Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp388-393)

Cluster Validity Measurement Techniques

CSABA LEGÁNY, SÁNDOR JUHÁSZ AND ATTILA BABOS

Department of Automation and Applied Informatics and HAS-BUTE Control Research Group
Budapest University of Technology and Economics
1111 Goldmann Gy. ter 3., Budapest
HUNGARY

Abstract: - Clustering is a process of discovering groups of objects such that the objects of the same group are similar, and the objects belonging to different groups are dissimilar. Several research fields deal with the problem of clustering: for example pattern recognition, data mining, machine learning. A number of algorithms exist that can solve the problem of clustering, but most of them 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 whether a clustering result is acceptable or not, thus several clustering validity techniques and indices have been developed. This paper deals with the problem of clustering validity. The most commonly used validity indices are introduced and explained, and they are compared based on experimental results.

Key-words: - Data mining, Clustering algorithms, Cluster validity, Validity indices

1 Introduction

One of the best known problems in the field of data mining is clustering. The problem of clustering is to partition a data set into groups (clusters) in such a way that the data elements within a cluster are more similar to each other than data elements in different clusters [1]. 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 [2].

Clustering is mostly unsupervised process, thus evaluating the result 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 cluster configuration found during the process is acceptable or not. Several clustering validity approaches have been developed [3].

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. Conclusion can be found in Section 5.

2 Related Work

There are different types of clustering algorithms and they can be classified into the following groups [2]:

- *Partitional Clustering*: These algorithms decompose the data set directly 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 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 aim of cluster validity is to find the partitioning that best fits the

underlying data. In most cases one or more clustering algorithms are executed multiple times with different input parameters on the same data set. Validity indices can be used in order to select the best clustering schema from the different results. Several validity indices were developed and introduced in various works [5][6][7][8][9].

Table 1 summarizes the commonly used notations of validity indices. Most widely used validity indices are introduced in the following section.

	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^T is a column vector
n_{ij}	Number of element in i[th] cluster j[th] dimension
n_j	Number of element in j[th] dimension in the whole data set
v_i	Centre point of the i[th] cluster
c_i	i[th] cluster
$\|c_i\|$	Number of element in the i[th] cluster

Table 1 Notation in validity indices

3 Validity Indices

In this section several validity indices are introduced. These indices are used for measuring the "goodness" of a clustering result compared to other ones that were created by other clustering algorithms, or by the same algorithm but using different input parameter values. These indices are usually suitable for measuring crisp clustering, where no overlapping between partitions is allowed.

3.1 Dunn and Dunn like indices

These cluster validity indices were introduced in [5]. The index definition is given by Equation 1.

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

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

The Dunn index compares the minimal cluster distance to the maximal cluster diameter. 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 followings: 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 were proposed [4]. These indices use different definitions for cluster distance and diameter.

3.2 Davies – Bouldin index

The Davies – Bouldin index [6] 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 the clusters (R_{ij}) can be defined freely but it has to satisfy the following conditions [6]:

- $R_{ij} \geq 0$
- $R_{ij} = R_{ji}$
- 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 $R_{ij} > R_{ik}$

In most cases cluster dispersion measure is the average distance to the cluster centre. Usually R_{ij} is defined in the following way (Equation 2):

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$
$$d_{ij} = d(v_i, v_j), \; s_i = \frac{1}{\|c_i\|}\sum_{x\in c_i} d(x, v_i) \quad (2)$$

The Davies–Bouldin index finds out for every cluster which cluster it is the most similar to. After it summarizes the maximum cluster similarities to create a single index DB (Equation 3):

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i, \text{ where}$$
$$R_i = \max_{j=1\ldots nc, i\neq j}(R_{ij}), \; i = 1\ldots n_c \quad (3)$$

If the Davies – Bouldin index is low, the clusters are not very similar to each other, which means that they are compact and well-separated.

3.3 SD validity index

The base measurements of SD validity index [10] are the average scattering and total separation of clusters. The scattering is given by calculating the variance of the clusters and the variance of the complete 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 4.

Variance of the dataset: Variance of a cluster:

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

$$\sigma(x) = \begin{bmatrix} \sigma_x^1 \\ \vdots \\ \sigma_x^d \end{bmatrix} \qquad \sigma(v_i) = \begin{bmatrix} \sigma_{v_i}^1 \\ \vdots \\ \sigma_{v_i}^d \end{bmatrix} \qquad (4)$$

The average scattering for clusters is defined as

$$Scatt = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{\|\sigma(v_i)\|}{\|\sigma(x)\|} \qquad (5)$$

If clusters are compact, the variance of the clusters is expected to be smaller than the variance of the dataset, thus the Scatt measure is low. 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 6.

$$Dist = \frac{\max_{i,j=1..n}\left(\|v_j - v_i\|\right)}{\min_{i,j=1..n}\left(\|v_j - v_i\|\right)}\sum_{t=1}^{k}\left(\sum_{z=1,z\neq t}^{k}\|v_t - v_z\|\right)^{-1} \qquad (6)$$

The SD index can be defined based on Equation 5 and 6 as follows

$$SD = \alpha Scatt + Dist \qquad (7)$$

where α is a weighting factor that is equal to Dist parameter in case of maximum number of clusters.

