A Survey on Internal Validity Measure for Cluster Validation

L.Jegatha Deborah[1], R.Baskaran[2], A.Kannan[3]

[1]Department of Computer Science & Engineering, Anna University of Technology, Chennai -25

blessedjeny@gmail.com

[2]Department of Computer Science & Engineering, Anna University, Chennai -25

baaski@annauniv.edu

[3]Department of Information Technology, Anna University, Chennai -25

akannan@annauniv.edu

Abstract

Data Clustering is a technique of finding similar characteristics among the data set which are always hidden in nature and grouping them into groups, called as clusters. Different clustering algorithms exhibit different results, since they are very sensitive to the characteristics of original data set especially noise and dimension. The quality of such clustering process determines the purity of cluster and hence it is very important to evaluate the results of the clustering algorithm. Due to this, Cluster validation activity had been a major and challenging task. The major factor which influences cluster validation is the internal cluster validity measure of choosing the optimal number of clusters. The main objective of this article is to present a detailed description of the mathematical working of few cluster validity indices and not all, to classify these indices and to explore the ideas for the future promotion of the work in the domain of cluster validation. In addition to this, a maximization objective function is defined assuming to provide a cluster validation activity.

Keywords: *Data clustering, cluster, cluster purity, cluster analysis, cluster validation, cluster validity indices.*

1. Introduction

Data Exploration is finding the hidden knowledge from the data sets. Data clustering is a tool which aids the data exploration process. Some of the data sets have natural groupings in them, whereas some others have to undergo the process of clustering in identification of specific groups. Data Clustering is a technique of partitioning the data set without known prior information. It finds its use in most of the applications where unsupervised learning occurs. A wide range of clustering algorithms is available in the market for grouping low dimensional data and data of higher dimensions. The different kinds of clustering algorithms when used for varying data sets produces different kinds of results based on the initial input parameters, environment conditions, nature of data set. In such a scenario, since there are no predefined classes or groups known in clustering process, it had been always an issue in finding an appropriate metric for measuring if found cluster configuration, number of clusters, cluster shapes, etc is acceptable or not.

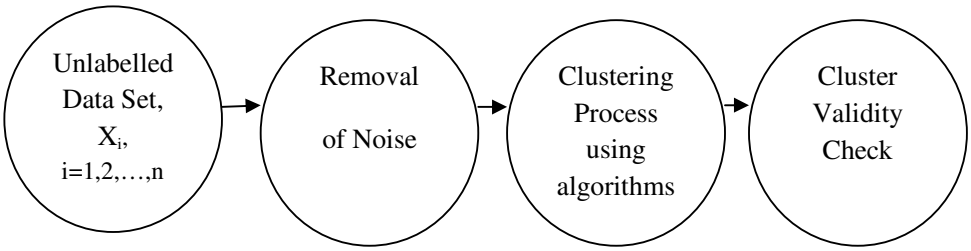

Fig. 1. Data Exploration

To resolve the above issue, Cluster analysis and in particular cluster validity issues have to be given much more attention. Three different techniques are available to evaluate the clustering results: External, Internal, and Relative. Most of the traditional cluster validity indices which are present could not work well on arbitrary shaped clusters. This is because of the fact that, these validity measures choose representative points from each clusters like mean values, they are found to be handicapped when other cluster structures are found. Two major parameters for the cluster validity indices are present for evaluation: Compactness and Separation. Both these metrics consider some data points which act as representatives of each cluster. Based on the three different techniques for evaluation as described earlier, the internal metric for evaluation is based on some metrics based on the input data set and the underlying clustering schema. In particular, the internal metric deals about the optimal number of clusters formed. Unfortunately, the internal metric has high computation complexity, because this evaluation technique considers all the data points within the cluster structure. Considering the internal metric for evaluation, the commonly used estimation measurements are compactness and separability, as defined below.

1.1 Cluster Validity Index Assessment Measures

Table 1. Properties of Cluster Validity Indices

Sl. No	Parameter in CVI	Measuring Principle	Definition	Implementation
1	Compactness	Intra-Cluster distance	1.Intra-Cluster distance measures Compactness. 2.The sum of the distances of the objects within the same cluster is minimized.	Summation/Minimum/Maximum /Average the distance between. 1. All pairs of point within the cluster. 2. Between centroid and all pairs of point within the cluster.

