Journal of Computer Science and Information Technology
June 2017, Vol. 5, No. 1, pp. 26-35
ISSN: 2334-2366 (Print), 2334-2374 (Online)
Copyright © The Author(s). All Rights Reserved.
Published by American Research Institute for Policy Development
DOI: 10.15640/jcsit.v5n1a3
URL: https://doi.org/10.15640/jcsit.v5n1a3

Data Clustering: An Approach for Evaluating the Adequate Number of Groups in Partitioned Techniques

Guillermo Molero-Castillo[1], Yaimara Céspedes-González,[2] and Alejandro Velázquez-Mena[3]

Abstract

The partitioned clustering techniques, such as k-means, have advantages in applications involving a large amount of data, but a particularity of this type of clustering is to establish a priori the number of input groups (k). So in practice, it is necessary to repeat the test by establishing different numbers of groups, choosing the solution that best suits the objective of the problem. Therefore, to validate the results obtained it is necessary to have validation mechanisms that allow evaluating the formation of the groups appropriately. An evaluation strategy is through validation indexes that help determine if the formation of the groups is adequate. These methods are based on estimates that identify how compact or separate the formed groups are. This paper presents validation indexes used as a strategy to determine the number of relevant groups. The results obtained indicate that this evaluation approach guarantees an adequate way the determination of the desired number of groups.

Keywords: Clustering, data mining, k-means, groups number, validation indexes.

1. Introduction

A current reality of data mining is its role as a supportive technology that can solve two major challenges: a) work with data sets to extract and discover information of interest, and b) use appropriate techniques to analyze, understand and identify trends and behaviors that facilitate a better understanding of the phenomena that surround us and help us in the decision-making process (Molero, 2008; Molero, 2014).

[1] CONACYT, Mexico, Universidad Veracruzana, Veracruz, Mexico, ggmoleroca@conacyt.mx
[2] Universidad Veracruzana, Veracruz, Mexico, yaimara.cespedes@gmail.com
[3] 3Universidad Nacional Autónoma de México, Mexico, mena@fi-b.unam.mx

One of the tasks of data mining and pattern recognition to construct models of knowledge extraction is clustering, whose objective is to evaluate similarities between the data to represent them in a few groups, that is, a heterogeneous population of data is divided into a number of homogeneous subgroups according to the similarities of their records (Berry and Linoff, 2004; Sumathi and Sivanandam, 2006).

Deciding the number of groups or partitions in which a data set should be divided is an important problem to be faced when working with clusters (Larose, 2005). In some cases, the obtained groups, after applying some algorithm of clustering, not represent the real structure that the data source owns. For this reason, it is necessary to have quantitative measures to evaluate the formation of groups.

This article presents the clustering as one of the significant tasks of data mining, which is addressed with the aim of publicizing the importance of the evaluation of groups obtained by partitional techniques, such as k-means. Validation indexes were used as a strategic method to evaluate if the formation of groups is the most appropriate. As a case of study, we used a set of clinical data generated from oncological variables of breast cancer, such as diagnosis, area, radius, texture, perimeter, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. The k-means partitional technique was used, and three types of validation indexes were applied: Silhouette (Rousseeuw, 1987), Dunn (Dunn, 1974) and Davies-Bouldin (Davies and Bouldin, 1979).

2. Background

2.1 Clustering

The clustering is an effective method to extract useful knowledge, which allows dividing a heterogeneous population of data into a few homogeneous subgroups according to the similarity of their records (Berry and Linoff, 2004; Sumathi and Sivanandam, 2006). A subgroup consists of one or more data vectors, which in turn comprise several variables (Molero, 2008). In the clustering, two main types stand out (Hand *et al*., 2001; Berry and Linoff, 2004; Larose, 2005; Witten and Frank, 2005): a) hierarchical and b) partitional. Hierarchical clustering is characterized by the recursive development of a tree-like structure. This type of clustering is divided into agglomerative or divisive (Larose, 2005). The agglomerative method begins with each element forming an independent group. In subsequent steps, the two the nearest groups are added to a new group, each time larger. In this way, the process continues until all elements are part of a single group. The divisive method considers all elements grouped into a single set and according to each iteration are divided into smaller and smaller independent subsets.

Some algorithms of hierarchical clustering are: Twostep, Cobweb, Birch (Balanced iterative reducing and clustering using hierarchical), Cure (Clustering using representatives), Rock (Robust clustering algorithm using links), Chameleon, among others.

