IJCERT

International Journal of Computer Engineering in Research Trends

Multidisciplinary, Open Access, Peer-Reviewed and fully refereed

Review Paper Volume-5, Issue-12 ,2018 Regular Edition E-ISSN: 2349-7084

A Review of Clustering and Clustering Quality Measurement

S. U. Patil1*, U. A. Nuli[2]

[1]Computer Science and Engineering department, M.Tech, Textile and Engineering Institute, Ichalkaranji, India
E-mail: shreyaupatil43@gmail.com, uanulil@yahoo.com

Available online at: http://www.ijcert.org

Received: 30/12/2018, Revised: 07/01./2019, Accepted: 10/01//2019, Published: 15/01 /2019

Abstract: - This paper presents a comparative study on clustering methods and developments made at various times. Clustering is defined as unsupervised learning where the objects are grouped on the basis of some similarity inherent among them. There are different methods for clustering objects such as hierarchical, partitioned, grid, density based and model-based. Many algorithms exist that can solve the problem of clustering, but most of them are very sensitive to their input parameters. Therefore it is essential to evaluate the result of the clustering algorithm. It is difficult to define whether a clustering result is acceptable or not; thus several clustering validity techniques and indices have been developed. Cluster validity indices are used for measuring the goodness of a clustering result comparing to other ones which were created by other clustering algorithms, or by the same algorithms but using different parameter values. The results of a clustering algorithm on the same data set can vary as the input parameters of an algorithm can extremely modify the behaviour and execution of the algorithm the intention of this paper is to describe the clustering process with an overview of different clustering methods and analysis of clustering validity indices.

Keywords: Cluster, Validity Index, Supervised, Data mining.

1. Introduction

The data mining technique refers to extraction of dedicated Particular statistical distributions. Clustering can, therefore, be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings information from a large, bulky and robust data set and transform it into an understandable structure for further supplementary use. It can also be termed as "knowledge discovery in databases". And in basic terminology, it can be called as classification. In classification, the guided classification refers to when the classes of the object are given in advance. Whereas the unguided classification refers to when the class label is not attached to an object in advance. This unguided classification is commonly known as Clustering.

These problems cannot be solved by one specific algorithm, but it needs various algorithms that differ significantly in their notion of what makes a cluster and how to depend on the individual data set and intended use of the results. Clustering is considered to be more difficult than supervised classification as there is no label attached to the patterns in clustering. The given label in the case of supervised classification becomes a clue to grouping data objects as a whole. Whereas in the case of clustering, it becomes difficult to decide, to which group a pattern will belong to in the absence of a label. One of the most important issues in cluster analysis is the evaluation of clustering results to find the partitioning that best fits the underlying data. This is the main subject of cluster validity. Cluster validity indices are used for measuring the goodness of a clustering result comparing to other ones which were

https://doi.org/10.22362/ijcert/2018/v5/i12/v5i1205

created to efficiently find the area of the data space, intervals or by other clustering algorithms, or by the same algorithms but using different parameter values. The results of a clustering algorithm on the same data set can vary as the input parameters of an algorithm can extremely modify the behaviour and execution of the algorithm.

2. Clustering Techniques

Here we will discuss various clustering approaches with basic techniques. As we have multiple techniques, multiple approaches are there. This is because there is no such precise definition for "cluster". That is why different clustering approaches have been proposed, each of which uses a different inclusion principle. In one of the related work [1], it is suggested dividing the clustering approaches into two distinct groups that are hierarchical and partitioning techniques. In [2] proposed the following three additional categories for applying clustering techniques: density, model and grid-based methods.

2.1. Hierarchical clustering methods

It is also known as connectivity based clustering. It is based on the idea of objects being more related to nearby objects than to objects farther away. Hierarchical clustering algorithms connect objects in clusters by their distance. A cluster can be described mainly by the maximum length needed to connect parts of the cluster. At different distances, different clusters will form [3].

Connectivity-based clustering is a family of methods that differ by the way distances are computed. It is based on the choice of distance functions. The hierarchical clustering can agglomerative or divisive. Hierarchical clustering techniques use various criteria to dedicate at each step which clusters should be joined as well as where the cluster should be partitioned into different clusters. It is based on measure of cluster proximity. There is three measure of cluster proximity: single-link, complete-link and common link.

