Métricas para la validación de Clustering

MINERIA DE DATOS

Elizabeth León Guzmán, Profesor Asociado

Universidad Nacional de Colombia
Ingeniería de Sistemas y Computación

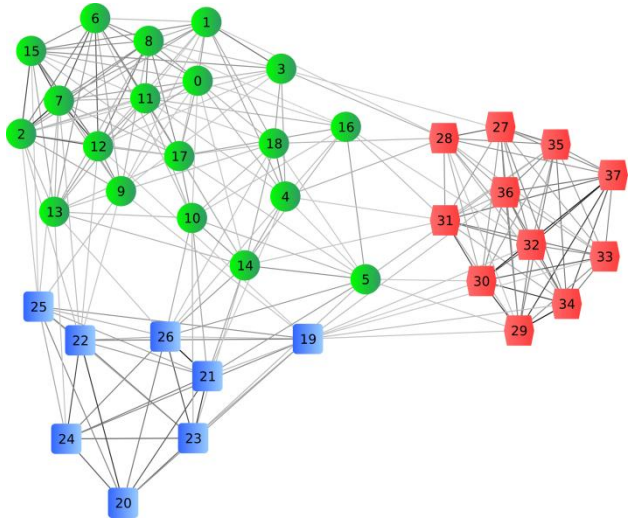
Contenido

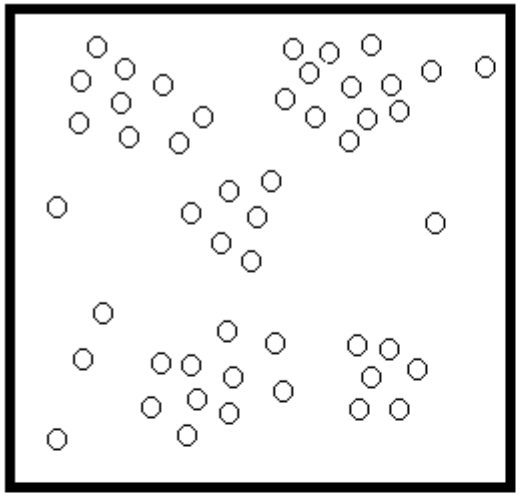
- Introducción
- Tipos de Validación
- Métricas de validación internas
  - Cohesión y separación
  - SSW
  - SSB
  - Sum of Squared base Indexes
  - Davies Bouldin
  - Coeficiente de Silhouette
- Métricas de validación externas
  - Precisión
  - Recall
  - Medida F
  - Entropía
  - Pureza
  - Mutual Information

Introducción

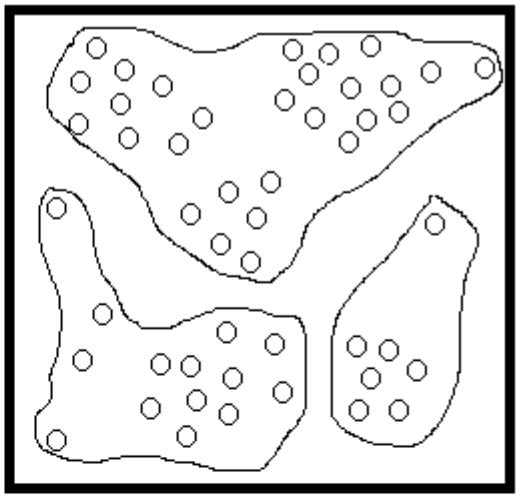
El Clustering (Agrupamiento) es un proceso no supervisado en Minería de Datos y en el Reconocimiento de Patrones, ampliamente utilizado y que es especialmente sensible a los parámetros de la entrada.

Es de importancia evaluar el resultado de los algoritmos de clustering, sin embargo, es difícil definir cuando el resultado de un agrupamiento es aceptable. Por esta razón existen técnicas e índices para la validación de un agrupamiento realizado.

