Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp372-377)

Cluster Validity Measurement for Arbitrary Shaped Clusters

FERENC KOVÁCS AND RENÁTA IVÁNCSY
Department of Automation and Applied Informatics
Budapest University of Technology and Economics
Goldmann Gyorgy ter 3
1111 Budapest
HUNGARY

Abstract: - Clustering is an unsupervised process in 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. In this pater a novel validity measurement index is introduced which can be used for evaluating arbitrary shaped clusters. The main advantage of this validity index are the following: it can compare and measure not only elliptical clusters but arbitrary shaped clusters as well. This validity index can evaluate the result of any clustering algorithms but the calculation of this index is very simple in case of density based algorithms as it can be calculated during the clustering process.

Key-Words: -clustering algorithms, arbitrary shaped clusters, density based clustering, cluster validity, validity indices

1 Introduction

One of the best known problem in the data mining is clustering. Clustering is the task of categorizing 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 dissimilar. 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 procedure thus evaluation process 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 whether the founded 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 (e.g.: variance). In fact there is no representative point in an arbitrary shaped cluster hence these validity indices cannot measure them properly.

The rest of the paper is organized as follows. Section 2 gives a general overview of clustering techniques and cluster validation methodes. Then the most commonly used cluster validity indices are introduced in Section 3. The fundamental density based clustering algorithms are described in Section 4. Section 5 introduces a novel cluster validity index that can measure arbitrary shaped cluster result. Finally some conclusion is given in Section 6.

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 classify into the following groups [1]:

- *Partitional Clustering:* These algorithms decompose directly the data set into a set of disjoint clusters (called partitions). They attempt to determine an integer number of partitions that optimize a certain criterion function. This optimization 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 they can find arbitrary shaped clusters.

- *Grid-based Clustering:* These type of algorithms are mainly proposed for spatial data mining. They quantize the search space into finite number of cells.

The result of an clustering algorithms can be very different from each other on the same data set as the other input parameters of the algorithms can extremely modify the behavior and execution of the algorithm. The aim of the cluster validity techniques 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 visualization and visual validation are not trivial therefore some formal methods are needed.

The procedure of evaluating the results of a clustering algorithms is called cluster validity. Two measurement criteria have been proposed for evaluating and selecting the 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 of the cluster.
- *Separation:* The clusters themselves should be widely separated. A simple measure for cluster separation is the cluster distance. 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 center 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 bases of the internal criteria are some metrics which are based on data set and the clustering schema. The main disadvantage of these two methods is their computational complexity.

The basis of the relative criteria is the comparison of the different clustering schema. The clustering algorithm is 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 relative criteria keep the possibility to compare clustering results independently of the clustering algorithms. 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

The validity indices are used for compering "goodness" of a clustering result to others which were created by other clustering algorithms, or by the same algorithms 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 j^{th} dimension
$\|X\|$	$\sqrt{X^T X}$, where X is column vector
n_{ij}	Number of element in the i^{th} cluster in the j^{th} dimension
n_j	Number of element in in the j^{th} dimension in the whole data set
v_i	Center point of the i^{th} cluster
c_i	i^{th} cluster
$\|c_i\|$	Number of element in the c_i cluster

Table 1. Notation in validity indices

3.1 Dunn and Dunn-like indices

These cluster validity indices were introduced in paper [7]. Equation 1 shows the definition of Dunn index where $d(c_i, c_j)$ is the dissimilarity function between two clusters and defined as $d(c_i, c_j) = \min_{x \in C_i, y \in C_j} (d(x,y))$. The diameter of a cluster can be defined in the following way: $diam(c_i) = \max_{x,y \in C_i} (d(x,y))$.

$$D = \min_{i=1,..,n_c} \left\{ \min_{j=i+1,...,n_c} \left\{ \frac{d(c_i,c_j)}{\max_{k=1,...,n_c} (diam(c_k))} \right\} \right\} \quad (1)$$

If a data set contains well-separated clusters, the distance between clusters is usually large and the di-

ameter of the clusters is 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 [13] [6]. These indices use different definition for cluster distance and cluster diameter.