2	Separability	Intra-Cluster distance	1.Intra-Cluster distance measures Separability. 2.The distance between any two cluster are maximized.	1. Sum of square distance between all pairs of cluster. [Distance indicates distance between centroid]
3	Exclusiveness ***(Proposed Metric)***	Probability density function	1.Probability density function measures irregularity. 2.All data values tend to cluster towards the mean value.	1. Apply Gaussian normal distribution functions in identifying outlier data.
4	Incorrectness ***(Proposed Metric)***	Loss function	1.Calculation of risk factor measures the degree of accuracy. 2.Risk factor should be minimized.	1. Calculate median value for each cluster. 2. Apply loss functions to calculate the risk percentage.

1.2 Defining an Maximization Objective Function

Considering the above major parameters to be used in cluster validity indices, an objective function is defined as follows. Choosing an optimal cluster validity index will depend on whether the above parameters are minimized or maximized. The appropriate cluster validity index for several application domains will be found based on maximizing the following objective function.

The objective function will take into account all the four parameters mentioned above, which is a combination of minimizing and maximizing various parameters. The objective function is generalized and defined for a data set of 'n' number of dimensions. The verbal function, on considering a single cluster, is defined as follows.

Objective Function (OBF) =

Min (Compactness) + Max (Separability) + Max (Exclusiveness) + Min (Incorrectness)

Mathematically the objective Function (OBF) defined for a cluster, r (consisting of n-dimensional data set) is,

$$OBF = \left\{ Min\left[C = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left\| x_{ij} - y_{ij} \right\|^2 \right] + Max\left[S = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left\| c_i - c_j \right\|^2 \right] + Max\left[\frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)} \right] + Min\left[I = E\left(L(x, \mu) \right) \right] \right\} \quad (1)$$

where,

$x_i \in X_i$ [i=1,2,3,…N)

$y_i \in Y_i$ [i=1,2,3,…,N)

x_{ij}, y_{ij} = data point coordinates of the i^{th} cluster in the j^{th} dimension

$\mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_n)$

k = total no. of dimensions of a data set

E() = Expected value of a function

C = Compactness Measure

S = Separability Measure

Ex = Exclusiveness Measure

I = Incorrectness Measure

n = total number of data items in the i^{th} cluster [i=1,2,3,…N]

N = Total number of Clusters

c_i, c_j = computer centroid values of cluster i and j

σ, Σ = Variance value of the entire data set

1.3 Vital Issues to be considered

The structure of the clusters obtained using any type of clustering algorithms is very important. All the evaluation parameters for cluster validation basically depends on intra-cluster distance metric, inter-cluster distance metric, mean, median, variance values. Hence, most of the existing indices are based on geometrical-based theory and only very few are based on statistics. The two new parameters included: Exclusiveness and Incorrectness are based on statistical theory

and estimation theory. The initial important issue to be considered is that none of the existing traditional validity indices perform very well on outlying data which are found to be a noisy data in most of the application domains. These noisy data have to be rejected, since their inclusion may invalidate the outcome results. Hence, in the objective function, the degree of excluding the outlier data must be maximized. The second important issues to be considered is that the impact on the degree of loss in a data item to be wrongly projected in a cluster, which is supposed to be not in its original cluster. This degree of wrongness is computed using the loss functions based on Statistical Theory and the available risk factor percentage is computed for each of the obtained clusters. Hence, for this parameter, in the objective function defined, the percentage of risk of a cluster should be minimized. The acceptable risk factor percentage shall be fixed with respect to the individual application domain. The threshold percentage for such domain shall be fixed by conducting several empirical tests on enormous, related data sets of the application domain.

1.4 This Paper

In this survey paper, Section 2 quickly surveys the problems and solutions with regard to the evaluation parameters of the cluster validity indices in several application domains. Section 3 finally gives the authors concluding remarks statements and the future research activities. Fig. 2 provides a clear picture of the complete stages handled in this survey paper.