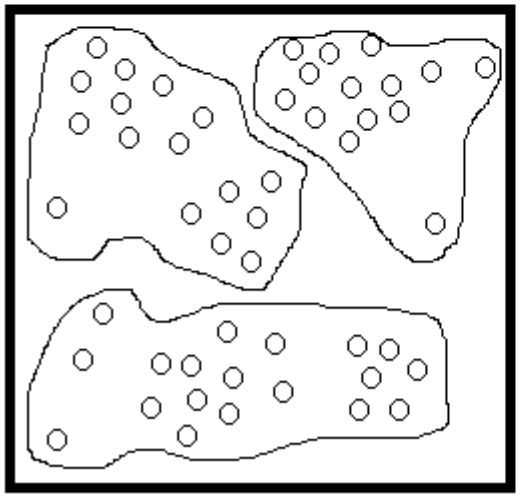




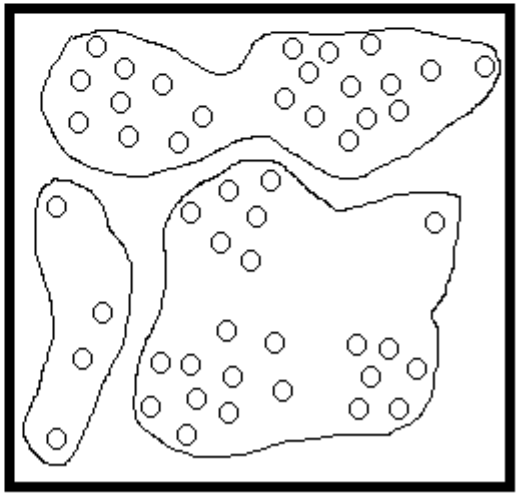
Puntos aleatorios



DBSCAN



K-means



Complete link

Tipos de Validación

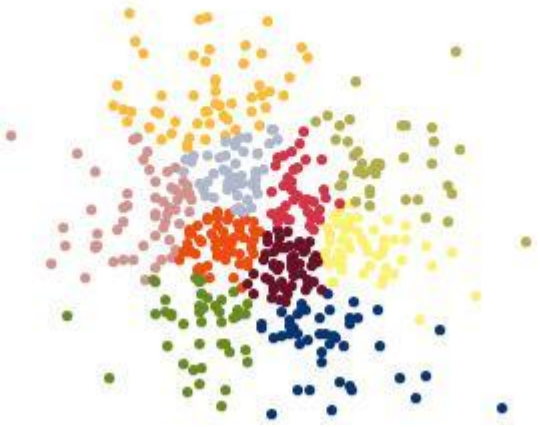
- La validación **externa** y la validación **interna** son las dos categorías más importantes para la validación de clustering. La principal diferencia es si se usa o no información externa para la validación, es decir, información que no es producto de la técnica de agrupación utilizada.

- A diferencia de técnicas de validación externas, las de validación interna miden el clustering únicamente basadas en información de los datos. Evalúan que tan buena es la estructura del clustering sin necesidad de información ajena al propio algoritmo y su resultado.

- Como la validación externa mide la calidad del agrupamiento conociendo información externa de antemano, es principalmente usada para escoger un algoritmo de clustering óptimo sobre un data set específico.

- Las métricas de validación interna pueden usarse para escoger el mejor algoritmo de clustering ,así como el número de clúster óptimo sin ningún tipo de información adicional.

- En la práctica , la información externa, como los labels de las clases , por lo general no se encuentra disponible en muchos escenarios de aplicación.



Métricas de Validación Interna

- Como el objetivo del clustering es agrupar objetos similares en el mismo clúster y objetos diferentes ubicarlos en diferentes clúster, las métricas de validación interna están basadas usualmente en los dos siguientes criterios:
  - **Cohesión**
  - **Separación**

Cohesión y Separación

- Cohesión: El miembro de cada clúster debe ser lo más cercano posible a los otros miembros del mismo clúster.

- Separación: Los clúster deben estar ampliamente separados entre ellos. Existen varios enfoques para medir esta distancia entre clúster: distancia entre el miembro más cercano, distancia entre los miembros más distantes o la distancia entre los centroides.

Sum of Squared Within (SSW)

Medida interna especialmente usada para evaluar la Cohesión de los clústeres que el algoritmo de agrupamiento generó.

$$SSE = \sum_{i=1}^{k} \sum_{x \epsilon C_i} dist^2(m_i, x)$$

Siendo k el número de clústeres, x un punto del clúster C_i y m_i el centroide del clúster C_i.

Sum of Squared Between (SSB)

Es una medida de separación utilizada para evaluar la distancia inter-clúster (Separación)

$$SSB = \sum_{j=1}^{k} n_j \, dist^2 \, (c_j - \bar{x})$$

Siendo k el número de clústeres, n_j el número de elementos en el clúster j, c_j el centroide del clúster j y $\bar{x}$ es la media del data set.