Partitional clustering organizes the elements into k groups. That is, it determines the number of partitions by an iterative procedure that optimizes the local or global structure of the pooled data (Vazirgiannis *et al*., 2003). Partitional methods have advantages in applications involving a large amount of data for which the construction of a tree is complicated (Witten and Frank, 2005). The problem of the partitional techniques is the decision of the desired number of output groups, so in practice it is necessary to repeat the test considering a different number of groups, choosing the solution that best suits the objective of the problem (Jain *et al*., 1999). Some algorithms within this type of clustering are k-means, k-medians, k-mode, Pam (Partitioning around medoids), Clara (Clustering large applications) and Clarams (Clustering large applications based on randomized search), among others.

Since clustering is an effective method for extracting useful knowledge, it uses algorithms that allow finding subgroups of data within a larger set of available data, maximizing the similarity of elements within groups, such as k-means, which is one of the best known clustering techniques used in data mining (Chou*et al*., 2003; Brock *et al*., 2011).

2.2 K-means

K-means is a partitional technique proposed by J. B. MacQueen in 1967 (Berry and Linoff, 2004). A characteristic of this type of clustering is to establish a priori the number of input groups (k), so in practice is necessary to repeat the test considering different groups numbers, until obtaining the solution that best suits the objective of the problem. The k-means procedure is as follows (Jain *et al*., 1999; Larose, 2005):

1. Randomly select k points or elements, making them represent the "centers" of groups.
2. Assign each of the remaining elements to the nearest center. This is the minimum distance between the element and the center. Usually, the distance measure used is Euclidean.
3. Once all elements have been assigned, the k centers are recalculated.
4. Repeat steps 2 and 3 until the centers are no longer modified.

In order to assign the records to the groups, whose center is the closest, we use the quadratic euclidean distance defined as (Clementine, 2006):

$$d_{ij} = \|X_i - C_j\|^2 = \sum_{q=1}^{Q}(x_{qi} - c_{qj})^2$$

where,

X_i: vector of the input variables for the record*i*

C_j: group center for region *j*

Q: number of input variables

x_{qi}: value of the *q-th* input variable for the *i-th* record

c_{qj}: value of the *q-th* input variable for the *j-th* record

To update the value of the centers in the groups, these are calculated as the average vector of the records established in that group: $C_j=X_j$, where the fields of the mean vector X_j are calculated according to the following equation:

$$\bar{x}_{qj} = \frac{\sum_{i=1}^{n_j} x_{qi}(j)}{n_j}$$

where,

n_j: is the number of records in the group *j*

$x_{qi}(j)$: is the *q-th* value for record*i* that is assigned to group *j*

With k-means it is expected to obtain results that reveal patterns of the data set, this is, the groups are formed with elements having similar characteristics.

3. Methodology

A qualitative and quantitative approach was used to conduct this study. As a method of evaluating the desired number of groups, validation indices were used, which are quantitative indicators that allow us to evaluate whether the formation of groups or partitions, obtained by partitional techniques, as k-means, is the most appropriate; representing the actual structure that the data source has. These indexes are Silhouette, Dunn, and Davies-Bouldin, which are based on estimates that identify how compact or separate the formed groups are (Chou *et al*., 2003; Brock *et al*., 2011).

3.1 Silhouette (Rousseeuw, 1987)

This index is used to estimate the desired number of groups, as well as to assess the assignment of records in the established groups (Brock *et al*., 2011). To estimate the desired number of groups the partition (k) is taken with the highest average, while to evaluate the assignment of records is calculated *s*(*i*) for the *i-th*record defined as (Bolshakova and Azuaje, 2003):

$$s(i) = \frac{b_i - a_i}{\max(a_i, b_i)}$$

where,

a_i: average distance between the *i-th* record and all others that are in the same group.

b_i: minimum average distance between the *i-th* record and the records that are in other groups.

The value *s*(*i*) is located in the interval -1 and 1. If *s*(*i*) is close to 1 it can be inferred that the *i-th* record was assigned to an appropriate group, if *s*(*i*) approaches zero indicates that the *i-th* record could be assigned to another nearest group, and if *s*(*i*) is close to -1 it can be inferred that the *i-th* record was poorly grouped.