Lower SD index means better cluster configuration as in this case the clusters are compacts and well-separated. We will use the inverse of the SD index in this article.

3.4 S_Dbw validity index

This validity index has been proposed in [9]. 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

$$Dens\_bw = \frac{1}{n_c(n_c-1)}\sum_{i=1}^{n_c}\left(\sum_{\substack{j=1,\\ i\neq j}}^{n_c}\frac{density(u_{ij})}{\max\{density(v_i), density(v_j)\}}\right) \qquad (8)$$

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

$$stdev = \frac{1}{n_c}\sqrt{\sum_{i=1}^{n_c}\|\sigma(v_i)\|} \qquad (9)$$

The S_Dbw index is defined in the following way:

$$S\_Dbw = Scatt + Dens\_bw \qquad (10)$$

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. We will use the inverse of S_Dbw index in this article. (7)

4 Index comparison

4.1. Experimental Results

The clustering algorithms and validity indices were evaluated with synthetical data set generated by a data set generator implemented by our research group. The validity indices were evaluated using the following datasets that are also depicted on Figure 1:

- Well-separated clusters: the cluster elements were generated around picked cluster centre points using normal distribution
- Ring shaped clusters: two clusters containing each other.

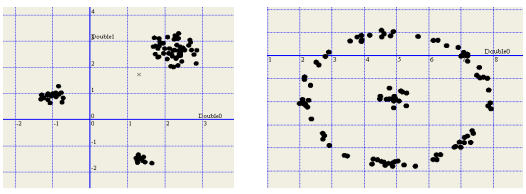

Figure 1 The used data sets in experimental evaluation

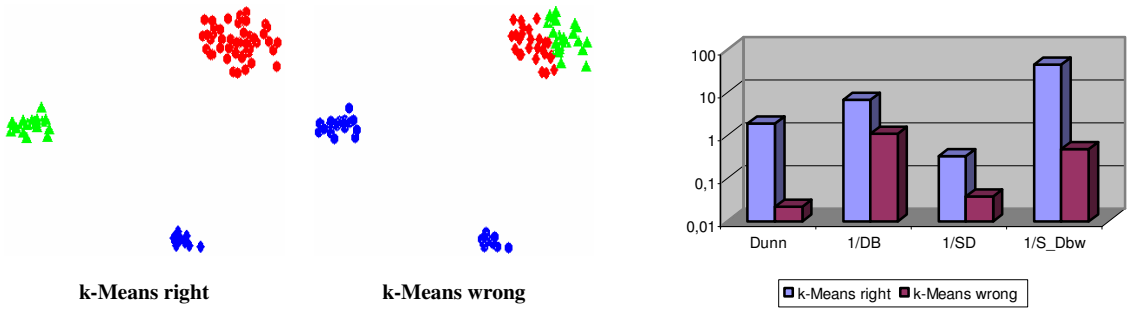

Figure 2 Validity indices on the first data set

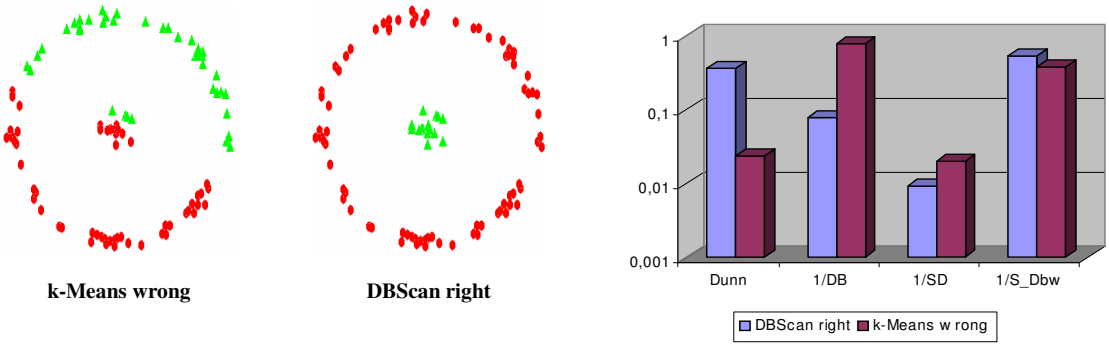

Figure 3 Validity indices on the second data set

Figure 2 shows two clustering results of the k-Means algorithm started from different random starting points on the first data set. The different index values are depicted on this figure as well. In this case it is easy to identify that the validity indices can properly compare the results of the clustering algorithm.

