Cluster Validity Measurement Techniques

Ferenc Kovács, Csaba Legány, Attila Babos

Department of Automation and Applied Informatics
Budapest University of Technology and Economics
Goldmann György tér 3, H-1111 Budapest, Hungary
{ferenc.kovacs, csaba.legany, attila.babos}@aut.bme.hu

Abstract: Clustering is an unsupervised process in the data mining and pattern recognition and most of the clustering algorithms are very sensitive to their input parameters. Therefore it is very important to evaluate the result of the clustering algorithms. It is difficult to define when a clustering result is acceptable, thus several clustering validity techniques and indices have been developed. In this paper the most commonly used validity indices are introduced and compared to each other.

Keywords: data mining, clustering algorithms, cluster validity, validity indices

1 Introduction

One of the best known problem in the data mining is the clustering. Clustering is the task of categorising objects having several attributes into different classes such that the objects belonging to the same class are similar, and those that are broken down into different classes are not. Clustering is the subject of active research in several fields such as statistics, pattern recognition, machine learning and data mining. A wide variety of clustering algorithms have been proposed for different applications [1].

Clustering is mostly unsupervised process thus the evaluation of the clustering algorithms is very important. In the clustering process there are no predefined classes therefore it is difficult to find an appropriate metric for measuring if the found cluster configuration is acceptable or not. Several clustering validity approaches have been developed [2] [3].

The main disadvantage of these validity indices is that they cannot measure the arbitrary shaped clusters as they usually choose a representative point from each cluster and they calculate distance of the representative points and calculate some other parameter based on these points (for example: variance).

The rest of the paper is organized as follows. General properties of clustering algorithms and cluster validity techniques are introduced in Section 2. The detailed investigation of the most commonly used cluster validity indices is given in Section 3. The experimental results and comparison of the indices are outlined in Section 4.

2 Related Work

The clustering problem is to partition a data set into groups (clusters) so that the data elements within a cluster are more similar to each other than data elements in different clusters [4]. There are different types of clustering algorithms and they can be classified into the following groups [1]:

- *Partitional Clustering:* These algorithms decompose directly data set into a set of disjoint clusters. They attempt to determine an integer number of partitions that optimise a certain criterion function. This optimisation is an iterative procedure.
- *Hierarchical Clustering:* These algorithms create clusters recursively. They merge smaller cluster into larger ones or split larger clusters into smaller ones.
- *Density-based Clustering:* The key point of these algorithms is to create clusters based on density functions. The main advantage of these algorithms is to create arbitrary shaped clusters.
- *Grid-based Clustering:* These types of algorithms are mainly proposed for spatial data mining. They quantise the search space into finite number of cells.

The result of a clustering algorithm can be very different from each other on the same data set as the other input parameters of an algorithm can extremely modify the behaviour and execution of the algorithm. The aim of the cluster validity is to find the partitioning that best fits the underlying data. Usually 2D data sets are used for evaluating clustering algorithms as the reader easily can verify the result. But in case of high dimensional data the visualisation and visual validation is not a trivial tasks therefore some formal methods are needed.

The process of evaluating the results of a clustering algorithm is called cluster validity assessment. Two measurement criteria have been proposed for evaluating and selecting an optimal clustering scheme [5]:

- *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 clusters themselves should be widely separated. There are three common approaches measuring the distance between two different clusters: distance between the closest member of the clusters, distance between the most distant members and distance between the centres of the clusters.

There are three different techniques for evaluating the result of the clustering algorithms [6]:

- *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 which are based on data set and the clustering schema. The main disadvantage of these two methods is its computational complexity.

The basis of the relative criteria is the comparison of the different clustering schema. One or more clustering algorithms are executed multiple times with different input parameters on same data set. The aim of the relative criteria is to choose the best clustering schema from the different results. The basis of the comparison is the validity index. Several validity indices have been developed and introduced [7] [8] [9] [10] [11] [12].

Most widely used validity indices are introduced in the following section.

3 Validity Indices

In this section several validity indices are introduced. These indices are used for measuring “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. These indices are usually suitable for measuring crisp clustering. Crisp clustering means having non overlapping partitions. Table 1 describes the used notation in validity indices.

Notation	Meaning
n_c	Number of clusters
d	Number of dimension
$d(x,y)$	Distance between two data element
$\overline{X_j}$	Expected value in the jth dimension
$\|X\|$	$\sqrt{X^T X}$, where X^T is a column vector
n_{ij}	Number of element in ith cluster jth dimension
n_j	Number of element in jth dimension in the whole data set
v_i	Centre point of the ith cluster
c_i	ith cluster
$\|c_i\|$	Number of element in the ith cluster

Table 1
Notation in validity indices

3.1 Dunn and Dunn like Indices

These cluster validity indices have been introduced in paper [7]. The index definition is given by Equation 1.

