International Journal on Recent Researches In Science, Engineering & Technology

ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105

Volume 4, Issue 6, June 2016.

JIR IF : 2.54
DIIF IF : 1.46
SJIF IF : 1.329

A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy. It is an absolutely free (No processing charges, No publishing charges etc) Journal Indexed in JIR, DIIF and SJIF.

Research Paper

Available online at: www.jrrset.com

Estimation of sufficient number of Groups in Partitioned Techniques using Data Clustering Approach

Dr.M.EDWIN JAYASINGH

Department of Computer Science and Engineering,Shadan College of Engineering and Technology HYD,T.S,INDIA

Abstract - The partitioned clustering techniques, such as k-means, have advantages in applications involving a large amount of data, however a particularity of this kind of clustering is to set up a priori the number of enter groups (k). So in practice, it is vital to repeat the test with the aid of organising one of a kind numbers of groups, deciding on the solution that exceptional fits the objective of the problem. Therefore, to validate the effects received it is crucial to have validation mechanisms that allow evaluating the formation of the agencies appropriately. An evaluation strategy is thru validation indexes that assist decide if the formation of the groups is adequate. These methods are based on estimates that pick out how compact or separate the fashioned businesses are. This paper offers validation indexes used as a approach to decide the wide variety of relevant groups. The effects got point out that this assessment strategy ensures an adequate way the dedication of the preferred quantity of groups.

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

1. Introduction

A modern-day reality of data mining is its role as a supportive technological know-how that can remedy two primary challenges: a) work with facts sets to extract and find out facts of interest, and b) use appropriate methods to analyze, apprehend and perceive developments and behaviors that facilitate a better understanding of the phenomena that encompass us and assist us in the decision-making procedure (Molero, 2008; Molero, 2014). One of the tasks of data mining and pattern cognizance to assemble models of expertise extraction is clustering, whose goal is to consider similarities between the records to signify them in a few groups, that is, a heterogeneous populace of statistics is divided into a quantity of homogeneous subgroups according to the similarities of their documents (Berry and Linoff, 2004; Sumathi and Sivanandam, 2006). Deciding the quantity of agencies or partitions in which a data set need to be divided is an vital problem to be faced when working with clusters (Larose, 2005). In some cases, the received groups, after making use of some algorithm of clustering, not represent the actual structure that the data supply owns. For this reason, it is fundamental to have quantitative measures to consider the formation of groups. This article offers the clustering as one of the tremendous duties of information mining, which is addressed with the aim of publicizing the importance of the evaluation of companies acquired through partitional techniques, such as k-means. Validation indexes had been used as a strategic approach to evaluate if the formation

of agencies is the most appropriate. As a case of study, we used a set of clinical records 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 fantastic approach to extract useful knowledge, which allows dividing a heterogeneous populace of facts into a few homogeneous subgroups in accordance to the similarity of their archives (Berry and Linoff, 2004; Sumathi and Sivanandam, 2006). A subgroup consists of one or greater data vectors, which in flip contain numerous variables (Molero, 2008). In the clustering, two predominant kinds stand out (Hand et al., 2001; Berry and Linoff, 2004; Larose, 2005; Witten and Frank, 2005): a) hierarchical and b) partitional. Hierarchical clustering is characterised by means of the recursive development of a tree-like structure. This type of clustering is divided into agglomerative or divisive (Larose, 2005). The agglomerative technique begins with each element forming an impartial group. In subsequent steps, the two the nearest agencies are brought to a new group, every time larger. In this way, the technique continues until all factors are part of a single group. The divisive approach considers all factors grouped into a single set and in accordance to each new release are divided into smaller and smaller impartial subsets. Some algorithms of hierarchical clustering are: Twostep, Cobweb, Birch (Balanced iterative decreasing and clustering using hierarchical), Cure (Clustering using representatives), Rock (Robust clustering algorithm the use of links), Chameleon, amongst others. Partitional clustering organizes the elements into k groups. That is, it determines the variety of partitions by means of an iterative process that optimizes the local or international shape of the pooled facts (Vazirgiannis et al., 2003). Partitional methods have blessings in applications involving a large amount of records for which the construction of a tree is difficult (Witten and Frank, 2005). The trouble of the partitional strategies is the selection of the favored wide variety of output groups, so in exercise it is vital to repeat the test considering a exceptional variety of groups, selecting the answer that nice fits the goal of the trouble (Jain et al., 1999). Some algorithms within this type of clustering are k-means, k-medians, k-mode, Pam (Partitioning round medoids), Clara (Clustering massive applications) and Clarams (Clustering giant purposes primarily based on randomized search), among others. Since clustering is an tremendous approach for extracting beneficial knowledge, it makes use of algorithms that allow finding subgroups of information within a larger set of accessible data, maximizing the similarity of elements within groups, such as k-means, which is one of the exceptional known clustering strategies used in information mining (Chouet al., 2003; Brock et al., 2011).