3.2 Davies-Bouldin Validity 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 $R_{ij} > R_{ik}$

Equation 2 shows the usual definition of the cluster similarity measure (R_{ij}).

$$\begin{aligned} R_{ij} &= \frac{s_i + s_j}{d_{ij}} \\ d_{ij} &= d\left(v_i, v_j\right) \\ s_i &= \frac{1}{\|c_i\|} \sum_{x \in c_i} d\left(x, v_i\right) \end{aligned} \tag{2}$$

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 clustering result. The formal definition of this index is described by Equation 3.

$$\begin{aligned} DB &= \frac{1}{n_c} \sum_{i=1}^{n_c} R_i \\ R_i &= \max_{j=1 \ldots n_c, j \neq i} (R_{ij}),\ i{=}1 \ldots n_c \end{aligned} \tag{3}$$

3.3 SD Validity Index

The bases of the 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 data set, thus it can measure the homogeneity and compactness of the clusters, as well. The variance of the data set and variance of a cluster are defined as follows:

$$\begin{array}{ll} \text{Variance of the dataset :} & \text{Variance of a cluster :} \\ \sigma_x^p = \frac{1}{n} \sum_{k=1}^{n} \left(x_k^p - \overline{x^p}\right)^2 & \sigma_{v_i}^p = \frac{1}{\|c_i\|} \sum_{k=1}^{n} \left(x_k^p - \overline{v_i^p}\right)^2 \end{array}$$

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

The average scattering definition is given by Equation 4. This can measure the average compactness of the clusters as it compares the variance of clusters and variance of the data set.

$$Scatt = \frac{1}{n_c} \sum_{i=1}^{n_c} \frac{\|\sigma\left(v_i\right)\|}{\|\sigma\left(x\right)\|} \tag{4}$$

The total separation of clusters is based on the distance of cluster center points thus it can measures the separation of clusters. Its definition is given by Equation 5

$$Dis = \frac{\max\limits_{i,j=1 \ldots n_c} (\|v_j - v_i\|)}{\min\limits_{i,j=1 \ldots n_c} (\|v_j - v_i\|)} \sum_{k=1}^{n_c} \left(\sum_{\substack{j=1, \\ i \neq j}}^{n_c} \|v_j - v_i\| \right)^{-1} \tag{5}$$

The SD index can be defined based on Equation 4 and 5 as follows

$$SD = \alpha \cdot Scatt + Dis \tag{6}$$

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.

3.4 S_Dbw Validity Index

The S_Dbw 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 4. The inter - cluster density is defined as follows:

$$Dens\_bw = \frac{1}{n_c\left(n_c - 1\right)} \sum_{i=1}^{n_c} \left(\sum_{\substack{j=1, \\ i \neq j}}^{n_c} \frac{ds(u_{ij})}{\max\left\{ds(v_i), ds(v_j)\right\}} \right) \tag{7}$$

where u_{ij} is the middle point of the line segment that is defined by the v_i andv_j clusters centers. The density function (ds) around a point is defined as follows: it counts the number of points in a hyper-sphere whoose radius is equal to the average standard deviation of clusters. The average standard deviation of clusters is defined as

$$stdev = \frac{1}{n_c}\sqrt{\sum_{i=1}^{n_c} \|\sigma(v_i)\|} \tag{8}$$

The S_Dbw index is defined in the following way:

$$S\_Dbw = Scatt + Dens\_bw \tag{9}$$

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 result.

4 Density Based Algorithms

The main advantage of the density based algorithms is that they can discover arbitrary shaped clusters. These algorithms investigate the local environment of the data points and they try to find dense areas in the data set. Several density based algorithms have been developed however DBSCAN [14] and DENCLUE [15] are the fundamental ones. These fundamental algorithms use different approaches for defining the density of a data set.

4.1 DBSCAN algorithm

The basis of the DBSCAN algorithm is the definition of density based connectivity. The algorithm has two input parameters: ϵ and MinPts. These input parameters are used for defining the density based connectivity in the following way:

- ϵ *neighborhood of a point:* The ϵ neighborhood of an point is the set of points whose distance is smaller than the given ϵ parameter.
- *Core object:* Core object is a data point whose ϵ neighborhood contains at least MinPts element.
- *Density reachable:* A point y density-reachable from a core object x if a finite sequence of core objects between x and y exists such that each next belongs to an - neighborhood of its predecessor.
- *Density connectivity:* Two points x, y are in density connectivity if they are density-reachable from a common core object

The defined density-connectivity is a symmetric relation and all the points reachable from core objects can be factorized into maximal connected components serving as clusters. The points that are not connected to any core point are declared to be outliers (they are not covered by any cluster).

