

A comparison of internal and external cluster validation indexes

ERÉNDIRA RENDÓN¹, ITZEL M. ABUNDEZ¹, CITLALIH GUTIERREZ, SERGIO DÍAZ ZAGAL¹, ALEJANDRA ARIZMENDI¹, ELVIA M. QUIROZ, AND ELSA ARZATE H.

División de Estudios de Postgrado e Investigación¹
Instituto Tecnológico de Toluca
Ex. Rancho la Virgen, Metepec, Edo. de México
MÉXICO
erendon@ittoluca.edu.mx; erendir@prodigy.net.mx

Abstract: The procedure of evaluating the results of a clustering algorithm is known under the term cluster validity. In general terms, cluster validity criteria can be classified in three categories: internal, external and relative. In this work we focus on the external and internal criteria. External indexes require a priori data for the purposes of evaluating the results of a clustering algorithm, whereas internal indexes do not. Consequently, different types of indexes are used to solve different types of problems and indexes selection depends on the kind of available data. It is interesting to note that, type of information or algorithm notwithstanding, they provided the highest degree of accuracy in group determining. That is why in this paper we show a comparison between external and internal indexes. Results obtained in this study indicate that internal indexes are more accurate in group determining in a given clustering structure. Five internal indexes were used in this study: BIC, CH, DB, SIL and DUNN. The groups that were used were obtained through clustering algorithms K-means and Bisecting-K-means.

Key-Words: Cluster validity, clustering algorithm, k-means, internal indexes, external indexes.

1 Introduction

The purpose of clustering is to determine the intrinsic grouping in a set of unlabeled data, where the objects in each group are indistinguishable under some criterion of similarity. Clustering is an unsupervised classification process fundamental to data mining (one of the most important tasks in data analysis). It has applications in several fields like bioinformatics [12], web data analysis [11], text mining and scientific data exploration [1]. Clustering refers to unsupervised learning and, for that reason it has no a priori data set information. However, to get good results, the clustering algorithm depends on input parameters. For instance, k-means [14] and CURE [13] algorithms require a number of clusters (k) to be created. In this sense, the question is: What is the optimal number of clusters? Currently, cluster validity indexes research has drawn attention as a means to give a solution [6]. Many different cluster validity methods have been proposed [7] [9] without any a priori class information.

Clustering validation is a technique to find a set of clusters that best fits natural partitions (number of clusters) without any class information.

Generally speaking, there are two types of clustering

techniques [15], which are based on external criteria and internal criteria.

- External validation: Based on previous knowledge about data.
- Internal validation: Based on the information intrinsic to the data alone.

If we consider these two types of cluster validation to determine the correct number of groups from a dataset, one option is to use external validation indexes for which a priori knowledge of dataset information is required, but it is hard to say if they can be used in real problems (usually, real problems do not have prior information of the dataset in question). Another option is to use internal validity indexes which do not require a priori information from dataset.

In the literature we can find different external and internal indexes, each approach has clear scope, in this paper we present a comparative study between these two approaches, analyzing four external indexes (the most referenced in literature): F-measure [16], NMIMeasure [17], Entropy [18], Purity and five internal indexes: BIC, CH, DB, SIL, DUNN [19,20,21,2]. We used *K-means* *Bisecting K-means* clustering algorithms to generate clusters.

The rest of the paper is organized as follows: section 2 presents surveys of related works. Section 3 offers a light analysis about some index validation. Section 4 contains details about clustering algorithms used. Section 5 presents the study comparative; results obtained and discuss some findings from these results. Finally, we conclude by briefly showing our contributions and further works.

2 Previous works

Almost every clustering algorithm depends on the characteristics of the dataset and on the input parameters. Incorrect input parameters may lead to clusters that deviate from those in the dataset. In order to determine the input parameters that lead to clusters that best fit a given dataset, we need reliable guidelines to evaluate the clusters; clustering validity indexes have been recently employed. In general, clustering validity indexes are usually defined by combining compactness and separability.

1.- *Compactness*: This measures closeness of cluster elements. A common measure of compactness is variance.

2.- *Separability*: This indicates how distinct two clusters are. It computes the distance between two different clusters. The distance between representative objects of two clusters is a good example. This measure has been widely used due to its computational efficiency and effectiveness for hypersphere-shaped clusters.

