,50

Cuadro 2: Resultados de los índices FM y oFM en la base de datos Iris, Wine, Yeast y Glass para soluciones de referencia  $C$  de 4 y 25 clusters y sin solapamiento ( $Vn = 0$ ) contra soluciones  $C'$  de 25 y 100 clusters tomando  $Vn = 0$  y  $Vn = 1$ .

Los ejemplos *VI*, *VII* y *VIII* se analizan en forma conjunta, ya que se observa cómo en cada una de las soluciones  $C'$  los clusters poseen progresivamente mayor solapamiento a medida que se avanza en cada ejemplo. Claramente se ve cómo FM decae cuando más solapamiento se considera, mientras que oFM se comporta de manera inversa. Este último comportamiento es el esperado, dado que al aumentar el nivel de solapamiento aumentan las posibilidades de obtener más pares de patrones agrupados juntos en la otra solución. En el ejemplo *IX* se observa una situación similar a la de *III*, en cuanto a que en una de las soluciones no se pueden agrupar patrones de a pares, y por ello oFM arroja una similitud nula; mientras que FM muestra algún grado de parecido que no se corresponde con la realidad. Por último, en el ejemplo *X* se ve una situación similar a la de *I* ya que ambas soluciones son las mismas, aunque ahora con clusters totalmente solapados. Se observa como oFM refleja correctamente esta semejanza entre las soluciones con un valor = 1 y FM falla debido al solapamiento.

## 4.2. Datos reales

Para los experimentos con datos reales se utilizaron 4 bases de datos: el conjunto de datos de la flor del Iris<sup>1</sup>, de Wine<sup>2</sup>, de Yeast<sup>3</sup> y de Glass<sup>4</sup> [2]. El conjunto de datos de Iris posee 4 atributos y 50 registros de cada una de 3 especies distintas de la flor, haciendo un total de 150 patrones [7]. De las tres clases existentes en el conjunto, sólo una es linealmente separable de las otras dos, teniendo estas últimas patrones de distinta clase muy cercanos entre sí en el espacio de atributos. Se ha seleccionado este conjunto por ser sencillo y pequeño, además del acceso libre y su aceptación en el ámbito científico. Por su parte, Wine representa la medición y análisis de 13 atributos químicos realizados sobre vinos de una misma región de Italia, pero tomados de diferentes cultivos. Este conjunto de datos consta de 178 patrones distribuidos en 3 grupos: cultivo A con 59 patrones, el B con 71 y el C con 48. La base

<sup>1</sup><http://archive.ics.uci.edu/ml/datasets/Iris>

<sup>2</sup><http://archive.ics.uci.edu/ml/datasets/Wine>

<sup>3</sup><http://archive.ics.uci.edu/ml/datasets/Yeast>

<sup>4</sup><http://archive.ics.uci.edu/ml/datasets/Glass+Identification>



de datos de Yeast, por su parte, representa un estudio sobre la levadura en donde se busca determinar la localización de sus proteínas en las células. Posee 1484 patrones distribuidos en 10 grupos; con 463, 429, 244, 163, 51, 44, 37, 30, 20 y 5 instancias en cada uno. A cada patrón se le realizaron 8 tipos de mediciones. Por último, el conjunto Glass posee 214 patrones distribuidos en 7 grupos o tipos de vidrios, en donde a cada objeto se le han medido 9 atributos. Entre ellos se encuentran en índice de refracción del vidrio y el contenido de óxido de distintos elementos como ser sodio, magnesio, aluminio, silicio, potasio, calcio, bario y hierro. Para el agrupamiento de los datos se utilizó un mapa auto-organizativo (SOM, de su nombre en inglés *Self-Organizing Map*) [12, 15].