2.1.1 Single-linkage clustering

This type of clustering is called the connectedness, the minimum method or the nearest neighbour method. In single-linkage clustering, the link between two clusters is made by a single element pair, namely those two elements (one in each cluster) that are closest to each other.

2.1.2. Complete-linkage clustering

In complete-linkage clustering also called the diameter, the maximum method or the furthest neighbour method, the distance between two clusters is determined by the longest distance from any member of one cluster to any member of the other cluster [2].

2.1.3Average-linkage clustering

In average linkage clustering also known as minimum variance method, the distance between two clusters is determined by the mapping of complete linkage clustering. average distance from any member of one cluster to any member of the other cluster[2].

2.2 Partition clustering methods

Partition clustering algorithm separates the data points into number of different partitions. These partitions are referred as clusters. The partition clustering organizes data into single partition instead of representing data into nested structure like hierarchical clustering. Partition clustering is more useful for large data set in which it is difficult to represent data in tree structure. The most commonly used criterion is the Euclidean distance, which finds the minimum distance between points with each of the available cluster and assigning the point to cluster.

2.2.1. Distance-based clustering

Distance-based is also known as centroid-based clustering; clusters are represented by a central vector, which may not necessarily be a member of the data set. Most K-means type algorithms require the number of clusters k to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms. Also, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centroid. K-means has a number of interesting theoretical properties:

a) It partitions the data space into a structure known as a Voronoi diagram.

b) It is conceptually close to nearest neighbour classification.

c) It can be seen as a variation of model based classification.

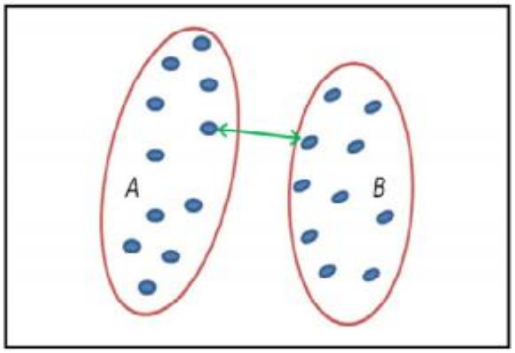

Fig.1. Mapping of single linkage clustering.

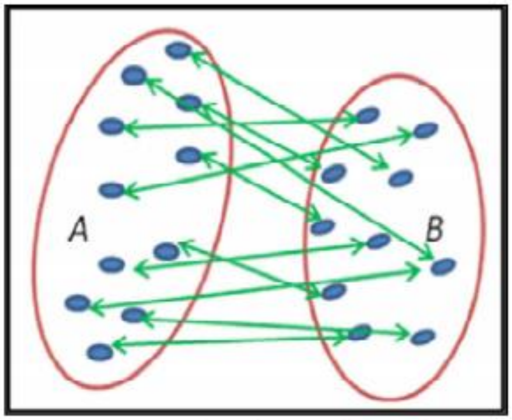

Figure 2. Mapping of average linkage clustering

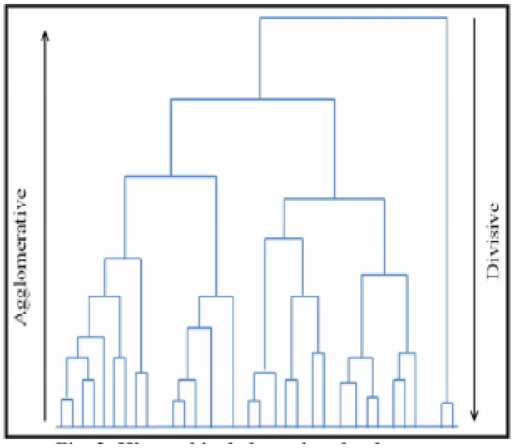

Figure 3. Hierarchical clustering dendrogram

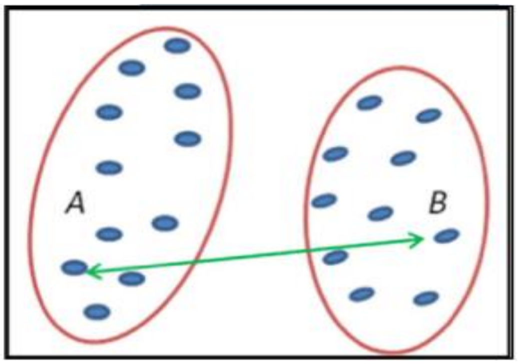

Fig. 4. Mapping of complete linkage clustering.