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

$$d\left(c_i,c_j\right) = \min_{x\in c_i, y\in c_j}\left\{d\left(x,y\right)\right\} \text{ and } diam\left(c_i\right) = \max_{x,y\in c_i}\left\{d\left(x,y\right)\right\} \quad (1)$$

If a data set contains well-separated clusters, the distances among the clusters are usually large and the diameters of the clusters are expected to be small [3]. Therefore larger value means better cluster configuration. The main disadvantages of the Dunn index are 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). Several Dunn-like indices have been proposed [6] [13]. These indices use different definition for cluster distance and cluster diameter.

3.2 Davies Bouldin Index

The Davies – Bouldin index [8] is based on similarity measure of clusters (R_{ij}) whose bases are the dispersion measure of a cluster (s_i) and the cluster dissimilarity measure (d_{ij}). The similarity measure of clusters (R_{ij}) can be defined freely but it has to satisfy the following conditions [8]:

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

Usually R_{ij} is defined in the following way:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

$$d_{ij} = d\left(v_i, v_j\right),\ s_i = \frac{1}{\left\|c_i\right\|} \sum_{x \in c_i} d\left(x, v_i\right) \tag{2}$$

Then the Davies – Bouldin index is defined as

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

$$R_i = \max_{j=1 \ldots nc, i \neq j} \left(R_{ij}\right),\ i = 1 \ldots n_c \tag{3}$$

The Davies – Boludin index measures the average of similarity between each cluster and its most similar one. As the clusters have to be compact and separated the lower Davies – Bouldin index means better cluster configuration.

3.3 RMSSDT and RS Validity Indices

Usually hierarchical clustering algorithms use these indices but they can be used for evaluating the results of any clustering algorithm. The RMSSTD (root – mean – square standard deviation) index [9] is the variance of the clusters, formally defined on Equation 4, thus it measures the homogeneity of the clusters. As the aim of the clustering process to identify homogenous groups the lower RMSSTD value means better clustering.

$$RMSSTD = \sqrt{\frac{\sum_{\substack{i=1\ldots nc \\ j=1\ldots d}} \sum_{k=1}^{n_{ij}} \left(x_k - \overline{x_j} \right)^2}{\sum_{\substack{i=1\ldots nc \\ j=1\ldots d}} \left(n_{ij} - 1 \right)}} \tag{4}$$

The motivation RS (R Squared) index [9], described on Equation 5, index is to measure the dissimilarity of clusters. Formally it measures the degree of homogeneity degree between groups. The values of RS range from 0 to 1 where 0 means there are no difference among the clusters and 1 indicates that there are significant difference among the clusters.

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

$$SS_t = \sum_{j=1}^{d} \sum_{k=1}^{n_j} \left(x_k - \overline{x_j} \right)^2, \; SS_w = \sum_{\substack{i=1\ldots nc \\ j=1\ldots d}} \sum_{k=1}^{n_{ij}} \left(x_k - \overline{x_j} \right)^2 \tag{5}$$

3.4 SD Validity Index

The bases of SD validity index [12] are the average scattering of clusters and total separation of clusters. The scattering is calculated by 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 and variance of a cluster are defined in Equation 6.

Variance of the dataset:

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

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

Variance of a cluster:

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

$$\sigma(v_i) = \begin{bmatrix} \sigma_{v_i}^1 \\ \vdots \\ \sigma_{v_i}^d \end{bmatrix} \tag{6}$$

The average scattering for clusters is defined as

$$Scatt = \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{\left\| \sigma(v_i) \right\|}{\left\| \sigma(x) \right\|} \tag{7}$$

The total separation of clusters is based on the distance of cluster centre points thus it can measures the separation of clusters. Its definition is given by Equation 8.

$$Dis = \frac{\max\limits_{i,j=1\ldots n_c}\left(\left\| v_j - v_i \right\|\right)}{\min\limits_{i,j=1\ldots n_c}\left(\left\| v_j - v_i \right\|\right)} \sum_{k=1}^{n_c} \left(\sum_{\substack{j=1,\\ i \neq j}}^{n_c} \left\| v_j - v_i \right\| \right)^{-1} \tag{8}$$

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

$$SD = \alpha \cdot Scatt + Dis \tag{9}$$

where α is a weighting factor that is equal to Dis parameter in case of maximum number of clusters. Lower SD index means better cluster configuration as in this case the clusters are compacts and separated.