There are three approaches to study cluster validity [10]. The first is based on external criteria. This implies that we evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a dataset, i.e. external information that is not contained in the dataset. The second approach is based on internal criteria. We may evaluate the results of a clustering algorithm using information that involves the vectors of the datasets themselves. Internal criteria can roughly be subdivided into two groups: the one that assesses the fit between the data and the expected structure and others that focus on the stability of the solution [22]. The third approach of clustering validity is based on relative criteria, which consists of evaluating the results (clustering structure) by comparing them with other clustering schemes.

In recent times, many indexes have been proposed in the literature, which are used to measure the fitness of the partitions produced by clustering algorithm [2]. The Dunn index [2] measures the ratio between the smallest cluster distance and the largest intra-cluster in a partitioning; several variations of

Dunn have been proposed [3][4]. DB measures the average similarity between each cluster and the one that most resembles it. [5]. The SD index [6] is defined based on the concepts of the average scattering for clustering and total separation among clusters. The S_Dbw index is very similar to SD index; this index measures the intra-cluster variance and inter-cluster variance. The index PS [7] uses nonmetric distance based on the concept of point symmetry [8], and measures the total average symmetry with respect to the cluster centers. Chow [9] proposes the CS index which obtains good clustering results when the densities and sizes are different, but its computational cost is elevated. The BIC index is derived from Bayes's theorem [19], and is used to determine which probability-based mixture is the most appropriate. Silhouette clustering structure quality; taking into account group compactness, separation between groups. On the other hand, external measures include Entropy, Purity, NMIMeasure and F-Measure.

3 Analysis of indexes

In this section, we offer an overview of internal and external validity indexes that were used in our study.

3.1 Internal validity indexes

- Bic index

The Bayesian information criterion (BIC) [19] is devised to avoid overfitting, and is defined as:

$$BIC = -\ln(L) + v\ln(n)$$

Where n is the number of objects, L is the likelihood of the parameters to generate the data in the model, and v is the number of free parameters in the Gaussian model. The BIC index takes into account both fit of the model to the data and the complexity of the model. A model that has a smaller BIC is better.

- Calinski-Harabasz index

This index is computed by

$$CH = \frac{\text{trace}(S_B)}{\text{trace}(S_w)} \cdot \frac{n_p - 1}{n_p - k}$$

Where (S_B) is the between-cluster scatter matrix, (S_w) the internal scatter matrix, n_p the number of clustered samples, and k the number of clusters.

- Davies-Bouldin index (DB)

This index aim to identify sets of clusters that are compact and well separated. The Davies-Bouldin index is defined as:

$$BD = \frac{1}{c} \sum_{i=1}^c \text{Max}_{i \neq j} \left\{ \frac{d(X_i) + d(X_j)}{d(c_i, c_j)} \right\}$$

Where c denotes the number of clusters, i, j are cluster labels, then $d(X_i)$ and $d(X_j)$ are all samples in clusters i and j to their respective cluster centroids, $d(c_i, c_j)$ is the distance between these centroid. Smaller value of DB indicates a “better” clustering solution.

- Silhouette index

For a given cluster, $X_j (j = 1, \dots, c)$, the silhouette technique assigns to the i th sample of X_j a quality measure, $s(i) = (i = 1, \dots, m)$, known as the silhouette width. This value is a confidence indicator on the membership of the i th sample in the cluster X_j and it is defined as:

$$s(i) = \frac{(b(i) - a(i))}{\text{Max}\{a(i), b(i)\}}$$

Where $a(i)$ is the average distance between the i th sample and all of samples included in X_j ; $b(i)$ is the minimum average distance between the i th sample and all of the samples clustered in $X_k (k = 1, \dots, c; k \neq j)$

- Dunn index

Dunn index is defined as:

Dunn =

$$\min_{1 \leq i \leq c} \left\{ \min \left\{ \frac{d(c_i, c_j)}{\max_{1 \leq k \leq c} (d(X_k))} \right\} \right\}$$

Where $d(c_i, c_j)$ defines the intercluster distance between cluster X_i and X_j ; $d(X_k)$ represents the intracluster distance of cluster (X_k) and c is the number of cluster of dataset. Large values of index Dunn correspond to good clustering solution.