A. K-means Clustering

K-means algorithm is one of the best-known, benchmarked and most straightforward clustering algorithms [4], which is mostly applied to solve the clustering problem. In this procedure, the given data set is classified through a user-defined number of clusters, k. The main idea is to determine k centroids, one for each cluster. The objective function J is given as follows

$$Minimize\ J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

In this context, prototype vectors are called code words, which constitute a codebook. VQ aims to represent the data with a reduced number of elements while minimizing information loss. Although K-means clustering is still one of the most popular clustering algorithms yet few limitations are associated with k-means clustering, include:

a) There is no efficient and universal method for identifying the initial partitions and the number of clusters k and

b)k-means is sensitive to outliers and noise. Even if an object is quite far away from the cluster centroid, it is still forced into a cluster and thus, distorts the cluster shapes.

Fuzzy c-means clustering

Fuzzy c-means (FCM) is a clustering method which allows one point to belong to two or more clusters, unlike K-means where only one cluster is assigned to each point. This method was developed in [5] . The procedure of fuzzy c-means [4] is similar to that of k- means. It is based on the minimization of the following objective function.

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}^m ||x_i - v_j||^2; 1 < m < \infty$$

Where *m* is fuzzy partition matrix exponent for controlling the degree of fuzzy overlap, with *m*> 1. Fuzzy overlap refers to how fuzzy the boundaries between clusters are, that is the number of data points that have significant membership in more than one cluster, u_{ij} is the degree of membership of x_i in the cluster *j*, *x i* is the *i* th pattern of D-dimension data , v_j is *j* th cluster centre of the D-dimension.

2.2.2 Density-based clustering

In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas that are required to separate clusters are usually considered to be noise and border points. Density-based clustering algorithm finds clusters based on density of data points in a region. The key idea is that each instance of cluster based on density of data points in a region. The key idea is that each instance of a cluster the neighbourhood of a given radius has to contain at least a minimum number of objects i.e. the cardinality of the neighbourhood has to exceed a given threshold. This is completely different from the partition algorithms that use iterative relocation of points given a certain number of clusters. One of the density-based clustering algorithms is the DBSCAN.

2.2.3 Model-Based clustering.

These algorithms find good approximate of model parameters that best fit the data. They can be either partition or hierarchical, depending on the structure or model they hypothesize about the data set and the way they refine this model to identify partitioning. They are closer to density-based algorithms, in that they grow particular cluster so that the preconceived model is improved. However, they sometimes start with a fixed number of cluster and they do not use the same concept of density. Some of the advantages of partitioned based algorithms include that they are (i) relatively scalable and straightforward and (ii)suitable for the database with compact spherical cluster that is well-separated.

3. Clustering validation Techniques

The procedure of evaluating the results of clustering algorithms is known under the term cluster validity. In general terms, there are three approaches to investigate cluster validate (Theodoridis and Koutroubs, 1999). Here the basic idea is the evaluation of a clustering structure by comparing it to other clustering schemes, resulting by the same algorithms but with different parameter values. There are two criteria proposed for clustering evaluation and selection of an optimal clustering scheme.

- Compactness: The member of each cluster should be as close to each other as possible. A common measure of compactness is the variance.
- Separation: The cluster themselves should be widely separated.

Three common approaches are measuring the distance between two different clusters: the distance between the closest member of the clusters, the distance between the most distant members and the distance between the centers of the cluster. There are three different techniques for evaluating the result of the clustering algorithms:

- External Criteria
- Internal Criteria
- Relative Criteria

Both internal and external criteria are based on statistical methods and they have high computation demand. The external validity methods evaluate the clustering based on some user specific intuition. The internal criteria are based on some metrics is its computational complexity. The basis of the relative criteria is the comparison of the different input parameters on same data set. The aim of the relative criteria is to choose the best clustering scheme from the different results. The basis of the comparison is the validity index. Several validity indices have been developed and introduced.

3.1 Validity Indices

In this section several validity indices are introduced. The indices are user for measuring the "goodness" of a clustering result compared to other ones that were created by other clustering algorithms, where no overlapping between partitions is allowed.