3.2 Dunn (Dunn, 1974)

This index indicates whether the formed groups are well compacted and separated. To estimate the desired number of groups, this indicator maximizes intergroup distance and minimizes intragroup distance (Saitta *et al*., 2007). Given a cluster partition, where C_i represents the *i-th* partition group, the Dunn index is defined as (Kovács *et al*., 2005):

$$D = \min_{1\le i\le n}\left\{\min_{1\le j\le n, j\ne i}\left(\frac{d(C_i,C_j)}{\max\limits_{1\le k\le n} diam(C_k)}\right)\right\}$$

where,

d(Ci,C_j): distance between groups $C_i$and C_j (intergroup distance)

diam(C_k): distance or intragroup diameter of the group C_k

The optimal number of groups is one that maximizes D.

3.3 Davies-Bouldin (Davies and Bouldin, 1979)

The Davies-Bouldin index (DB) estimates the desired number of groups through a measure of dispersion and dissimilarity of the established groups (Halkidi *et al*., 2002). Like the Dunn index, this index reveals whether the groups formed are compact and well separated (Bolshakova and Azuaje, 2003). The Davies-Bouldin index is defined as (Boutin and Hascoêt, 2004):

$$DB = \frac{1}{K}\sum_{i=1}^{K}\max_{i\ne j}\left\{\frac{diam(C_i) + diam(C_j)}{d(C_i,C_j)}\right\}$$

where,

diam(C_i): intragroup distance of the group C_i

diam(C_j): intragroup distance of the group C_j

d(C_i,C_j): distance between groups C_i and C_j (intergroup distance)

The configuration that minimizes DB is taken as the desired number of groups. At present, this index is considered as one of the best validation indexes by its better approximation of the desired number of groups (Saitta *et al*., 2007).

3.4 Data source

The data source from which the clustering process was performed corresponds to clinical studies of Wisconsin Diagnostic Breast Cancer (WDBC), compiled and reviewed in November 1995 by W. H. Wolberg, W. N. Street and O. L. Mangasarian of the University of Wisconsin and Madison Hospital (Lichman, 2013). The original database was compiled between January 1991 and November 1994. Clinical records are derived from digitized images. The total number of records used in this study was 50. Table 1 shows the oenological variables that are part of the available clinical cases.

Table 1. Available oncology variables

Variable	Description	Type
ID number	Identifies the patient	Discrete
Diagnosis	It is the diagnosis (M=malignant, B=benign)	Discrete
Radius	Average distances of the center and points of the perimeter	Continuous
Texture	Standard deviation of gray-scale	Continuous
Perimeter	Value of breast cancer perimeter	Continuous
Area	Value of breast cancer area	Continuous
Smoothness	Variation of the radius length	Continuous
Compactness	Perimeter ^ 2 / Area–1	Continuous
Concavity	Fall or severity of the contours	Continuous
Concave points	Number of concave contour sectors	Continuous
Symmetry	Symmetry of the image	Continuous
Fractal dimension	Border approach–1	Continuous

As an identifier of the data source, the ID number field was chosen as the reference that uniquely identifies each of the clinical cases evaluated in this study.

4. Results

For the validation process, k-means was used with different input configurations (k), that is, seven clustering were executed (k = 2, 3, ..., 8). Table 2 shows the results of the groups obtained, where the labels 1, 2, 3, 4, 5, 6, 7 and 8 represent the membership of the clinical case of breast cancer (ID number) at corresponding group, and II, III, IV, V, VI, VII and VIII correspond to the seven clusters defined as entry in k-means.