3.2 External validity indexes

- F-measure

Combines the precision and recall concepts from information retrieval. We then calculate the recall and precision of that cluster for each class as:

$$\text{Recall}(i, j) = \frac{n_{ij}}{n_i}$$

And

$$\text{Precision}(i, j) = \frac{n_{ij}}{n_j}$$

Where n_{ij} is the number of objects of class i that are in cluster j , n_j is the number of objects in cluster j , and

n_i , is the number of objects in class i . The F – Measure of cluster j and class i is given by the following equation:

$$F(i, j) = \frac{2\text{Recall}(i, j)\text{Precision}(i, j)}{\text{Precision}(i, j) + \text{Recall}(i, j)}$$

The F – Measure values are within the interval $[0, 1]$ and larger values indicate higher clustering quality.

- Nmimeasure

Is called Normalized Mutual Information (NMI). The NMI of two labeled objects can be measured as:

$$\text{NMI}(X, Y) = \frac{I(X, Y)}{\sqrt{H(X)H(Y)}}$$

Where, $I(X, Y)$ denotes the mutual information between two random variables X and Y and $H(X)$ denotes the entropy of X , X will be consensus clustering while Y will be the true labels.

- Purity

Purity is very similar to entropy. We calculate the purity of a set of clusters. First, we cancel the purity in each cluster. For each cluster, we have the purity $P_j = \frac{1}{n_j} \text{Max}_i (n_j^i)$ is the number of objects in j with class label i . In other words, P_j is a fraction of the overall cluster size that the largest class of objects assigned to that cluster represents. The overall purity of the clustering solution is obtained as a weighted sum of the individual cluster purities and given as:

$$\text{Purity} = \sum_{j=1}^m \frac{n_j}{n} P_j$$

Where n_j is the size of cluster j , m is the number of clusters, and n is the total number of objects.

- Entropy

Entropy measures the purity of the clusters class labels. Thus, if all clusters consist of objects with only a single class label, the entropy is 0. However, as the class labels of objects in a cluster become more varied, the entropy increases. To compute the entropy of a dataset, we need to calculate the class distribution of the objects in each cluster as follows:

$$E_j = \sum_i p_{ij} \log(p_{ij})$$

Where the sum is taken over all classes. The total entropy for a set of clusters is calculated as the weighted sum of the entropies of all clusters, as shown in the next equation

$$E = \sum_{j=1}^m \frac{n_j}{n} E_j$$

Where n_j is the size of cluster j , m is the number of clusters, and n is the total number of data points.

4 K-means and Bisecting K-means algorithms

4.1 K-means clustering

K-means is a popular nonhierarchical clustering technique. In this case the k representative objects are called centroids. K-means is an iterative algorithm, and its basic idea is to start with an initial partition and assign objects to clusters so that the squared error decreases. The algorithm follows simple ways to classify a given data set into k clusters fixed a priori. The algorithm proceeds as follows:

1. Select an initial k clusters centroid.
2. Assign each object to its closest cluster centroid. That generates a new partition.
3. Compute the centroid of the new partition.
4. Repeat steps 2, and 3 until convergence is obtained. Typical convergence criteria are: no reassignment of patterns to new cluster centers, or minimal decrease in squared error.
5. Repeat steps 2 and 3 until convergence is obtained.

4.2 Bisecting K-means Algorithm

The bisecting K-means algorithm [10]. Begins with a single cluster of all dataset and works as follows:

1. Pick a cluster to split.
2. Find two sub-clusters using the basic K-means algorithm.
3. Repeat step 2, the bisecting step, for a fixed number of times and take the split that produces the clustering with the highest overall similarity.(for each cluster, its

similarity is the average pairwise objects similarity), and we seek to minimize that sum over all clusters number clusters.

4. Repeat steps 1, 2 and 3 until the desired number of clusters is reached.

5. Study Comparative

In this section, we show experimentally tested using the K-means and Bisecting K-means algorithms. We used 12 synthetic data sets. These data sets were used by Maria Halkidi [6] and Chien-Hsing Chou [7][9].

To find the best partition, we have used the K-means and Bisecting K-means algorithms with its input parameters (K) ranging between 2 and 8.