Sum of Squares based Indexes

Los índices o medidas basadas en las «sumas de cuadrados» presentadas anteriormente se caracterizan por medir o cuantificar la dispersión de los puntos a nivel inter-cluster e intra-cluster. Los índices son:

Ball y Hall (1965) $\frac{SSW}{k}$

Calinski y Harabasz (1974) $\frac{SSB/(k-1)}{SSW/(n-k)}$

Hartigan (1975) $log(\frac{SSB}{SSW})$

Xu (1997) $\mathrm{d} * \log\left(\sqrt{\frac{SSW}{dN^2}}\right) + \log(k)$

Siendo k el número de clústeres, N el número de datos y d la dimensión de los datos.

Otros Índices Internos

Existe otro grupo de índices o métricas internas que no tiene relación con los mencionados anteriormente y que se basan en otros criterios:

- Índice Davies-Bouldin (DB)
- Coeficiente de Silhouette

Davies-Bouldin index (DB)

Éste índice está definido como

$$DB = \frac{1}{k} \sum_{i=1, i \neq j}^{k} \max \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

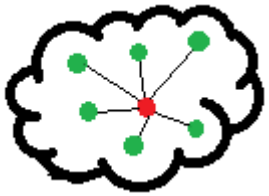
Donde k es el número de clústeres, σ_i es la distancia promedio entre cada punto en el clúster i y el centroide del clúster, σ_j es la distancia promedio entre cada punto del clúster j y el centroide del clúster, y $d(c_i, c_j)$ es la distancia entre los centroides de los 2 clústeres .

Valores pequeños para el índice DB indica clústeres compactos, y cuyos centros estas bien separados los unos de los otros. Consecuentemente el número de clústeres que minimiza el índice DB se toma como el óptimo.

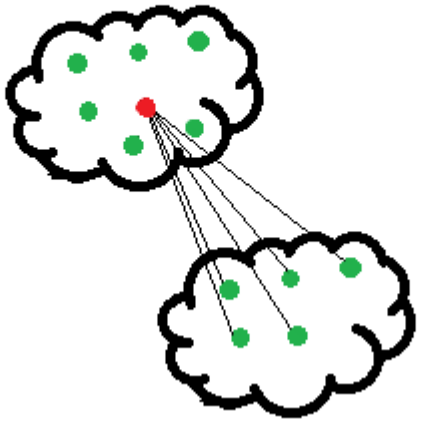
Coeficiente de Silhouette

Dado un punto x del conjunto de datos :

- Cohesión a(x) : distancia promedio de x a todos los demás puntos en el mismo clúster.



- Separación b(x) : distancia promedio de x a todos los demás puntos en el clúster más cercano.



Coeficiente de Silhouette

El coeficiente de silhouette para el punto x está definido como:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

- Donde el valor de s(x) puede variar entre -1 y 1. –
  - -1 = mal agrupamiento
  - 0 = indiferente
  - 1 = bueno

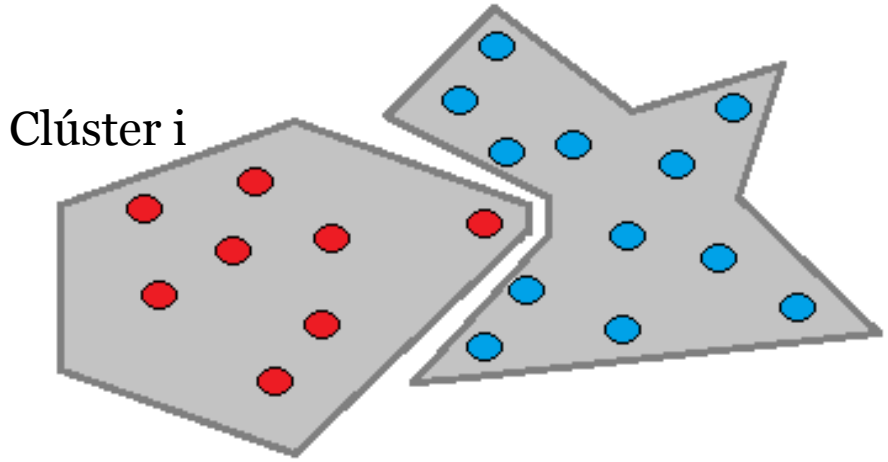
El coeficiente de Silhouette para todo el agrupamiento es :