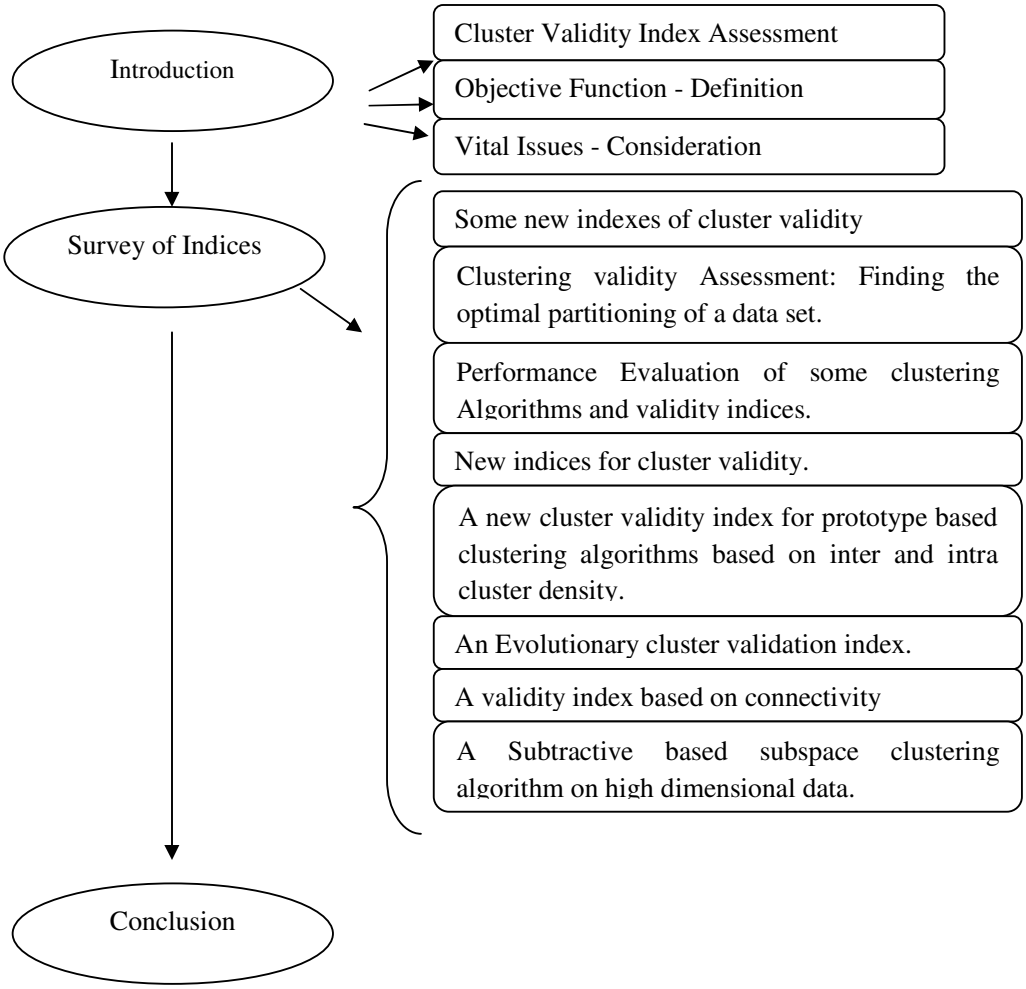

Fig. 2. Paper Overview

2. A Quick Study and Detailed Analysis of Some Indices

Hypothesis: Minimal Intra-Cluster and Maximal Inter-Cluster Distance

2.1 Some New Indexes of Cluster Validity [1]

Dunn's Index is based on geometrical considerations that have the basic rationale of designing a criterion to identify sets of clusters that are compact and well separated. To understand this index let S and T be non-empty sets of $\Re^p$ and let d: $\Re^p \times \Re^p \mapsto \Re^+$ be any metric on the non-empty sets. The standard definitions of the diameter Δ of S and the set distance δ between S and T are

$$\Delta(s) = \max_{x,y \in s}\{d(x,y)\} \text{ and}$$
$$\delta(S,T) = \min_{x \in S, y \in T}\{d(x,y)\} \qquad (2)$$

The quantity $\Delta(s)$ measures the distance between clusters directly on the points in the cluster. The latter $\delta(S,)$ measures the distance between two clusters based on their centroid values. Dunn's defined an index assuring compactness and separability relative to distance d, called as Compact Well Separated Clusters, if only if the following property is satisfied: for all s,q,r with q ≠ r, any pair of points $x, y \in Xs$, y∈conv(X_s) are closer together as measured by d, than any pair u,v , u∈ X_s and v ∈ conv(X_s), where conv(S) is the convex hull [2-3] of S in $\Re^p$. Dunn's index for Compact Well Separated Clusters is obtained by the equation

$$V_D(S) = \max_{1 \le i \le c}\left\{\min_{1 \le j \le c}\left\{\frac{\delta(X_i, conv(X_j))}{\max_{1 \le k \le c}\{\Delta(X_k)\}}\right\}\right\} \qquad (3)$$

where $i \neq j$. However, Dunn proved that data set X can be partitioned into CWS clusters relative to the distance d, if and only if $\max_{0 \le x \le 1}\{V_D(U)\} > 1$. V_D sets very attractive geometrical requirements for good CWS clusters.