3.1.1 Dunn and Dunn like indices

These clusters validate indices were introduced in [2]. The index definition is given by Equation

$$D = \min_{i=1\ldots n_c}\left\{ \min_{j=i+1\ldots n_c}\left(\frac{d(c_i,c_j)}{\max\limits_{k=1\ldots n_c}(diam(c_k))} \right)\right\}, \text{where}$$

$$d(c_i,c_j) = \min_{x\in c_i, y\in c_j}\{d(x,y)\} \text{ and } diam(c_i) = \max_{x,y\in c_i}\{d(x,y)\}$$

If a data set contains well-speared clusters, the distances among the cluster are usually large and the diameters of the cluster are accepted to be small [6]. Therefore larger value means better cluster configuration, the main disadvantage if the Dunn index is the following: the calculation of the index is time-consuming and this index is very sensitive to noise (as the maximum cluster diameter can be large in a noisy environment).

3.1.2 Davies Bouldin Index The Davies

Bouldin index [7] is based on the similarity measure of clusters (Rij) whose bases are the dispersion measure of the cluster (si) and the cluster dissimilarly measure (dij). The similarity measure of clusters (Rij) can be defined freely but it has to satisfy the following conditions [7]:

- $R_{ij} \geq 0$
- $R_{ij} = R_{ji}$
- $\text{if } \mathrm{s_i} = 0 \text{ and } \mathrm{s_j} = 0 \text{ then } \mathrm{R_{ij}} = 0$
- $\text{if } \mathrm{s_j} > \mathrm{s_k} \text{ and } \mathrm{d_{ij}} = \mathrm{d_{ik}} \text{ then } \mathrm{R_{ij}} > \mathrm{R_{ik}}$
- $\text{if } \mathrm{s_j} = \mathrm{s_k} \text{ and } \mathrm{d_{ij}} < \mathrm{d_{ik}} \text{ then } \mathrm{R_{ij}} > \mathrm{R_{ik}}$

Then the Davis- Bouldin index is defined as

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i, \text{ where}$$

$$R_i = \max_{j=1\ldots nc, i\neq j}\left(R_{ij}\right),\ i = 1\ldots n_c$$

3.1.3. RMSSDT and RS validity Indices

Usually, hierarchical clustering algorithms use these indices, but they can be used for evaluating the results of any clustering algorithms. The RMSSTD (root - mean - square standard deviation) index [9] is the variance of the clusters. As the aim of the clustering process to identify homogenous groups the lower RMSSTD value means better clustering.

$$RMSSTD = \sqrt{\frac{\sum_{\substack{i=1\ldots nc \\ j=1\ldots d}}\sum_{k=1}^{n_{ij}}\left(x_k - \overline{x_j}\right)^2}{\sum_{\substack{i=1\ldots nc \\ j=1\ldots d}}\left(n_{ij}-1\right)}}$$

The motivation RS (R Squared) index [9], described on Equation, the index is to measure the dissimilarity of the cluster. Formally it measures the degree of homogeneity degree between groups. The values of RS range from 0 to 1 where 0 means there is no difference between the cluster and 1 indicates that there is a significant difference among the clusters.

$$RS = \frac{SS_t - SS_w}{SS_t}, \text{ where}$$

$$SS_t = \sum_{j=1}^{d}\sum_{k=1}^{n_j}\left(x_k - \overline{x_j}\right)^2,\ SS_w = \sum_{\substack{i=1\ldots nc \\ j=1\ldots d}}\sum_{k=1}^{n_{ij}}\left(x_k - \overline{x_j}\right)^2$$

3.1.4 SD validity Index

The bases of SD validity index [10] are the average scattering of clusters and total separation of clusters. The scattering is calculated by the variance of the clusters and variance of the dataset, thus it can measure the homogeneity and compactness of the clusters. The variance of the dataset variance of a cluster is defined in Equation.