Table 2. Clustering obtained by k-means

No.	ID number	II	III	IV	V	VI	VII	VIII
1	P842302	1	1	1	1	1	1	1
2	P842517	1	3	3	3	3	3	3
3	P84300903	1	3	3	3	3	7	7
4	P84348301	1	1	1	4	6	6	6
5	P84358402	1	3	3	3	3	3	3
6	P843786	1	1	1	4	4	4	4
7	P844359	1	3	3	3	3	3	3
8	P84458202	1	1	4	4	4	4	4
9	P844981	1	1	1	4	4	4	4
10	P84501001	1	1	1	4	4	4	4
11	P845636	2	3	4	5	5	5	5
12	P84610002	1	3	4	5	5	5	5
13	P846226	1	1	3	1	1	7	7
14	P846381	2	3	4	5	5	5	5
15	P84667401	1	1	1	4	4	4	4
16	P84799002	1	1	4	4	4	4	4
17	P848406	2	3	4	5	5	5	5
18	P84862001	1	1	1	4	4	4	4
19	P849014	1	3	3	3	3	3	3
20	**P8510426**	**2**	**2**	**2**	**2**	**2**	**2**	**8**
21	**P8510653**	**2**	**2**	**2**	**2**	**2**	**2**	**8**
22	**P8510824**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
23	P8511133	1	1	1	1	1	1	1
24	P851509	1	3	3	3	3	3	3
25	P852552	1	3	3	3	3	7	7
26	P852631	1	1	1	1	1	1	1
27	P852763	1	1	4	4	4	4	4
28	P852781	1	3	3	3	3	3	3
29	P852973	1	1	4	4	4	4	4
30	P853201	1	3	3	3	3	3	3
31	P853401	1	3	3	3	3	7	7
32	P853612	2	1	4	4	4	4	4
33	P85382601	1	1	3	3	3	7	7
34	P854002	1	3	3	3	3	7	7
35	P854039	1	3	4	3	4	4	4
36	P854253	1	3	4	5	5	5	5
37	P854268	2	3	4	5	5	5	5
38	**P854941**	**2**	**2**	**2**	**2**	**2**	**2**	**8**
39	P855133	2	3	4	5	5	5	5
40	P855138	2	3	4	5	5	5	5
41	P855167	2	3	4	5	5	5	5
42	P855563	2	1	4	4	4	4	4

Table 2. Clustering obtained by k-means

43	P855625	1	3	3	3	3	7	7
44	P856106	2	1	4	4	4	4	4
45	P85638502	2	3	4	5	5	5	5
46	P857010	1	3	3	3	3	3	3
47	**P85713702**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
48	P85715	1	1	4	4	4	4	4
49	**P857155**	**2**	**2**	**2**	**2**	**2**	**2**	**8**
50	**P857156**	**2**	**2**	**2**	**2**	**2**	**2**	**8**

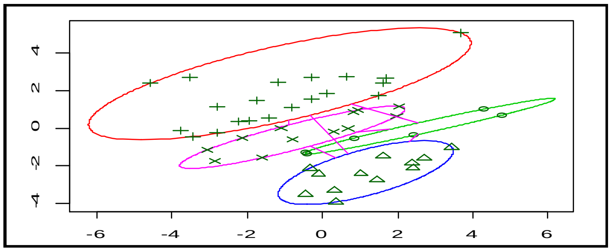
The validation indexes, Silhouette, Dunn and Davies-Bouldin, were applied to the seven clusters obtained by the k-means. The following results were obtained (Table 3):

Table 3. Desired number of groups through validation indexes

Validation indexes	Number of groups						
	k=II	k=III	k=IV	k=V	k=VI	k=VII	k=VIII
Silhouette	0.28	0.34	**0.44**	0.40	0.29	0.24	0.22
Dunn	0.32	0.37	**0.42**	**0.43**	0.33	0.31	0.21
Davies-Bouldin	1.22	1.18	**1.04**	1.08	1.12	1.16	1.23

It was observed that the desired number of groups suggested by the validation indices is 4, Silhouette (the highest value = 0.44) and Davies-Bouldin (the smaller value = 1.04). In the case of the Dunn validation index (the highest value = 0.43), this indicates that the desired number of groups is 5, but the next cluster approaching the other two indices –Silhouette and Davies-Bouldin– is also 4.Thus, the group that meets the validation indexes is 4 –Silhouette (0.44), Dunn (0.42) and Davies-Bouldin (1.04)–. This validates the four groups of clinical studies of patients with breast cancer (Figure 1). A significant factor reinforcing this validation is mainly based on the Davies-Bouldin index, which is one of the most recognized indexes for its best approximation.

Figura 1. Formación de los cuatro grupos de casos de cáncer de mama



In general, based on the results (Table 2) and the study variables Area, Radius, Perimeter and Texture of the four groups obtained the following were observed:

- Group 1 (pink) presents 9 cases of malignant breast cancer with an average Perimeter of 98 pixels, and average Area of 650 pixels, which is the number of pixels inside the cancerous nucleus, including the edges. The tumor size in this group of patients is significantly large.
- Group 2 (green) corresponds to 7 clinical cases of benign breast cancer with an average Area of 442 pixels and average Perimeter of 76 pixels. This is the only group of patients with benign breast cancer.
- Group 3 (blue) comprises 15 clinical cases of malignant breast cancer with an average Area of 1129 pixels and average Perimeter of 126 pixels. In this group, we have patients with a large tumor size, compared to patients in other groups.
- Group 4 (red) has 19 cases of malignant breast cancer with an average Area of 640 pixels and average Perimeter of 94 pixels. From this, it can be inferred that the tumor size in this group of patients is moderately large.

