VOL. 10, NO. 9, MAY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

CLUSTER VALIDITY MEASURES DYNAMIC CLUSTERING ALGORITHMS

S. Angel Latha Mary, A. N. Sivagami and M. Usha Rani
Department of Computer Science Engineering, Karpagam College of Engineering, Coimbatore, India
E-Mail: xavierangellatha@gmail.com

ABSTRACT

Cluster analysis finds its place in many applications especially in data analysis, image processing, pattern recognition, market research by grouping customers based on purchasing pattern, classifying documents on web for information discovery, outlier detection applications and act as a tool to gain insight into the distribution of data to observe characteristics of each cluster. This ensures that cluster places its identity in all domains. This paper presents the clustering validity measures which evaluates the results of clustering algorithms on data sets with the three main approaches of cluster validation techniques namely internal, external and relative criteria. Also it validates the cluster using the cluster indices namely Dunn's index, Davies- Boludin index and Generalized Dunn Index using K-mean and Chameleon algorithm.

Keywords: Cluster, algorithm.

1. INTRODUCTION

Clustering process involves grouping of data objects based on the likeness among them and aims at attaining high intra-clustering similarity and low inter-clustering similarity. The success of clustering application resides on the cluster validation measures which evaluate the goodness of clustering results. External cluster validation, internal cluster validation and relative cluster validation are the main categories of cluster validation.

In this paper we present a comparative study between these approaches, analysing 11 internal clustering validation measures, 7 external clustering validation measures and giving the general view of relative clustering validation measures. The main criteria for comparison **is**

Compactness: This criteria denotes how close the objects in the cluster are. Compactness is evaluated depending upon the variance of objects within a cluster. Lower variance denotes the cluster is highly compact.

Separation: This criteria denotes how distinct the object in the cluster are. Separation can be measured based on distance between representative objects of the two clusters. [1]

2. DIFFERENT ASPECTS OF CLUSTER VALIDATION

1. Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data without reference to external information. - Use only the data
4. Comparing the results of two different sets of cluster analyses to determine which is better.
5. Determining the 'correct' number of clusters.

3. CLUSTER VALIDATION METHODS

Determining the correct number of clusters in a data set has been, by far, the most common application of cluster validity. In general, indices of cluster validity fall into one of three categories following categories.

Internal Cluster Validation – Based on the information intrinsic to the data alone.

External Cluster Validation – Based on previous knowledge about the data

Relative Cluster Validation- Based on repeated analysis of same algorithm on different parameters to obtain stable result.

Internal cluster validation measures

Root Mean Square Standard Deviation (RMSSTD) index is the square root of the pooled sample variance of all the attributes within each cluster. It measures the homogeneity of the formed clusters. Thus in simple terms RMSSTD is the with-in group sum of squares of each cluster by the product of number of variables and the number of elements in the cluster. Also it is mainly used in hierarchical clustering algorithms where lower RMSSTD value denotes that the formed cluster is optimal where as higher value of RMSSTD at each hierarchical steps denotes that the formed cluster is worse. This index is valid for rectangular data. If the dissimilarity matrix is available then the index is only valid if the methods used are average, centroid and ward [2].

Root Squared (RS) index is aimed at measuring the dissimilarity of clusters. It is calculated by sum of squares between clusters to the total sum of squares of the whole data set.

The value of RS range from 0 to 1. If RS value is 0, it indicates there is no difference among clusters whereas 1 indicates clusters are considerably distinct [3].

Hubert's Γ statistic finds the distinct clusters by counting the correlation between them. For this it uses the

square matrix of same size. In a **ModifiedHubert's Γ Statistic**, for a data set with N points and M= N(N-1)/2 pairwise comparison of points, we define P as N*N where P(i,j) is the proximity matrix of all points in the data set from i to j. Similarly Q(i,j) is the proximity matrix of the centre of the cluster to which each point belongs. If the representative member of the cluster is its centre then,
If the value ofΓ is high, then the cluster is said to be well-distinct. But the fact here is when the number of cluster increases Γ also increases, thus a normalised version of Hubert's Γ statistic is used.