Table 1 presents a summary of the tests carried out in the datasets that were clustered with the K-means Bisecting algorithm, using internal clustering validity indexes; from there we can see that DB, SIL and DUNN indexes identified the correct number of groups in 11 of the trials, and CH index gave wrong results in datasets 2, 8 and Cmc. BIC index gave wrong results in datasets 2, 7, 8, 12 and Cmc. Additionally, none of the indexes identified the correct number of clusters of the datasets 2 and 8. Table 2 presents a summary of the tests carried out in the datasets that were clustered with the Bisecting K-means algorithm, using external cluster validity indexes; from where we can see that the F-measure index correctly identified the number of groups in all trials and Entropy only gave correct results in Dataset6. In conclusion, from these tests it can be said that internal validity indexes are more accurate in the identification of correct group numbers in clusters formed with the Bisection K-means algorithm.

DB and SIL indexes gave out better results (see table 3) when K-means algorithm and internal validity indexes were used; accurate results were obtained in 11 out of 13 trials. The CH index was accurate in 10 out of 13 trials and BIC DUNN was accurate in 9 out of 13 trials. When the K-means algorithm and external validity indexes were used (see table 4), the F-measure and NMIMeasure indexes gave correct results in 10 out of 13 trials and Entropy was accurate only in 3 out of 13 trials.

As mentioned earlier, internal validity indexes do not require a priori information about the dataset in question and, in spite of that, are the most accurate.

Table 1. Overview of the results obtained with internal validity indexes applied to the Bisection K-means algorithm.

Datasets	BIC	CH	DB	SIL	DUNN
Dataset1	♦	♦	♦	♦	♦
Dataset2					
Dataset3			♦	♦	♦
Dataset4	♦	♦	♦	♦	♦
Dataset5	♦	♦	♦	♦	♦
Dataset6	♦	♦	♦	♦	♦
Dataset7		♦	♦	♦	♦
Dataset8					
Dataset9	♦	♦	♦	♦	♦
Dataset10	♦	♦	♦	♦	♦
Dataset11	♦	♦	♦	♦	♦
Dataset12			♦	♦	♦
Cmc		♦	♦	♦	♦
Total	7	9	11	11	11

Table 2. Overview of the results obtained with external validity indexes applied to the Bisection K-means algorithm.

Datasets	F-Measure	NMI-Measure	Purity	En- tropy
Dataset1	♦	♦	♦	
Dataset2				
Dataset3	♦	♦	♦	
Dataset4	♦	♦	♦	
Dataset5	♦			
Dataset6	♦	♦	♦	♦
Dataset7	♦	♦		
Dataset8				
Dataset9	♦	♦	♦	
Dataset10	♦	♦	♦	
Dataset11	♦	♦		
Dataset12	♦	♦		
Cmd	♦			
Total	11	9	6	1

Table 3. Overview of the results obtained with internal validity indexes applied to the K-means algorithm.

Datasets	BIC	CH	DB	SIL	DUNN
Dataset1	♦	♦	♦	♦	♦
Dataset2	♦	♦			
Dataset3			♦	♦	
Dataset4	♦	♦	♦	♦	♦
Dataset5	♦	♦	♦	♦	♦
Dataset6	♦	♦	♦	♦	
Dataset7		♦	♦	♦	♦
Dataset8	♦				
Dataset9	♦	♦	♦	♦	♦
Dataset10	♦	♦	♦	♦	♦
Dataset11	♦	♦	♦	♦	♦
Dataset12			♦	♦	♦
Cmc		♦	♦	♦	♦
Total	9	10	11	11	9

Table 4. Overview of the results obtained with external validity indexes applied to the K-means algorithm.

Datasets	F-Measure	NMI-Measure	Purity	En- tropy
Dataset1	♦	♦	♦	♦
Dataset2			♦	
Dataset3	♦	♦	♦	♦
Dataset4	♦	♦	♦	
Dataset5	♦	♦	♦	
Dataset6	♦	♦	♦	
Dataset7		♦		
Dataset8			♦	
Dataset9	♦	♦	♦	♦
Dataset10	♦	♦	♦	
Dataset11	♦	♦		
Dataset12	♦	♦		
Cmd	♦			
Total	10	10	9	3

6. Conclusions and further work

This paper presents a comparison between two clustering validity index approaches, internal and external; carrying out analyses of four external indexes and four internal indexes. 13 datasets were used, which were clustered using the K-means and Bissection K-means algorithms. Each dataset was clustered with different K values ($K = 1, \dots, 8$ groups). Out of 65 (13×5) cases where the results of the Bissection K-means algorithm using internal indexes, correct group numbers were obtained 86% of the time, and in 51.9% when external indexes (13×4) were used. When clusters of the K-means algorithm were clustered using internal indexes, 76.9% of accuracy was obtained; and 61.5% with external indexes. From which we can infer that internal indexes are more precise in real group number determination than external indexes, or at least with the used datasets.