Figure 3 compares the clustering results of the k-Means and the DBScan algorithm. The different index values are depicted on this figure as well. The result is a little bit surprising as the Dunn and S_Dbw index can identify the right clustering result but the other indices offer wrong decision. The main disadvantage of the current validity indices is that they cannot identify the right clustering schema unless the clusters are well separated.

4.2. Runtime comparison

Four validity indices have been compared on the same data set. The data set contained a few randomly created groups of points. The number of points was increased by adding further points to each group. Figure 4 displays the runtime of four indices for different number of points. All four indices have a complexity of $O(n^2)$, and the most time-consuming to compute is the Dunn index.

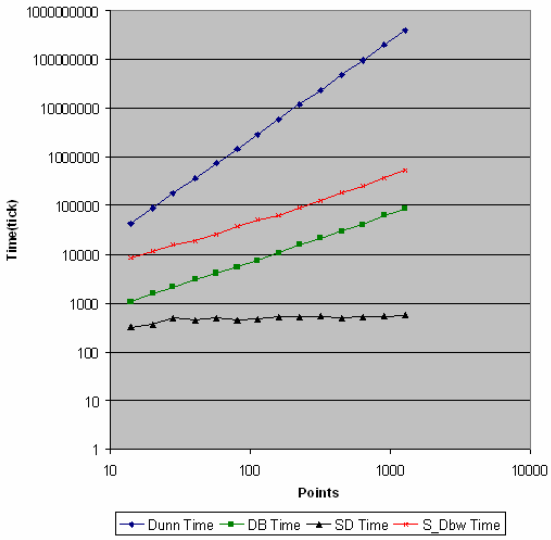

Figure 4 Runtime comparison of several validity indices

4.3. Finding optimal clustering results with indices

In this section three indices will be compared by the ability of finding proper clusters. Figure 5 shows the sample dataset.

Figure 5 The data set used in optimal cluster finding

The DBScan algorithm was used for index comparison, which is very sensitive to its input parameters (Eps and MinPts). Eps was chosen to be between 0 and 0.02, MinPts was between 0 and 19.

Indices favour compact and well-separated clusters. However, clusters containing only few points also satisfy these conditions, which means that every index will favour such clusters. A possible solution to this problem is to omit such cases, when every cluster is very small (i.e. only contain one or two points).

Figure 6 compares three indices. In all cases, the plateau (Eps is between 0.09 and 0.15, MinPts is between 0 and 8) represents right clustering results, while the peak represents a clustering, in which every point belongs to a separate cluster. The surface diagram of DB contains a lot of false peaks, while the other surface have only one.

Figure 7 compares clustering results belonging to different points (e.g. Eps,MinPts pairs) of the Dunn-surface. Clusterings based on plateau points are considered to be right, while others belonging to peak points are wrong.