The DBSCAN algorithm identify the core points of the data set than it grows the clusters from the core points. The main advantage of this algorithm is that it investigates the local environment of the data points thus it can provide not only elliptical clusters but arbitrary shaped clusters as well. On the other hand the disadvantage of this algorithms is that it is very sensitive to its input parameters.

4.2 DENCLUE algorithm

The basis of the DENCLUE algorithm is the density function of the data points. Each data point has own influence (effect to its environment) and the global density of the data set is the superposition of the influence functions of the data elements. Formally, the density function in a data point can be defined in the following way:

$$f^D(x) = \sum_{y \in D} f(x, y) \tag{10}$$

The influence function($f(x, y)$) can be defined in several way e.g: square wave function, Gaussian influence function etc. The DENCLUE algorithm concentrates on local maximums of density functions called density-attractors and uses a flavor of gradient hill-climbing technique for finding them. The local maximums define the clusters and a data element is put into a cluster if a local maximum directly can be reached by hill-climbing techniques. The found local maximum must be larger than a given input parameter (ξ) hence this parameter keep the possibility to define the outlier points. This algorithm is very sensitive to its input parameters, as well and the influence functions can have multiple input parameters and ,of course, the ξ parameter can extremely modify the result of the algorithm.

5 Validity Index for Arbitrary Shaped Clusters

As mentioned in the previous section the density based algorithms are very sensitive to their input parameters. The main disadvantage of the introduced validity indices is that they need reference points of the clusters and calculates parameters based on these reference points. As the main part of the clustering algorithms provides elliptical clusters this behavior of the validity indices was not disturbing. But the density based algorithms can provide arbitrary shaped clusters thus these validity indices cannot measure properly the clustering quality of the density based algorithms.

5.1 Variance of the nearest neighbor distance

The main problem of the current validity indices is that they do not investigate the local environment of the data points they only take into consideration the global reference points of the clusters. The novel approach is to investigate the local environment of a data element, formally, the deviation of the nearest neighbor distances is investigated in every cluster. This validity index is based on the variance of the nearest neighbor in a cluster, this can be defined as follows:

$$\begin{aligned} d_{min}(x_i) &= \min_{y \in C_i}(d(x_i,y)) \\ \overline{d_{min}(C_i)} &= \frac{\sum_{x_i \in C_i} d_{min}(x_i)}{\|C_i\|} \\ V(C_i) &= \frac{1}{\|C_i\|-1} \sum_{x_i \in C_i} \left(d_{min}(x_i) - \overline{d_{min}(C_i)}\right)^2 \end{aligned} \quad (11)$$

Based on Equation 11 it is possible to define novel validity index, called Variance of the Nearest Neighbor Distance (VNND):

$$VNND = \sum_{i=1}^{n_c} V(C_i) \quad (12)$$

This validity index measures the homogeneity of the clusters. Lower index value means more homogenous clustering. This validity index does not use global representative points thus it can measure arbitrary shaped clusters, as well. This validity index can evaluate results of any clustering algorithm but in some cases the calculation of the index can be time consuming. But in case of density based algorithms this validity index can be calculated during the clustering process. As these algorithms investigate the local environment of the data element during the clustering hence it is possible to find easily the nearest neighbor of a point.

Though the calculation of this validity index can be time consuming if the clusters contains huge number of elements. In this case this index can be estimated by sampling data points from the clusters.

5.2 Experimental Result

The cluster validity indices were evaluated by synthetically generated data sets. The data sets were generated by our synthetical data generator. Figure 1 depicts the generated data sets. The main properties of the generated data sets are the following:

- *DS 1 - Well separated clusters:* the cluster elements were generated around the cluster centers points using normal distribution
- *DS 2 - Ring shaped clusters:* two cluster, which contains each other
- *DS 3 - Arbitrary shaped clusters:* some arbitrary shaped clusters close to each other

We used DBSCAN and k-mean algorithms during the evaluation process. The DBSCAN algorithm was able to find the right clustering schema in all cases. The k-mean algorithm was unable to provide the right clustering result during the evaluation of the second and third data set. Table 2 shows the offered number of clusters based on the clustering result. It is very important to notice although the k-mean algorithm provide wrong clustering result in case of second and third data set the validity indices can compare the results and offer a result.