Intricacy

The estimation of various parameters is found to be very difficult in terms of computation complexity. The order of magnitude increases sharply on addition or deletion of new data items in a data set. However, estimation of conv(X_j) in the above equation, for even lower dimension data set in computationally very expensive. Hence, the above indices work well for data sets of fewer dimensions.

2.2 Clustering Validity Assessment: Finding the Optimal Partitioning of a Data set [5]

The majority of the clustering algorithms depend on certain assumptions in order to define the sub groups present in a data set. Such clustering schemes require some sort of cluster validation when considering many sub groups present in the entire data set. This paper deals with the definition of cluster validity index enabling the selection of the optimal input parameters values for a clustering algorithm that best partitions the data set.

In addition to the criteria widely accepted for partitioning the data set which are compactness and separability, the data set might be falsely partitioned in most of the cases when considering these two measures alone. The optimal set of input parameters leads to well separated

partitions of the data. DBSCAN algorithm defines clusters based on density variations, considering values for the cardinality and radius of an object's neighborhood which gives the best partitions for the given input values. The proposed cluster validity index in this paper considered clusters compactness (intra-cluster variance) and the density between clusters (inter-cluster density).

Description 1: Defining Inter-cluster density: This defines the evaluation of the average density in the region among clusters in relative with the density of the clusters.

Let D = {vi / i=1,…,c}, a partitioning of a data set S into c convex clusters where vi is the center of each cluster which results from applying a clustering algorithm to data set S. Let the average standard deviation (stdev) of clusters is given by the following equation

$$stdev = \frac{1}{c}\sqrt{\sum_{i=1}^{c}\|\sigma(v_i)\|} \qquad (4)$$

$\|x\|$ is calculated by the expression $\sqrt{(X^T X)}$

The goal of this index is that the density among clusters is to be significantly low in comparison with the density in the considered clusters. The inter-cluster density is defined by the equation

$$Dens\_bw(c) = \frac{1}{c\,.\,(c-1)}\sum_{i=1}^{c}\left(\sum_{j=1}^{c}\frac{density(U_{ij})}{max\{density(v_i),density(v_j)\}}\right) \; and \; j \neq i$$

where v_i, v_j are the centers of clusters c_i, c_j respectively. U_{ij} is the middle point of the line segment defined by the clusters centers v_i, v_j. With respect to this, the term density (u) is defined by the equation

$$density(u) = \sum_{i=j}^{u_{ij}} f(x_l, u)\,\pi r^2 \qquad (5)$$

where N_{ij} represents the number of tuples that belongs to clusters C_i and C_j and $x_l \in C_i \cup C_j \subseteq S$ .The function f(x,u) is defined by the following equation

$$f(x,u) = \begin{cases} 0, & if \;\; d(x,u) > stdev \\ 1, & otherwise \end{cases}$$

A point belongs to the neighborhood of u if its distance from u is smaller than the average standard deviation of clusters. The data set considered have been scaled to consider all the dimensions in picture.

Description 2: Defining Intra-cluster variance: The intra-cluster variance is the average scattering for clusters. This is a parameter defined for "within clusters".

Scattering for clusters is defined by

$$Scat(c) = \frac{1}{c}\sum_{i=1}^{c}\frac{\|\sigma(v_i)\|}{\|\sigma(S)\|} \qquad (6)$$

Where $\sigma(S)$ is the variance of the data set and $\sigma(v_i)$ is the variance of cluster c_i. The proposed validity index S_Dbw is found using the equation

$$S\_Dbw(c) = Scat(c) + Dens\_bw(c) \qquad (7)$$

From the above equation, a smaller value of Scat(c) is an indication of compact clusters and a smaller value of Dens_bw(c) indicates well-separated clusters. The number of clusters, c, that

minimizes the above index can be considered as an optimal value for the number of clusters that are present in the data set.

Intricacy

The S_Dbw proposed cluster validity index had known to work well on data sets with intra-cluster variance and inter-cluster density for very compact and well-separated cluster. The experimental results had considered data sets of multiple dimensions. However, the considered index performs well on non-standard geometry shaped cluster structure. The index does not work properly for ring shaped or extraordinarily curved shaped cluster structures. In addition to this, the authors of this paper have discussed some aspects with regard to the time complexity. The total complexity is found to be O(n). This has not been proved with the time complexity of other existing indices.