2.2 K-means

K-means is a partitional technique proposed by means of skill of J. B. MacQueen in 1967 (Berry and Linoff, 2004). An attribute of this kind of clustering is to establish a priori the vary of entering companies (k), so in exercise is fundamental to repeat the check wondering about one of a kind organizations numbers, until obtaining the answer that well fits the purpose of the problem. The k-means method is as follows (Jain et al., 1999; Larose, 2005): Randomly pick ok points or elements, making them signify the "centers" of groups.

1. Assign each of the final factors to the nearest center. This is the minimum distance between the issue and the center. Usually, the distance measure used is Euclidean.
2. Once all elements have been assigned, the k centers are recalculated.

3. Repeat steps 2 and three until the facilities are no longer modified. In order to assign the data to the groups, whose middle 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 keep posted the value of the centers in the groups these are deliberated as the average vector of the records recognized in that group: $C_j=X_j$, where the fields of the mean vector Xj 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 allocated to group j With k-means it is predictable to obtain results that disclose patterns of the data set, this is, the groups are formed with elements having alike characteristics.

3. Methodology

A qualitative and quantitative approach was used to conduct this study. As a method of evaluating the preferred 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 approximate the preferred number of groups, as well as to evaluate the assignment of records in the recognized groups (Brock et al., 2011). To estimate the preferred number of groups the partition (k) is taken with the uppermost average, while to assess the assignment of records is intended s(i) for the i-threcord 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 point out whether the shaped groups are well compressed and alienated. To estimate the preferred number of groups, this pointer maximizes intergroup remoteness and reduces intragroup distance (Saitta et al., 2007). Given a cluster partition, where Ci represents the i-th partition group, the Dunn index is defined as (Kovács et al., 2005):

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

where,
$d(Ci,Cj)$: distance between groups Ci and Cj (intergroup distance)
$diam(Ck)$: distance or intragroup diameter of the group Ck
The optimal number of groups is one that maximizes D.

3.3 Davies-Bouldin (Davies and Bouldin, 1979)

The Davies-Bouldin index (DB) approximates the preferred number of groups through a measure of dispersion and difference of the recognized 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,
$diam(Ci)$: intragroup distance of the group Ci
$diam(Cj)$: intragroup distance of the group Cj
$d(Ci,Cj)$: distance between groups Ci and Cj (intergroup distance)

The configuration that minimizes DB is taken as the favored number of groups. At present, this index is regarded as one of the nice validation indexes by using its better approximation of the desired range of agencies (Saitta et al., 2007).

3.4 Data source

The facts source from which the clustering method was carried out corresponds to medical studies of Wisconsin Diagnostic Breast Cancer (WDBC), compiled and reviewed in November 1995 by means of W. H. Wolberg, W. N. Street and O. L. Mangasarian of the University of Wisconsin and Madison Hospital (Nilashi, Mehrbakhsh, 2013). The unique database used to be compiled between January 1991 and November 1994. Clinical data are derived from digitized images. The complete variety of records used in this find out about was once 50. Table 1 shows the oenological variables that are section of the reachable 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 facts source, the ID range field used to be chosen as the reference that uniquely identifies every of the clinical cases evaluated in this study.

4. Results

For the validation process, k-means was once used with distinctive enter configurations (k), that is, seven clustering had been done (k = 2, 3, ..., 8). Table 2 indicates the consequences of the groups obtained, the place the labels 1, 2, 3, 4, 5, 6, 7 and eight characterize the membership of the scientific case of breast cancer (ID number) at corresponding group, and II, III, IV, V, VI, VII and VIII correspond to the seven clusters described 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	2	1	4	4	4	4	4
33	P85382601	1	1	3	3	3	7	7
34	P854002 1	1	1	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	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
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**

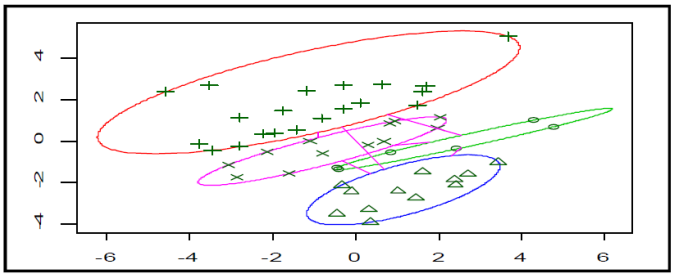
The corroboration indexes, Silhouette, Dunn and Davies-Bouldin, were functional to the seven clusters acquired by the k-means. The following results were obtained Table 3:

Table 3. Preferred 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 used to be observed that the preferred range of corporations advised with the aid of the validation indices is 4, Silhouette (the absolute best fee = 0.44) and Davies-Bouldin (the smaller value = 1.04). In the case of the Dunn validation index (the very best cost = 0.43), this indicates that the preferred wide variety of businesses is 5, but the subsequent cluster drawing near the different two indices – Silhouette and Davies-Bouldin– is also 4.Thus, the crew that meets the validation indexes is 4 – Silhouette (0.44), Dunn (0.42) and Davies-Bouldin (1.04)–. This validates the 4 agencies of clinical studies of sufferers with breast most cancers (Figure 1). A widespread component reinforcing this validation is in general primarily based on the Davies-Bouldin index, which is one of the most diagnosed indexes for its satisfactory approximation.