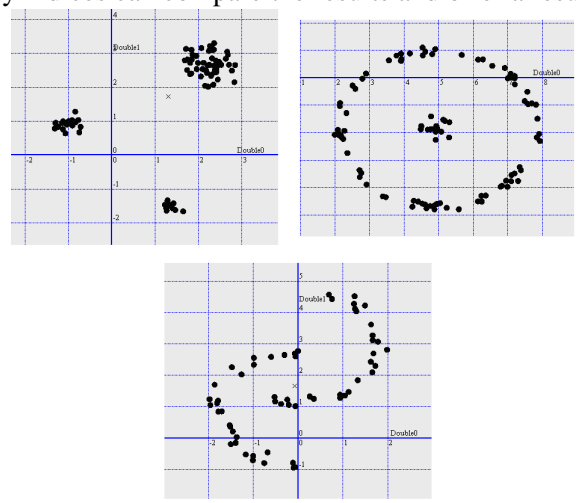

Figure 1. The used data set in experimental evaluation

DS 1		Dunn	SD	S_Dbw	VNND
	k-mean	3	3	3	3
	DBSCAN	3	3	3	3
DS 2					
	k-mean	4	3	3	3
	DBSCAN	2	2	2	2
DS 3					
	k-mean	3	2	2	2
	DBSCAN	2	3	2	2

Table 2. The offered number of cluster

Table 3 shows more interesting result. If we compare the clustering results independently of the algorithms there are several cases when the traditional validity indices make wrong decision and choose wrong clustering result. But the novel VNND index can compare the clustering result better and it does not make any mistake during the comparison process.

	Dunn	SD	S_Dbw	VNND
DS 2	DBSCAN 2	k-mean 2	DBSCAN 2	DBSCAN 2
DS 3	k-mean 3	DBSCAN 2	k-mean 2	DBSCAN 2

Table 3. The globally optimal clustering result

6 Conclusion

In this paper the most commonly used cluster validity indices have been investigated and a novel one has been introduced. The general disadvantage of the current validity indices is that they are unable to measure non elliptical clusters properly as they work global representative points of the clusters. The introduced novel validity index does not need global representative points as it investigates the local environment of the data elements. During the evaluation process only the VNND validity index was able to measure correctly the clustering results and it was able to choose the right configuration. However the calculation of this validity index can time consuming but in case of density based algorithm it can be calculated easily.

7 Acknowledgement

This work has been supported by the Mobile Innovation Center Hungary, the fund of the Hungarian Academy of Sciences for control research and the Hungarian National Research Fund (grant number: T042741).

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, "Clustering validity checking 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," in *ACM SIGMOD International Conference on Management of Data*, pp. 73–84, June 1998.

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

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

[7] J. Dunn, "Well separated clusters and optimal fuzzy partitions," *Journal of Cybernetica*, vol. 4, pp. 95–104, 1974.

[8] D.L. and D. Bouldin, "Cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 224–227, 1979.

[9] S. Sharma, *Applied multivariate techniques*. New York, NY, USA: 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," in *ICDM*, pp. 187–194, 2001.

[12] M. Halkidi, M. Vazirgiannis, and Y. Batistakis, "Quality scheme assessment in the clustering process," in *PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, (London, UK), pp. 265–276, Springer-Verlag, 2000.

[13] J. B. N. R. Pal, "Cluster validation using graph theoretic concepts," *Pattern Recognition*, vol. 30, no. 4, 1997.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Second International Conference on Knowledge Discovery and Data Mining* (E. Simoudis, J. Han, and U. Fayyad, eds.), (Portland, Oregon), pp. 226–231, AAAI Press, 1996.

[15] A. Hinneburg and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," in *Knowledge Discovery and Data Mining*, pp. 58–65, 1998.