If normalised Hubert's statistic has values scaled from -1 to 1, large absolute value on it denotes a well separated clusters [4].

Calinski-Harabasz index (CH) is sometimes called the variance ratio criterion (VRC). Well-defined clusters have a large between-cluster variance and a small within-cluster variance. The larger the VRC ratio, the better the data partition. To determine the optimal number of clusters, maximize VRC_k with respect to *k*. The optimal number of clusters is the solution with the highest Calinski-Harabasz index value. The Calinski-Harabasz criterion is best suited for *k*-means clustering solutions with squared Euclidean distances. CH criterion is most suitable in case when clusters are more or less spherical and compact in their middle (such as normally distributed, for instance). Other conditions being equal, CH tends to prefer cluster solutions with clusters consisting of roughly the same number of objects [5].

Index I measures separation based on the maximum distance between the cluster centres and measures compactness based on the distances between objects and their cluster centres. If the value of I is higher, then the clusters are said to be well separated and compact [1].

Dunn's index is defined as the minimum of the ratio of the dissimilarity measure between two clusters to the diameter of cluster, where the minimum is taken over all the clusters in the data set. It aims at inter-cluster separation and intra-cluster compactness. This index is valid for both rectangular and dissimilarity data. One of the drawbacks of using this is the computational cost as the number of clusters and dimensionality of the data increase [6].

Generalized Dunn's index *(*vGD) larger value of indicates good clusters and the number of clusters that maximizes number of clusters [6].

The Silhouette index (S) validates the clustering based on the pairwise difference of between- and within-cluster distances. [1] The technique provides a succinct graphical representation of how well each object lies within its cluster. If s(i) is close to 1, then the data i is appropriately clustered. If s(i) is close to -1, then the data i would be more appropriate if it was clustered in its neighbouring cluster. If s(i) is near zero, then the data i is on the border of two natural clusters. The average s(i) over all data of a cluster is a measure of how tightly grouped all the data in the cluster are. Thus silhouette plots and averages may be used to determine the natural number of clusters within a dataset. [7]

Davies- Boludin (DB) index is well known for its better partition capability. Similar to the Dunn index, Davies-Bouldin index identifies clusters which are far from each other and compact. The DB index is obtained by calculating the average of all cluster similarity.

The Davies Boludin index measures the average of similarity between each cluster and its most similar one. The optimal clustering solution has the smallest Davies-Bouldin index value. [8]

Xie- Beni Index (XB) defines the inter-cluster separation as the minimum square distance between cluster centres, and the intra-cluster compactness as the mean square distance between each data object and its cluster centre. The optimal cluster number is reached when the minimum of XB is reached. They are mainly used in fuzzy clustering as it can validate fuzzy partitions considering the geometric features of clusters, which suit human feelings in most cases.The minimum value of XB indicates good clustering [1].

The SD validity index definition is based on the concepts of average scattering for clusters and total separation between clusters. In the sequel, we give the fundamental definition for this index. The optimal number of clusters can be obtained by minimizing the value of SD. [9]

S_Dbw Validity Index is based on the density of the cluster in addition to common criteria like compactness and separation towards which SD index also resides. It measures the intra-cluster variance and the inter-cluster variance. The intra cluster variance measures the average scattering of clusters. The density function around a point is defined as follows: it counts the number of points in a hyper-sphere whose radius is equal to the average standard deviation of clusters. Lower index value indicates better clustering schema [9].

External cluster validation measures

As discussed in the introduction, external cluster validity metrics use some predefined knowledge like class labels or number of cluster, for quality evaluation. In this scenario, good cluster structure means the same as predefined class structure in the dataset.

F- Measure is the harmonic mean of precision and recall values for each cluster. F-measure tries to balance the precision and recall values across all the clusters. The $F\text{−}Measure$ values are within the interval [0, 1] and larger values indicate higher clustering quality. The maximum value of F-measure is thus one [10].