$$SC = \frac{1}{N}\sum_{i=1}^{N} s(x)$$

Métricas de Validación Externa

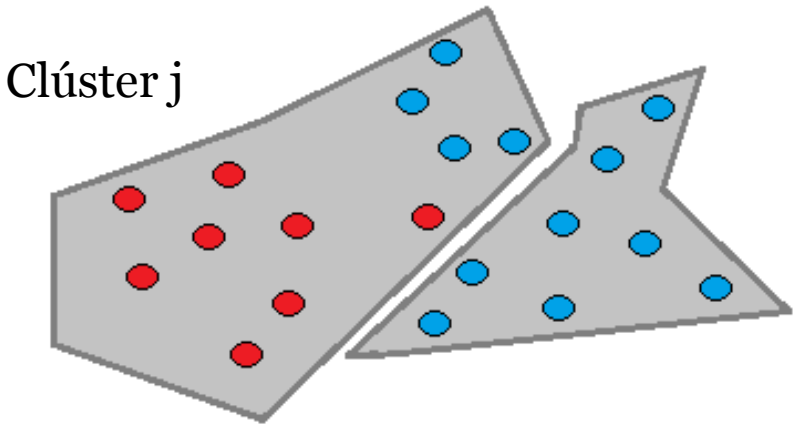
- Cuando se tiene información externa tal como la clase de cada dato, es común y ampliamente utilizado el siguiente análisis:

Se tiene la clase de cada dato en el data set, es decir, se tiene de antemano el número de clúster y a cual clúster pertenece cada dato.



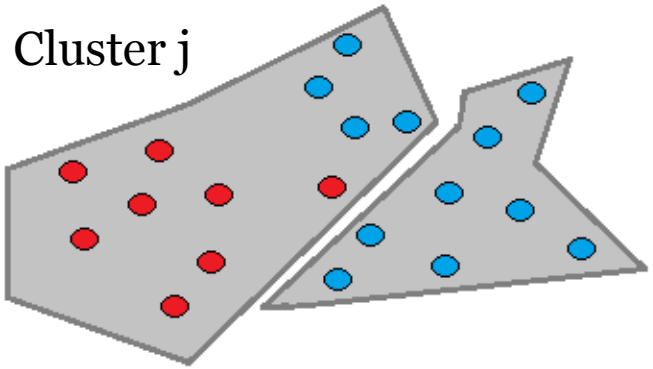
Métricas de Validación Externa

- Una vez realizado el agrupamiento mediante algún algoritmo para ese propósito (k-means, DBSCAN), el algoritmo puede sugerir un nuevo agrupamiento de los datos , diferente al que indicaran las clases conocidas de antemano:



Métricas de Validación Externa

- Teniendo el agrupamiento sugerido por el algoritmo utilizado, el pasos siguiente es la construcción de una tabla como la siguiente

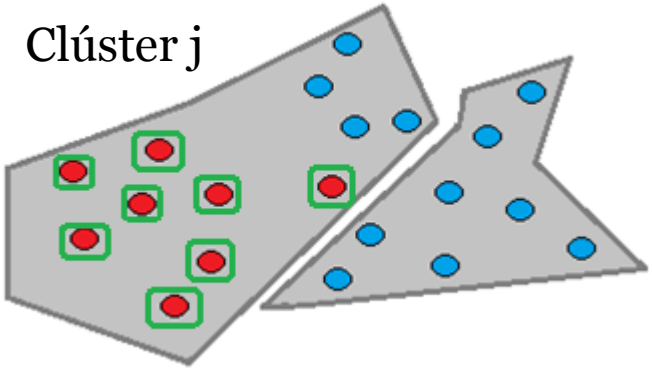


		Verdad	
		P	**N**
Hipótesis	P		
	N		

- Se ubica en las columnas la información referente a la «Verdad», es decir, las etiquetas de las clases que se conocen de antemano. A nivel de las filas se maneja la información correspondiente al resultado del algoritmo