3.5 S_Dbw Validity Index

This validity index has been proposed in [11]. Similarly to SD index its definition is based on cluster compactness and separation but it also takes into consideration the density of the clusters. Formally the S_Dbw index measures the intra-cluster variance and the inter-cluster variance. The intra cluster variance measures the average scattering of clusters and it is described by Equation 7. The inter – cluster density is defined as follows

$$Dens\_bw = \frac{1}{n_c\left(n_c - 1\right)} \sum_{i=1}^{n_c} \left(\sum_{\substack{j=1,\\ i \neq j}}^{n_c} \frac{density(u_{ij})}{\max\left\{ density(v_i), density(v_j) \right\}} \right) \tag{10}$$

where u_{ij} is the middle point of the line segment that is defined by the v_i and v_j clusters centres. 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. The average standard deviation of clusters is defined as

$$stdev = \frac{1}{n_c} \sqrt{\sum_{i=1}^{n_c} \left\| \sigma\left(v_i\right) \right\|} \tag{11}$$

The S_Dbw index is defined in the following way:

$$S\_Dbw = Scatt + Dens\_bw \tag{12}$$

The definition of S_Dbw indicates that both criteria of “good” clustering are properly combined and it enables reliable evaluation of clustering results. Lower index value indicates better clustering schema.

4 Experimental Results

The clustering algorithms and validity indices were evaluated synthetically generated data set. These data were generated by our data set generator. The validity indices were evaluated using the following datasets:

- Well separated clusters: the cluster elements were generated around the cluster centres points using normal distribution.
- Ring shaped clusters: Two cluster, which contains each other.
- Arbitrary shaped clusters: some arbitrary shaped clusters close to each other.

The used data sets are depicted on Figure 1.

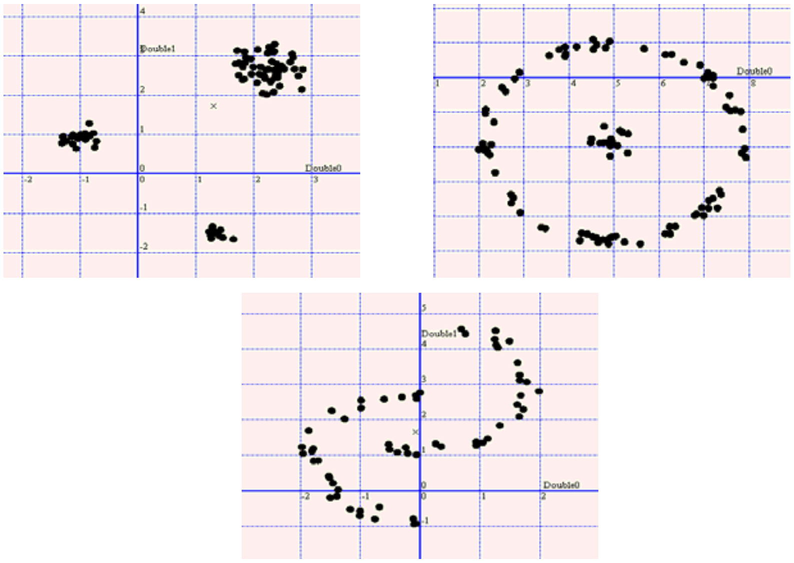

Figure 1

The used data set in experimental evaluation

Figure 2 shows a comparison of the validity indices on the first data set. The used clustering algorithm is k-means algorithm and in the first case it found the right clustering schema but in the second case it generates wrong cluster configuration. In this case it easy to identify that the validity indices can compare, in appropriate way, the result of the clustering algorithm.

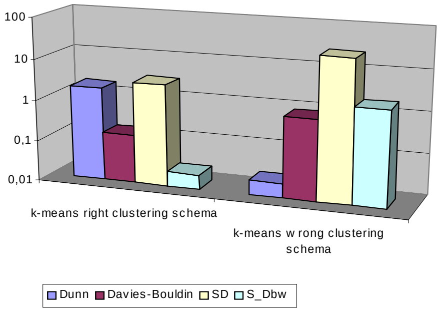

Figure 2
Validity indices on the first dataset

Figure 3 shows the validity indices based clustering of the second data set. Two clustering results are compared: a right clustering result (using DB-Scan algorithm) and a one (using k-means algorithm). A result a little bit surprising as the Dunn and S_Dbw index can identify the right clustering result but the other indices offer wrong decision.