NMI measure is called Normalized Mutual Information (NMI). Here the mutual information tries to quantify the amount of shared information between the clustering and the partition. The NMI value lies in the range [0, 1]. Values close to 1 indicate a good clustering. [10]

Entropy measures the purity of the clusters class labels. The entropy value becomes zero if the object in a cluster have same class label. The entropy value increases

when the class labels of objects in a cluster become more varied. For a perfect clustering, entropy value is zero, whereas the worst possible entropy value is $\log_2 m$ [10].

Purity is very similar to entropy. Purity quantifies the extent to which a cluster C_i contains entities from only one partition. In other words it measures how "pure" each cluster is.Thus, the purity of clustering C is defined as the weighted sum of the cluster-wise purity values [10]

General Rand index Rand index measures the similarity between data clustering. It corresponds to accuracy even when the class labels are not found. Adjusted rand index is a form of rand index that is adjusted for the chance grouping of elements.

The Rand index ranges from 0 to 1, with 0 indicating that the two data clusters do not agree on any pair of points and 1 ndicating that the data clusters are exactly the same.

Adjusted Rand index The adjusted Rand index is the corrected for chance version of the Rand index. Though the Rand Index may only yield a value between 0 and +1, the Adjusted Rand Index can yield negative values if the index is less than the expected index [11].

Jaccard Coefficient Jaccard index is a name often used for comparing similarity, dissimilarity, and distance of the data set.Measuring the Jaccard similarity coefficient between two data sets is the result of division between the number of features that are common to all divided by the number of properties.Jaccard distance is non-similar measurement between data sets. It can be determined by the inverse of the Jaccard coefficient which is obtained by removing the Jaccard similarity from the coefficient. It is equal to a number of features that are all minus by number of features that are common to all divided by the number of features.

Fowlkes and Mallows index Another method for comparing clusters was proposed by Fowlkes and Mallows as an alternative for Rand index. The Fowlkes-Mallows index computes the similarity between the clusters returned by the clustering algorithm and the benchmark classifications. The higher the value of the Fowlkes-Mallows index the more similar the clusters and the benchmark classifications are.The Fowlkes and Mallows index, when results of two clustering algorithms is used to evaluate the results, can be defined as TP is the number of true positives, FP is the number of false positives, FN is the number of false negatives[12].

Mirkin Metrics This coefficient assumes null value for identical clustering and positive values otherwise. It corresponds to the Hamming distance between the binary vector representations of each partition. It provides an alternative adjusted form of Rand index. However, unlike Hubert and Arabie's adjusted Rand (Hubert, 1985) it doesn't provide a correction for chance agreement. Meila (2005) also proposed a bounded version of this index [13].

Relative cluster validation Both internal and external cluster validation indices use statistical analysis of data which will increase the computational cost. Hence a criteria without statistical analysis is proposed which is relative cluster validation. [14]

The fundamental idea of this approach is to choose the best clustering scheme of a set of defined schemes according to a pre-specified criterion. The criterion has two major sections.

Optimization-like criteria, which are those for which higher (maximization) or lower (minimization) values naturally, indicate the best partitions.

Difference-like criteria, which are those primarily designedto assess the relative improvement between two consecutive partitions produced by a hierarchical clustering algorithm [14].

4. RESULTS AND DISCUSSION

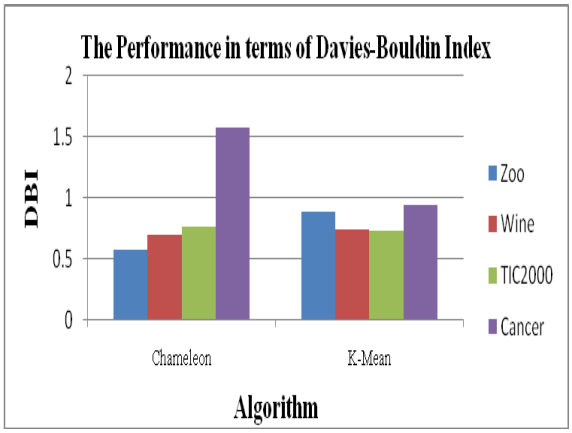

Figure-1. Performance in terms of Davies-Bouldin index.