  In each of the four groups, it was observed that the clinical cases of the patients share similar characteristics in size (area and perimeter) and type of disease (benign and malignant).

5. Conclusions

The validation indices, Silhouette, Dunn and Davies-Bouldin (this one recognized by their best approximation) were shown to be useful in determining the desired number of groups. The study was focused on the type partitional technique k-means. The Silhouette(0.44), Dunn (0.42) and Davies-Bouldin (1.04) indexes made it possible to determine that the case of study, over breast cancer, it can be divided into four clinically similar groups. The obtained results indicate that this evaluation approach, through validation indexes, guarantees adequate and with a high degree of reliability the obtaining of the desired number of groups. This work involved the analysis of clinical data, the management of a clustering technique to identify similar clinical cases of bed cancer, and the use of validation indexes, allowing to extend the vision of the data mining and its application to problems of diverse nature, in this case, applied to Health.

6. References

Berry, M. and Linoff, G. (2004). Data Mining Techniques: for marketing, sales, and customer relationship management. Indiana, United States: Wiley Publishing.

Bolshakova, N. and Azuaje, F. (2003). Improving expression data mining through cluster validation. Proceedings of the 4th International EMBS Special Topic Conference on Information Technology Applications in Biomedicine(pp. 19-22). Ireland: IEEE.

Boutin, F. and Hascoêt, M. (2004). Cluster validity indices for graph partitioning. Proceedings of the 8th International Conference on Information Visualization, (pp. 376-381). Washington, United States: IEEE.

Brock, G., Pihur, V., Susmita, D. and Somnath, D. (2011). clValid, an R package for cluster validation. Department of Bioinformatics and Biostatistics, University of Louisville, pp. 32, United States.

Chou, C., Su, M. and Lai, E. (2003). A new cluster validity measure for clusters with different densities. Proceedings of the International Conference on Intelligent Systems and Control (pp. 276-281).

Clementine (2006). Clementine 10.1 Algorithms Guide. (User manual, Integral Solutions Limited).United States: SPSS.

Davies, D. and Bouldin, D. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224-227.

Dunn, J. (1974). Well-separated clusters and optimal fuzzy partitions.Journal of Cybernetics, 4(1), 95–104.

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

Hand, D., Mannila, H. and Smyth, P. (2001). Principles of Data Mining. Massachusetts, United States: The Massachusetts Institute of Technology Press.

Jain, A., Murty, M. and Flynn, P. (1999). Data Clustering: A Review. ACM Computing Surveys, 31(3), 264-323.

Larose, D. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. New Jersey, United States: John Wiley & Sons.

Molero, G. (2008). Desarrollo de un modelo basado en técnicas de minería de datos para clasificar zonas climatológicamente similares en el estado de Michoacán (Master's thesis). National Autonomous University of Mexico, Mexico.

Molero, G. (2014). Clasificador bayesiano para el pronóstico de la supervivencia y mortalidad de casos de cáncer de mama en mujeres de origen hispano (Doctoral dissertation). University of Guadalajara, Mexico.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

Saitta, S., Raphael, B. and Smith, I. (2007). A bounded index for cluster validity. In P. Perner (Eds.), Machine Learning and Data Mining in Pattern Recognition, 174-187.

Sumathi, S. and Sivanandam, S. (2006). Introduction to Data Mining and its Applications. Studies in Computational Intelligence, 29. Heidelberg, Germany: Springer.

Vazirgiannis, M., Halkini, M. and Gunopulos, D. (2003). Uncertainty handling and quality assessment in data mining. Advanced Information and Knowledge Processing, Heidelberg, Germany: Springer.

Lichman, M. (2013). Breast Cancer Wisconsin Data Set. UCI Machine Learning Repository, Center for Machine Learning and Intelligent Systems. Available: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Witten, I. and Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. United States: Morgan Kaufmann.