The variance of the dataset:

Variance of a cluster:

$$\sigma_x^p = \frac{1}{n}\sum_{k=1}^{n}\left(x_k^p - \overline{x^p}\right)^2 \qquad \sigma_{v_i}^p = \frac{1}{\left\|c_i\right\|}\sum_{k=1}^{n}\left(x_k^p - \overline{v_i^p}\right)^2$$

$$\sigma(x) = \begin{bmatrix}\sigma_x^1 \\ \vdots \\ \sigma_x^d\end{bmatrix} \qquad \sigma(v_i) = \begin{bmatrix}\sigma_{v_i}^1 \\ \vdots \\ \sigma_{v_i}^d\end{bmatrix}$$

The average scattering for clusters is defined as

$$Scatt = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{\left\|\sigma(v_i)\right\|}{\left\|\sigma(x)\right\|}$$

The total separation of cluster is based on the distance of cluster center points thus it can measure
the separation of clusters. Its definition is given by Equation.

$$Dis = \frac{\max_{i,j=1\ldots n_c}\left(\left\|v_j - v_i\right\|\right)}{\min_{i,j=1\ldots n_c}\left(\left\|v_j - v_i\right\|\right)}\sum_{k=1}^{n_c}\left(\sum_{\substack{j=1, \\ i\neq j}}^{n_c}\left\|v_j - v_i\right\|\right)^{-1}$$

The SD index can be defined based on equation 7 and 8 as follows

SD= α . Scatt + Dis

Where *α* is a weighting factor that is equal to D is a parameter in case of the maximum number of clusters. Lower SD index means better cluster configuration as in fact the cluster are compacted and separated.

4. Conclusion and Future Scope

The classification of objects is primary in many data processing application including data mining, medical diagnostics, pattern recognition and social paradigms. The objects already labelled are placed in supervised classified groups whole non labelled are grouped in unsupervised classified groups. This paper presented the method used for clustering and their limitation. In the hierarchical type of clustering methods, clusters are formed by iterative dividing the patterns into top-down or bottom up manner accordingly agglomerative and divisive or splitting hierarchical clustering methods are discussed. As opposed to hierarchical clustering, partitioned clustering assign data into k-clusters without any hierarchical structure by optimizing some criterion function. The most common criterion is finding Euclidean distance between the points with each of the available clusters and assigning the point to the cluster with minimum distance.

The benchmark k-means clustering validity index. Cluster analysis is one of the major tasks in the various research area. However, it may be found under different names in different contexts. Thus based on a certain clustering criterion the data are grouped so that data points in a cluster are more similar to each other than points in different clusters presented in a data set, there is a need of some kind of clustering result validation. Here, we have discussed several validity indices like Dunn and Dunn like, Davies Bouldin index, RMSSDT and RS validity index and SD validity index. There is an abundant amount of literature available in clustering and its applications, it is possible to cover the entirely, only basics and few important methods are included in this paper with their merits and demerits.

5. References

[1] J. Han , M. Kamber , J. Pei , Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2011 .

[2] A . Nagpal , A . Jatain , D. Gaur , Review based on data clustering algorithms, in: Proceedings of the IEEE Conference on Information and Communication Technologies, 2013 .

[3] G.P. Zhang , Neural networks for classification: a survey, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 30 (4) (2002) 451–462 .

[4] A.K. Jain , Data clustering: 50 years beyond *k* -means, Pattern Recognit. Lett. 31 (8) (2010) 651–666 .

[5] F.S. Marzano , D. Scaranari , G. Vulpiani , Supervised fuzzy-logic classification of hydrometeors using C-band weather radars, IEEE Trans. Geosci. Remote Sens. 45 (11) (2007) 3784–3799 .

[6] Guha, S, Rastogi, R., and Shim K. . ROCK: A Robust Clustering Algorithm for Categorical Attributes. In Proceedings of the IEEE Conference on Data Engineering, (1999)

[7] Rezaee, R., Lelieveldt, B.P.F., and Reiber, J.H.C. (1998). A New Cluster Validity Index for the Fuzzy c-Mean. *Pattern Recognition Letters*, 19, 237–246.

[8] M. Halkidi, Y. Batistakis and M. Vazirgiannis: On Clustering Validation Techniques, Journal of Intelligent Information Systems, Vol. 17, No. 2-3, pp. 107-145, 2001

[9] Xie, X.L. and Beni, G. (1991). A Validity Measure for Fuzzy Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4), 841–846.

[10] M. Halkidi and M. Vazirgiannis and Y. Batistakis: Quality Scheme Assessment in the Clustering Process, Proc. Of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, pp. 265-276, 2000.