This paper uses the cluster validation metrics Generalized Dunn Index (GDI) and Davies-Bouldin Index (DBI) real data sets Zoo data set, Wine data set, TIC2000 data set and Wisconsin Breast Cancer data set for the algorithms chameleon and k-mean.

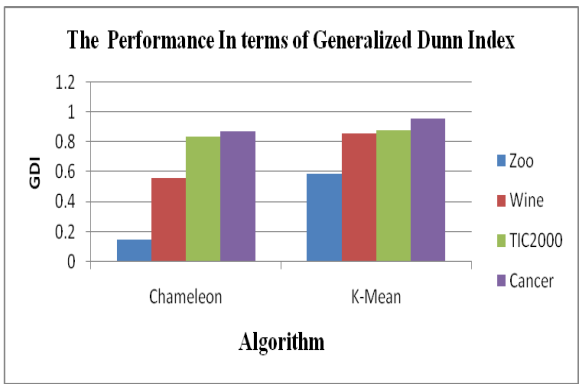

Figure-2. The Performance in terms of Generalized Dunn index.

In Figure-1 show the performance of the cluster accuracy in terms of Davies-Bouldin Index using the

algorithms chameleon and K-mean. Figure-2 shows the Performance of cluster accuracy in terms of Generalized Dunn Index.

The performance of the Chameleon is good when the data size is small in terms of cluster accuracy. It is poor when the data size is large, even though it posses the capabilities of dynamic concept.

5. CONCLUSIONS

The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage. In this paper, we investigated the validation properties of a suite of 11 existing internal clustering validation measures and 7 external clustering validation measures along with the overview of relative clustering validation. This helps in selection of appropriate validation index for different applications.

REFERENCES

[1] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao and Junjie Wu. 2010. “Understanding of Internal Cluster Validation Measures “IEEE International Conference on Data Mining.

[2] Cluster Analysis by Leland Wilkinson, Laszlo Engelman, James Corter, and Mark Coward.

[3] Cluster Validity in Clustering Methods by Qinpei Zhao.

[4] Emerging Technologies in Industry by Dominik RyzkoIntelligent.

[5] Calinski T., and J. Harabasz. 1974. "A dendrite method for cluster analysis." Communications in Statistics. Vol. 3, No. 1, pp. 1–27.

[6] Angel Latha Mary S., Shankar Kumar K. R. 2012. “Density Based Dynamic Data Clustering Algorithm based on Incremental Dataset” Journal of Computer Science” 8 (5): 656-664, ISSN 1549-3636.

[7] Peter J. Rousseeuw. 1987. "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics. 20:53–5.doi:10.1016/0377-0427(87)90125-7.

[8] Davies, David L., Bouldin, Donald W. 1979. “A Cluster Separation Measure". IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1(2):224–27.doi:10.1109/TPAMI. 1979.4766909.

[9] Clustering Validity Checking Methods: Part II by Maria Halkidi, YannisBatistakis, Michalis Vazirgiannis.

[10] Internal versus External cluster validation indexes by Erendira Rendon, Itzel Abundez, Alejandra Arizmendi and Elvia M. Quiroz.

[11] W. M. Rand. 1971. "Objective criteria for the evaluation of clustering methods". Journal of the American Statistical Association (American Statistical Association). 66 (336): 846–850.doi:10.2307/2284239. JSTOR 2284239

[12] Fowlkes E. B. Mallows C. L. (1 September 1983. “A Method for Comparing Two Hierarchical Clusterings". Journal of the American Statistical Association. 78 (383): 553.doi:10.2307/2288117

[13] http://darwin.phyloviz.net/ Comparing Partitions/ index.php? link=Tut11.

[14] On Clustering Validation Techniques by Maria Halkidi, YannisBatistakis, Michalis Vazirgiannis.