2.3 Performance Evaluation of Some Clustering Algorithms and Validity Indices [8]

The paper introduces some validity indices for the data sets that have distinct sub structures. When subgroups are present in the data sets, finding the optimal number of clusters is really a challenging task. The index provided in this literature gives a solution to the data set with distinct sub structures. The proposed new index I is defined by the following equation

$$I(k) = \left(\frac{1}{K} \times \frac{E_1}{E_K} \times D_K\right)^p \quad (8)$$

where k is the total number of clusters. In the above equation, the parameter E_K and D_K are given by

$$E_K = \sum_{k=1}^{K} \sum_{j=1}^{n} u_{kj} \left\| x_j - z_k \right\| \quad (9)$$

$$D_K = max_{i,j=1}^{K} \left\| z_i - z_j \right\| \quad (10)$$

where n is the total number of points in the data set. $U(X) = [\![u_{kj}]\!]_{K\times n}$ is a partition matrix for the data set and Z_k is the center of the k^{th} cluster. The value of k for which I(k) is maximized is considered to be the correct number of clusters.

Description 1: Compact Clusters:

The data points within each cluster should be close to each other. The index I is a composition of three factors, namely, $\frac{1}{K}, \frac{E_1}{E_K} and\ D_K$. The first factor will try to reduce index I as K is increased. The second factor consists of the ratio of E_1, which is constant for a given data set, and E_K which decreases with increase in K. Hence, index I increases as E_K decreases. This indicates that more number of clusters are formed which are compact in nature. The distance between points in the data set within a group is small.

Description 2: Maximum Separation:

The separation between two clusters over all possible pairs of clusters is maximized. The third factor in the index I, measures the maximum separation between two clusters over all possible pairs of clusters. D_K will increase the value of K. But this value is upper bounded by the maximum separation between two points in the data set. The power p in equ (8) is used to control the contrast between the different cluster configurations. The considered value of p in this paper is 2.

Intricacy:

The proposed index I is found to work well on any type of data sets in providing the appropriate number of clusters. The data set under consideration is based on medical grounds. The new index of this paper is not compared with other types of already existing indices. Since, there is no formal comparisons the speed with respect to convergence has not been well depicted. The distance measure used in the proposed index is the traditional Euclidean distance metric, wherein the index could be investigated for other distance metric measures also like Manhattan distance, Mahanolobis distance, etc. The data set with outliers and the time complexity of the working index had not been discussed. The clustering algorithms and the cluster validity indices described in this article functions for crisp data sets.

2.4 New indices for cluster validity assessment [11]

A cluster validity index is used to validate the outcome. This article presents an analysis of design principles implicitly used in defining cluster validity indices and reviews the limitations of existing cluster validity indices. New cluster validity indices are proposed in this paper which is found to face the limitation of other indices. In the summation-type CVIs the intra-cluster and inter-cluster distances are considered. The intra-cluster distance increases sharply as nc decreases from $nc_{optimal}$ to $nc_{optimal}-1$. The inter-cluster distance decreases sharply as nc decreases from $nc_{optimal}+1$ to $nc_{optimal}$. This assumption is not taken into account in the traditional CVIs, wherein it is taken to be a valid factor in the case of the proposed CVI.

Description 1: Compactness

The measure of decompactness is obtained, to obtain the degree of compactness. The problem of unnecessary merging is also taken into account for the new index. The average decompactness measure which is the mean absolute deviation in this case is obtained by the following equations.

$$v_u(nc) = \frac{1}{nc}\sum_{i=1}^{nc}\left(\frac{1}{n_i}\sum_{x\in C_i} d(x, c_i)\right) \tag{11}$$

$$v_o(\mathrm{nc}) = \frac{nc}{\mathrm{d_{min}}}, \qquad d_{min} = \min_{i\neq j} d(c_i, c_j)$$

$$v_{sv}(nc) = v_{uN}(nc) + v_{oN}(nc) \tag{12}$$

$v_{uN}(nc)$ and $v_{oN}(nc)$ are min-max normalized versions of $v_u(nc)$ and $v_o(nc)$ respectively. This approach tends to hide the effect of a cluster by unnecessary merging. In order to solve this problem, the new proposed indices follow the equations below

$$v_u^* = \max_{i=1,2,..nc}\{v_{u,i}(nc)\}$$

$$= \max_{i=1,2,...nc}\left\{\frac{1}{n_i}\sum_{x\in C_i} d(x, c_i)\right\}$$

$$v_{sv}^*(nc) = v_{uN}^*(nc) + v_{oN}(nc) \tag{13}$$

v_u^* is given as a new variant of the already existing v_u.