Métricas de Validación Externa

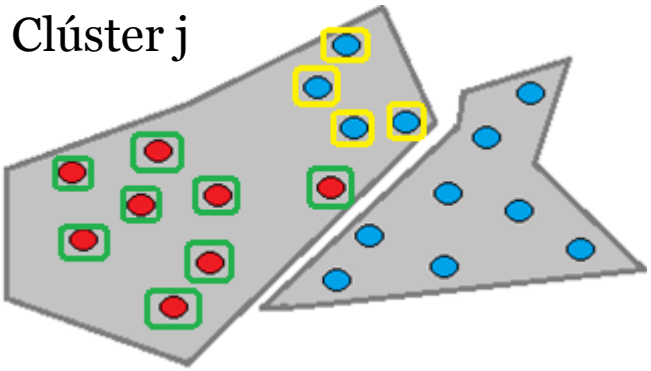
- Se comienza a trabajar con el concepto de «VP» o «Verdadero Positivo». Este termino hace referencia a aquellos puntos que fueron ubicados por el algoritmo en el mismo clúster que indicaba la clase con la que se contaba de antemano.



		Verdad	
		P	**N**
Hipótesis	P	**VP**	
	N		

Métricas de Validación Externa

- El siguiente concepto es el de «Falso Positivo» (FP) que hace referencia a aquellos puntos que fueron ubicados por el algoritmo en el clúster j y que en realidad pertenecían a otro clúster.

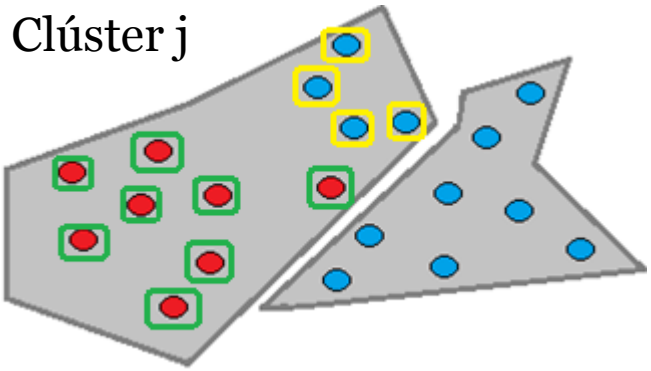


		Verdad	
		P	**N**
Hipótesis	P	**VP**	**FP**
	N		

Métricas de Validación Externa

- Los «Falsos Negativos» (FN) hacen referencia a aquellos elementos del clúster j que fueron ubicados en un clúster diferente al que indicaba su etiqueta. En el ejemplo presentado, el clúster j tiene todos sus elementos asignados correctamente, luego no hay falsos negativos.

Clúster j

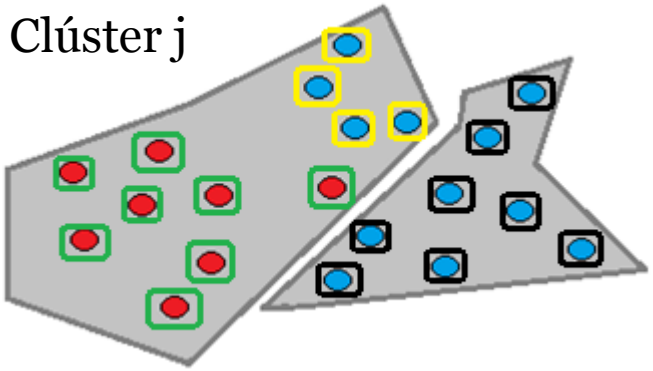


		Verdad	
		P	**N**
Hipótesis	P	**VP**	**FP**
	N	**FN**	

Métricas de Validación Externa

- El último término a aclarar es : «Verdadero Negativo» (VN). Este hace referencia a aquellos elementos que fueron ubicados correctamente fuera del clúster j, es decir, aquellos elementos ajenos al clúster en cuestión y que efectivamente no correspondían a este.

Clúster j



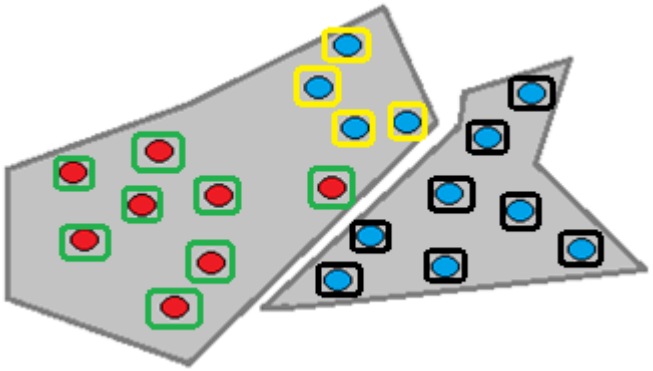
Verdad

Hipótesis	P	N
P	VP	FP
N	FN	VN

Precisión y Recall

- Con la terminología anterior aclarada es posible introducir las siguientes métricas ampliamente utilizadas y provenientes del campo de Information Retrieval: la Precisión y el Recall.