Para el agrupamiento se entrenaron mapas con distintas cantidades de neuronas, es decir, clusters. Todos los mapas utilizados tienen una topología rectangular en forma de grilla y la cantidad de iteraciones de entrenamiento fue fijada en 100 épocas. La inicialización del mapa fue determinística, utilizando PCA [17] sobre los datos. Para considerar solapamiento en las neuronas del mapa se tomaron las neuronas vecinas inmediatas, es decir la de la izquierda, derecha, arriba y abajo, de cada neurona. Así, al considerar vecindad  $Vn = 0$  cada neurona es un único cluster y  $Vn = 1$  cada neurona y sus cuatro vecinas forman un mismo cluster. Por ello, varios patrones pueden llegar a pertenecer a más de un cluster y de esta forma pertenecer a clusters solapados.

En el Cuadro 2 se observa una primer columna en donde se indican las cantidades de clusters considerados para las soluciones  $C$  y  $C'$ . Luego el cuadro se divide en dos grandes columnas que representan los valores de los índices analizados, FM y oFM. Cada una de estas columnas se vuelven a dividir para mostrar los resultados de las pruebas con soluciones  $C'$  que poseen y no poseen solapamientos. La cantidad de clusters de las soluciones fueron de 4, 25 y 100. Para la solución tomada como referencia  $C$  no se consideró solapamiento. En el caso de Iris, cuando se toma  $k_1 = 4$  FM disminuye su valor no sólo al tomar solapamiento sino también al pasar de  $k_2 = 25$  a  $k_2 = 100$ . Por otro lado oFM aumenta notablemente al considerar solapamiento, pero también disminuye al aumentar la cantidad de clusters en la solución  $C'$ . Se observa como FM vuelve a decaer al considerar solapamiento cuando  $k_1 = 25$  y  $k_2 = 100$ . Se puede visualizar además cómo oFM mantiene la tendencia de aumentar al considerar solapamiento pero, al igual que FM, aumenta al pasar de  $k_1 = 4$  a  $k_1 = 25$  cuando  $k_2 = 100$ . Para la base de datos Wine se percibe, de la misma forma, que en todas las pruebas el valor del índice FM experimenta un descenso de su valor cuando se pasa de no considerar solapamiento a considerarlo. Se observa además cómo el valor de oFM empieza a subir cuando se toma en cuenta el solapamiento. Al observar los valores de Yeast, se percibe cómo baja el índice de FM cuando se toma  $Vn = 1$  en vez de  $Vn = 0$  en  $C'$  y cómo, opuestamente, el índice propuesto comienza lentamente a crecer cuando se pasa a considerar vecindad en las soluciones  $C'$ . Por el último, al ver la evolución de los valores para Glass, se puede apreciar, nuevamente, cómo los valores que arroja para FM decrecen en todos los casos que se aplicó solapamiento, y una tendencia opuesta en el índice oFM.

En todos los experimentos, considerando los distintos conjuntos de datos, el valor de FM decae cuando aumenta el solapamiento para la solución  $C'$ , de forma totalmente opuesta a lo que sucede con los valores de oFM. Se ve además que siempre los dos índices mostraron mayor valor, es decir mayor semejanza de las soluciones comparadas, a medida que la cantidad de clusters de  $C$  se asemejaba a la de  $C'$ . En el caso de FM este valor disminuía cuando se tenía en cuenta solapamiento, mientras que se mostraba un incremento notorio en el valor que arrojaba el índice que se propone. Esto es debido a que a mayor cantidad de clusters hay mayor dispersión de los patrones en los mismos y por ello menor es el valor de los índices calculados. Esto último es consistente con lo que se observa en todas las comparaciones donde  $k_1 = 4$  y  $k_2 = 100$ . En dicho caso los valores no disminuyen o aumentan prácticamente, sea para FM o para oFM respectivamente, puesto que existe mucha dispersión de los patrones en el mapa  $C'$ . De esta forma, el efecto de considerar vecindad hace que el solapamiento sea más pequeño en comparación con los casos en donde los mapas no tienen tanta diferencia en cuanto a tamaño. Es por ello que tanto FM apenas baja y oFM incrementa poco en la consideración de solapamiento con soluciones con tanta diferencia en cantidad de clusters.