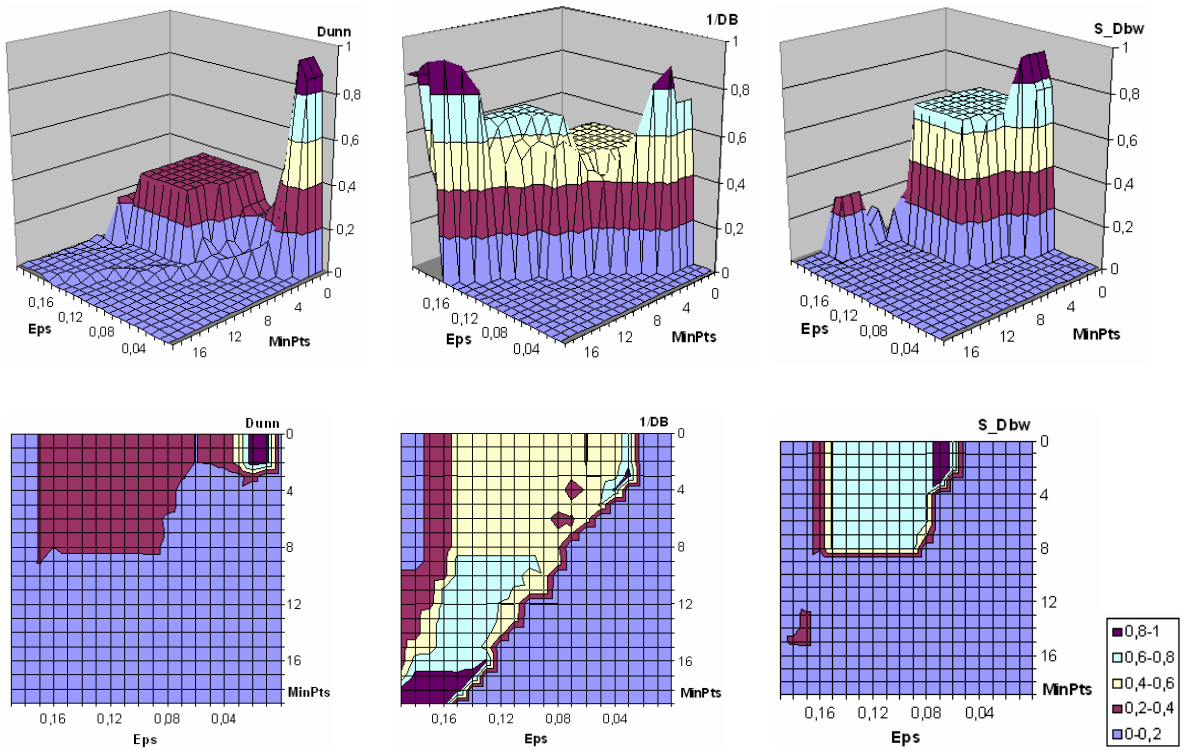

Figure 6 Comparison of clustering results with Dunn, Davies-Bouldin and S_Dbw indices

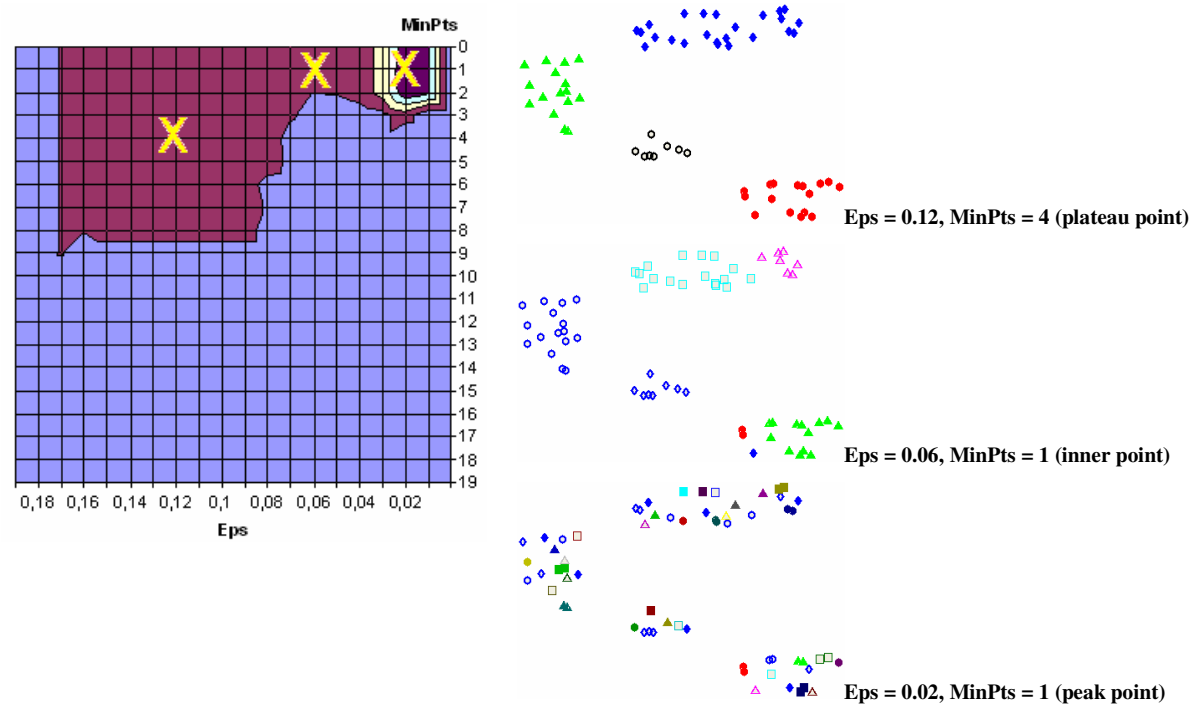

Figure 7 Clustering results belonging to plateau, inner and peak points

5 Conclusions

In this paper several cluster validity indices have been summarised. These validity indices have been evaluated with various different input dataset and we tried to compare the efficiency of these validity indices. The result of this comparison shows Dunn and S_Dbw are able to find not well-separated clusters, while the others cannot identify them. All of them have a complexity of $O(n^2)$, the Dunn is the most time-consuming, while the SD is the fastest. Indices are able to compare clustering results to find the best one, but they favour the very small clusters, so peaks should be eliminated from their surface diagrams. In fact, the surface diagram of DB index contained a lot of peaks. Taking all these factors into consideration, we would recommend the use of S_Dbw index. These indices measure the variance of clusters around some representative points, but arbitrary shaped clusters do not have representative centre points. 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] 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

[2] 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

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

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

[6] 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

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

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

[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

[11] 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

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