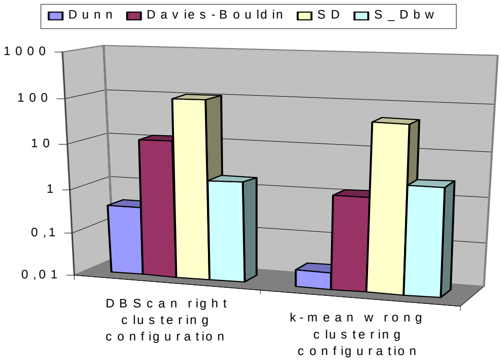

Figure 3
Validity indices on the second data set

Two clustering results, which are based on third data set, are depicted in Figure 4. The comparison of the validity indices are given in Figure 5. It is possible to realise that only the Dunn index can identify the right clustering schema. The main disadvantage of the current validity indices is that they cannot identify the right clustering schema unless the clusters are well separated.

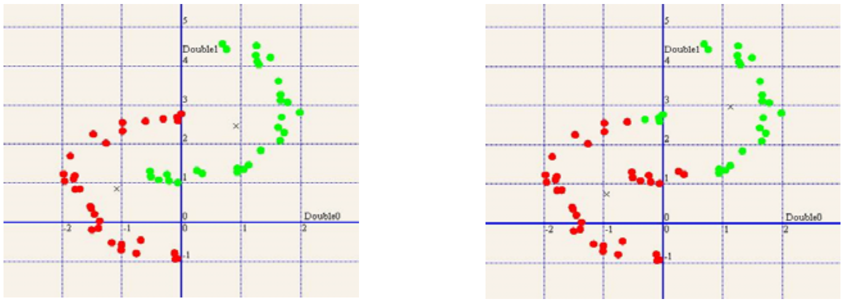

Figure 4
Clustering results on third data set

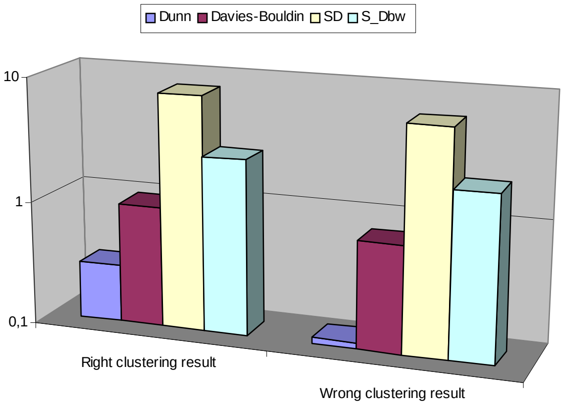

Figure 5
Clustering results on third data set

Conclusions

In this paper several cluster validity indices have been summarised. These validity indices have been evaluated with several different input dataset and we tried to compare the efficiency of these validity indices. The result of this comparison is that these indices can identify only the well separated hyper sphere shaped clusters. As these indices measure the variance of the clusters around some representative points but some clusters, especially the arbitrary shaped clusters, do not have representative centre point. Thus it is important to define novel validity indices which can measure arbitrary shaped clusters.

Acknowledgment

This work has been supported by the Mobile Innovation Center, Hungary.

References

[1] A. K. Jain, M. N. Murty and P. J. Flynn: Data clustering: a review, ACM Computing Surveys, Vol. 31, No. 3, pp. 264-323, 1999

[2] M. Halkidi, Y. Batistakis and M. Vazirgiannis: Cluster validity methods: part I, SIGMOD Rec., Vol. 31, No. 2, pp. 40-45, 2002

[3] M. Halkidi, Y. Batistakis and M. Vazirgiannis: Cluster validity methods: part II, SIGMOD Rec., Vol. 31, No. 3, pp. 19-27, 2002

[4] S. Guha, R. Rastogi and K. Shim: CURE: an efficient clustering algorithm for large databases, Proc. of ACM SIGMOD International Conference on Management of Data, pp. 73-84, 1998

[5] M. J. A. Berry and G. Linoff: Data Mining Techniques for Marketing, Sales and Customer Support, John Wiley & Sons, Inc., 1996

[6] S. Theodoridis and K. Koutroubas: Pattern Recognition, Academic Press, 1999

[7] J. C. Dunn: Well Separated Clusters and Optimal Fuzzy Partitions, Journal of Cybernetica, Vol. 4, pp. 95-104, 1974

[8] D. L. Davies and D. W. Bouldin: Cluster Separation Measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 1, No. 2, pp. 95-104, 1979

[9] Subhash Sharma: Applied multivariate techniques, John Wiley & Sons, Inc., 1996

[10] 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

[11] M. Halkidi and M. Vazirgiannis: Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set, Proc. of ICDM 2001, pp. 187-194, 2001

[12] 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

[13] N. R. Pal and J. Biswas: Cluster Validation using graph theoretic concepts, Pattern Recognition, Vol. 30, No. 4, 1997