## 5. Conclusiones y trabajos futuros

En este trabajo se propuso un nuevo índice para análisis de estabilidad entre soluciones solapadas, parcial o completamente. Se realizó un análisis del índice FM para mostrar sus falencias al momento de aplicarlo sobre soluciones solapadas y a partir de éste se derivó uno nuevo. El índice propuesto refleja correctamente la similitud entre soluciones que pueden o no poseer clusters solapados. Esto se alcanzó teniendo especial cuidado en la forma de contar casos extremos en los que se pudiesen agrupar los patrones, permitiendo así una normalización más cuidadosa. Luego, en los experimentos se mostró como el índice FM es incapaz de tratar con soluciones solapadas, mientras que oFM puede sortear dicho problema.

Con respecto a trabajos futuros, se espera realizar mayor cantidad de pruebas de los índices utilizando otros algoritmos y posiblemente nuevos conjuntos de datos. También se desea incorporar pruebas de significancia estadística a los resultados para poder arribar a conclusiones más sólidas. También sería deseable a futuro la incorporación al análisis de nuevos tipos de índices y derivar así nuevas métricas que se puedan comparar con las de este trabajo.

## Referencias

- [1] E. Amigó, J. Gonzalo, J. Artilles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. 12:461–486. doi: 10.1007/s10791-008-9066-8.
- [2] K. Bache and M. Lichman. UCI machine learning repository, 2013.
- [3] A Bayá. *Aplicación de algoritmos no supervisados a datos biológicos*. PhD thesis, Universidad Nacional de Rosario, Marzo 2011.
- [4] S. Ben-David, U.V. Luxburg, and D. Pál. A sober look at clustering stability. In *COLT, Springer*, pages 5–19, 2006. doi: 10.1007/11776420\_4.
- [5] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. pages 6–17, 2002.
- [6] A. Ben-Hur and I. Guyon. Detecting stable clusters using principal component analysis. pages 159–182, 2003. doi: 10.1385/1-59259-364-X:159.
- [7] R.A. Fisher. The use of multiple measurements in taxonomic problems. pages 179–188, 1936.
- [8] E. Fowlkes and C. Mallows. A method for comparing two hierarchical clusterings. 78:553–569, 1983.
- [9] Mark K. Goldberg, Mykola Hayvanovych, and Malik Magdon-Ismail. Measuring similarity between sets of overlapping clusters. pages 303–308, 2010. doi: 10.1109/SocialCom.2010.50.
- [10] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. pages 107–145, 2001.
- [11] J. Handl, J. Knowles, and D.B. Kell. Computational cluster validation in post-genomic data analysis. pages 3201–3212, 2005. doi: 10.1093/bioinformatics/bti517.
- [12] T. Kohonen. Self-organized formation of topologically correct feature maps. 43:59–69, 1982.
- [13] M. Meilă. Comparing clusterings—an information based distance. pages 873–895, May 2007. doi: 10.1016/j.jmva.2006.11.013.
- [14] M. Meilă and D. Heckerman. An experimental comparison of model-based clustering methods. pages 9–29, January 2001.
- [15] D. Milone, G. Stegmayer, L. Kamenetzky, M. Lopez, J. Giovannoni, J.M. Lee, and F. Carrari. \*omesom: a software for integration, clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants. pages 438–448, 2010. doi: 10.1186/1471-2105-11-438.

- 
- [16] X. V. Nguyen, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. pages 2837–2854, 2010.
- [17] K. Pearson. On lines and planes of closest fit to systems of points in space. pages 559–572, 1901.
- [18] Ohad Shamir and Naftali Tishby. Model selection and stability in  $k$ -means clustering. pages 213–243, 2010. doi: 10.1007/s10994-010-5177-8.
- [19] David Skillicorn. *Understanding complex datasets. Data mining with matrix decompositions*. Ed. Chapman & Hall / CRC, 2007.
- [20] J. Wu, H. Yuan, H. Xiong, and G. Chen. Validation of overlapping clustering: A random clustering perspective. 180:4354–4369, November 2010. doi: 10.1016/j.ins.2010.07.028.
- [21] R. Xu. and D. Wunsch. *Clustering*. IEEE Press Series on Computational Intelligence. Ed. Wiley, 2009.