References:

- [1] Jain, A. K., Murty, M.N. Flynn, P.J. *Data clustering: A review*. ACM Computer. Surveys 31 (3), 1999, pp.264-323.
- [2] Dunn, J. C. *A Fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters*. J. Cyber. 3, 1973, pp. 32-57.
- [3] Pal N. R. and Biswas J. *Cluster Validation using graph theoretic concepts*, Pattern Recognition, Vol.30, No. 6. 1997. pp. 847-857.
- [4] Bezdek J. C. Pal N.R., *Some new indexes of cluster validity*. IEEE Trans. Syst.Man, Cyber. Part B 28 (3), 1998, pp. 301-315.
- [5] Davis D. L., Bouldin D.W. *A cluster separation measure*. IEEE Trans. Pattern Anal. Mach.Intel. (PAMI)1 (2), 1998, pp.224-227
- [6] Halkidi M., Vazirgiannis, M. *Quality scheme assessment in the clustering process*. In Proc. PKDD (Principles and Practice of Knowledge in databases). Lyon, France. Lecture Notes in Artificial Intelligence. Spring –Verlag Gmbh, vol.1910, 2000, pp. 265-279.
- [7] Chow C.H, Su M.C and Lai Eugene. *Symmetry as A new measure for Cluster Validity*. 2 th. WSEAS Int.Conf. scientific Computation and Soft Computing, Crete, Greece, 2002, pp. 209-213.
- [8] Su M.C, Chow C.H. *A Modified Version of the K-Means Algorithm with a Distance Based on Cluster Symmetry*. IEEE Trans. Pattern Anal. And Machine Intelligence, vol. 23. No.6, 2001, pp.674-680.
- [9] Chow C.H, Su M.C and Lai Eugene. *A new Validity Measure for Clusters with Different Densities*. Pattern Anal. Applications, 7, 2004, pp.2005-2020.
- [10] Theodoridis, S., Koutroubas, K. *Pattern Recognition*, Academic Press, USA, 1999.
- [11] Athena Vakali, Jaroslav Pokorný and Theodore Dalamagas. *An Overview of Web Data Clustering Practices*, Lecture Notes Computer Science, Vol. 3268, 2005, pp.597-606.
- [12] M.J.L. Hoon, S. Imoto, J. Nolan and S. Miyano. *Open source clustering software*. Bioinformatics, Vol. 20 No. 9, 2004, pp. 1453-1454.
- [13] Guha Sudipto, Rastogi Rajeev, Shim Kyuseok. *CURE: An Efficient Clustering Algorithm for Large DataBases*. In Proceedings of the CAM SIGMOD Conference on Magnagement of Data, Seatle, Washington, U.S.,01-04 Jun., 1998, pp. 73-83.
- [14] J.B. MacQueen. *Some Methods for classification and Analysis of Multivariable Observations*, Proceeding of 5th Berkeley on Mathematical Statistics and Probability, University of California Press, 1967, pp. 281-297.
- [15] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. *Cluster Validity methods: Part I*, SIGMOD Record, Vol.31(2), 2002, pp. 40-45.
- [16] Larsen and C. Aone. *Fast and effective text mining using linear-time document clustering*. In Proc. 5th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 1999, pp. 16-22.
- [17] Strehl A and Ghosh J. *Cluster ensembles – a knowledge reuse framework for combining multiple partitions*. Journal on Machine Learning Research (JMLR) 3, 2002, pp.583–617.
- [18] Shannon C. E. *A Mathematical Theory of Communication*, The Bell System Technical Journal, 1948, pp. 379 – 423.
- [19] Raftery A. *A note on Bayes factors for log-linear contingency table models with vague prior information*. Journal of the Royal Statistical Society. 48(2), 1986, pp. 249-250.
- [20] Calinski, T. and J. Harabasz. *A dendrite method for cluster analysis*. Commun. Stat. 3: , 1974, pp. 1-27.
- [21] P.J. Rousseeuw. *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, J. Comp App. Math, vol. 20, 1987, pp. 53-65.
- [22] Pacual D., Pla F., Sánchez J.S. *Cluster validation using information stability measures*, Pattern Recognition Letters 31, 2010, pp.454-461.