Fig 1. Training of the Four Groups of Breast Cancer Cases



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

1. Group 1 (pink) presents 9 instances of malignant breast most cancers with an average Perimeter of 98 pixels, and average Area of 650 pixels, which is the number of pixels inner the cancerous nucleus, inclusive of the edges. The tumor size in this group of patients is considerably large.
2. Group 2 (green) corresponds to 7 clinical cases of benign breast most cancers with an common Area of 442 pixels and average Perimeter of seventy six pixels. This is the solely group of sufferers with benign breast cancer.

3. Group 3 (blue) comprises 15 medical instances of malignant breast cancer with an common Area of 1129 pixels and average Perimeter of 126 pixels. In this group, we have patients with a giant tumor size, in contrast to sufferers in different groups.
4. Group four (red) has 19 instances of malignant breast cancer with an average Area of 640 pixels and common Perimeter of ninety four pixels. From this, it can be inferred that the tumor measurement in this group of patients is reasonably large. In each of the four groups, it was once located that the clinical cases of the patients share comparable characteristics in dimension (area and perimeter) and type of disorder (benign and malignant).

5. Conclusions

The validation indices, Silhouette, Dunn and Davies-Bouldin (this one diagnosed by their high-quality approximation) had been proven to be beneficial in deciding the preferred wide variety of groups. The find out about was centered on the type partitional technique kmeans. The Silhouette(0.44), Dunn (0.42) and Davies-Bouldin (1.04) indexes made it feasible to decide that the case of study, over breast cancer, it can be divided into four clinically similar groups. The got results point out that this assessment approach, via validation indexes, guarantees sufficient and with a high degree of reliability the acquiring of the preferred range of groups. This work worried the analysis of medical data, the management of a clustering technique to become aware of similar medical cases of mattress cancer, and the use of validation indexes, permitting extending the vision of the data mining and its application to problems of numerous nature, in this case, applied to Health.

REFERENCES

[1] Linoff, Gordon S., and Michael JA Berry. Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons, 2011.
[2] Bolshakova, N., and F. Azuaje. "Improving expression data mining through cluster validation." Information Technology Applications in Biomedicine, 2003. 4th International IEEE EMBS Special Topic Conference on. IEEE, 2003.
[3] Boutin, Francois, and Mountaz Hascoët. "Cluster validity indices for graph partitioning." Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on. IEEE, 2004.
[4] Brock, Guy, et al. "clValid, an R package for cluster validation." Journal of Statistical Software (Brock et al., March 2008) (2011).
[5] Chou, Chien-Hsing, Mu-Chun Su, and Eugene Lai. "A new cluster validity measure for clusters with different densities." IASTED International Conference on Intelligent Systems and Control. 2003.
[6] SPSS, Clementine. "10.1, Algorithms Guide." Integral Solutions Limited, USA (2006).
[7] Davies, David L., and Donald W. Bouldin. "A cluster separation measure." IEEE transactions on pattern analysis and machine intelligence 2 (1979): 224-227.
[8] Dunn, Joseph C. "Well-separated clusters and optimal fuzzy partitions." Journal of cybernetics 4.1 (1974): 95-104.
[9] Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. "Clustering validity checking methods: part II." ACM Sigmod Record 31.3 (2002): 19-27.
[10] Koller, Daphne, et al. Introduction to statistical relational learning. MIT Press, 2007.
[11] Larose, Daniel T., and Chantal D. Larose. Discovering knowledge in data: an introduction to data mining. John Wiley & Sons, 2014.
[12] Molero, G. "Development of a model based on data mining techniques to classify climatologically similar zones in the state of Michoacán." MS, National University Atunóma of Mexico, Mexico (2008).

[13] Molero-Castillo, Guillermo, Yaimara Céspedes-González, and Alejandro Velázquez-Mena. "Data Clustering: An Approach for Evaluating the Adequate Number of Groups in Partitioned Techniques." Journal of Computer Science 5.1 (2017): 26-35.
[14] Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics 20 (1987): 53-65.
[15] Saitta, Sandro, Benny Raphael, and Ian FC Smith. "A bounded index for cluster validity." International Workshop on Machine Learning and Data Mining in Pattern Recognition. Springer, Berlin, Heidelberg, 2007.
[16] Sumathi, S., and S. N. Sivanandam. "Data mining in customer value and customer relationship management." Introduction to Data Mining and its Applications (2006): 321-386.
[17] Vazirgiannis, Michalis, Maria Halkidi, and Dimitrious Gunopulos. Uncertainty handling and quality assessment in data mining. Springer Science & Business Media, 2003.
[18] Nilashi, Mehrbakhsh, et al. "A knowledge-based system for breast cancer classification using fuzzy logic method." Telematics and Informatics 34.4 (2017): 133-144.
[19] Witten, Ian H., et al. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.