Clúster j



Verdad

Hipótesis	P	N
P	VP	FP
N	FN	VN

$$Precisión = \frac{a}{a + b}$$

$$Recall = \frac{a}{a + c}$$

Medida F

- Con los conceptos de precisión y recall es posible definir otro tipo de métrica llamada «Medida F». Esta se da en función de las dos métrica ya vistas y puede ser interpretada como la media harmónica de ambas. En particular la medida F maneja un parámetro «α» de la siguiente manera:

$$F_\alpha = \frac{1+\alpha}{\frac{1}{precisión} + \frac{\alpha}{recall}}$$

α = 1	media harmónica
α ε (0:1)	preferencia por la precisión
α > 1	preferencia por el recall

Entropía y Pureza

- Como se mencionó, la idea de la validación externa es , una vez finalizado el algoritmo de agrupación, se compara el clúster en el que fue asignado cada elemento, con la etiqueta de clase que traía de antemano (información externa).

- Para determinar que tan bueno fue el agrupamiento realizado, existen otras dos métricas externas llamadas Entropía y Pureza.

- *Sea C el conjuntos de clases en el data set D , C = (c_1, c_2,…., c_k). El algoritmo de agrupamiento produce k clústeres, que particionan a D en k distintos subconjuntos D_1, D_2,…, D_k.*

Entropía

Para cada clúster se asume la entropía como:

$$entropía(D_i) = -\sum_{j=1}^{k} Pr_i(c_j)\, log_2 Pr_i(c_j)$$

Dónde $Pr_i(c_j)$ es la proporción de puntos de la clase c_j ubicados en el clúster i o D_i. La entropía total de todo el agrupamiento (que considera todos los clústers) es:

$$entropía_{total}(D) = \sum_{i=1}^{k} \frac{|D_i|}{|D|}\ x\ entropía(D_i)$$

Pureza

Mide el hecho de que un clúster contenga solo una clase entre sus datos. La pureza de cada clúster se calcula con:

$$pureza(D_i) = max_j(Pr_i(c_j))$$

La pureza total de todo el agrupamiento (considerando todos los clusters) es :

$$pureza_{total}(D) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} \; x \; pureza(D_i)$$

Mutual Information

$$MI = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{p_i p_j}$$

Siendo $p_{ij} = \frac{n_{ij}}{n}$

$$p_i = \frac{n_i}{n}$$

$$p_j = \frac{n_j}{n}$$

Bibliografía

- LIU, Bing. *Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data*. Springer.
- IBM. *Sum of Squared Error (SSE) (cluster evaluation algorithms)*. Disponible en: http://pic.dhe.ibm.com/infocenter/spssstat/v20r0m0/index.jsp?topic=%2Fcom.ibm.spss.statistics.help%2Falg_cluster-evaluation_goodness_sse.htm
- PARDO, Mateo. *Clustering*. Nation Research Council (CNR) , Berlín Alemania, Disponible en: http://lectures.molgen.mpg.de/algsysbio10/clustering.pdf
- WEINGESSEL, Andreas. DIMITRIADOU, Evgenia. DOLNICAR, Sara. *An examination of Indexes For Determining the Number of Clusters in Binary Data Sets*. Working Paper No. 29, Enero 1999, Disponible en: http://epub.wu.ac.at/1542/1/document.pdf
- ZHAO, Qinpei. XU, Mantao. FRÄNTI Pasi. *Sum-of-Squares Based Cluster Validity Index and Significance Analysis*. Universidad de Joensuu, Finlandia. Disponible en:http://www.cs.joensuu.fi/~zhao/PAPER/ICANNGA09.pdf
- BioMed Central. *Cluster Validity Measures*. Disponible en: http://www.biomedcentral.com/content/supplementary/1471-2105-9-90-s2.pdf