The "Scat" value in the following equation where,

$$D_{max} = \max_{i \neq j} d(c_i, c_j)$$

and $D_{min} = \min_{i \neq j} d(c_i, c_j)$ resorts to the averaging behavior, the same problem is expected and hence a new variant is proposed. The existing and the new modified values are given as follows

$$Scat(nc) = \frac{1}{nc} \sum_{i=1}^{nc} \frac{\|\sigma(c_i)\|}{\|\sigma(X)\|} \quad (14)$$

$$\mathrm{Dis}(nc) = \frac{D_{max}}{\mathrm{D_{min}}} \sum_{i=1}^{nc} \left(\sum_{j=1}^{nc} d(c_i, c_j) \right)^{-1}$$

$$SD(nc) = a.Scat(nc) + Dis(nc)$$

$$a = Dis(nc_{max})$$

$$Scat^*(nc) = \max_{i=1,2,\ldots,nc} \left\{ \frac{\|\sigma(c_i)\|}{\|\sigma(X)\|} \right\} \quad (15)$$

$$SD^*(nc) = a\,.Scat^*(nc) + Dis(nc)$$

$$a = Dis(nc_{max})$$

Intricacy

The design principles of the new indices are explained and verified for performance against the existing indices. However, the traditional index S_Dbw is not considered for comparisons because of the high computational cost. The data set considered are 2-D data where the new indices had performed well and there had been no discussion on whether these new indices will perform well on high-dimensional data set and data set with noise.

2.5 A new cluster validity index for prototype based clustering algorithms based on inter- and intra-cluster density [13]

This article faces the common challenges of how to evaluate, without auxiliary information and to what extent the obtained clusters fit the natural partitions of the data set. A new validity index, conn_index for prototype based clustering of data sets is applicable with a wide variety of cluster characteristics (i.e.) clusters of different shapes, sizes, densities and even overlaps. conn_index is based on weighted Delaunay triangulation called "connectivity matrix".

For crisp clustering, the Davies-Bouldin index [8] and the generalized Dunn Index [10] are some of the most commonly used indices. Both depend on a *separation* measure between clusters and a measure for *compactness* of clusters based on distance. Even though these two indices work satisfactorily for well-separated clusters, they may fail for complicated data structures with clusters of different shapes or sizes or with overlapping clusters. When the clusters have homogeneous density distribution, one effective approach to correctly evaluate the clustering of data sets is CDbw (composite density between and within clusters) [16]. CDbw finds prototypes for clusters instead of representing the clusters by their centroids, and calculates the validity measure based on inter- and intra-cluster densities, and cluster separation. The densities are calculated as the number of data samples within a standard deviation from the prototypes. However, it fails to represent true inter- and intra-cluster densities when the clusters have inhomogeneous density distribution.

Definition 1: Let CADJ be an N × N matrix where N is the number of prototypes. The cumulative adjacency, CADJ(i,j) of two prototypes v_i and v_j is the number of data vectors for which v_i is the Best Matching Unit (BMU) and v_j is the second BMU. Each prototype is BMU for the data vectors in its Receptive Field (RF) (Voronoi polyhedron). By this definition, $|RF_i| =$

$\sum_{k=1}^{N} CADJ(i,k)$ and RF is the receptive field in Voronoi polyhedron. The size of the RF indicates how the data is distributed among the prototypes.

Definition 2: The level of connectedness (similarity) of two prototypes v_i and v_j is

$$CONN(i,j) = CADJ(i,j) + CADJ(j,i) \quad (16)$$

By the definition, CONN is symmetric and shows how similar two prototypes are by indicating the number of data vector for which they are the BMU and the second BMU pair.

Description 1: Compactness of Clusters

Assuming k number of clusters, N prototypes *v* in a data set, C_k and C_l are two different clusters where $1 \leq k, l \leq K$, the new proposed CONN_Index will be defined with the help of Intra_Conn and Inter_Conn quantities which are considered as compactness and separation. The compactness of C_k, Intra_Conn(C_k) is the ratio of the number of data vectors in C_k whose second BMU is also in C_k, to the number of data vectors in C_k. The Intra_Conn(C_k) is defined by,

$$Intra\_Conn(C_k) = \frac{\sum_{i,j}^{N} \{CADJ(i,j): v_i\, v_j \in C_k\}}{\sum_{i,j}^{N} \{CADJ(i,j): v_i \in C_k\}} \quad (17)$$

and Intra_Conn $\in [0,1]$. The greater the value of Intra_Conn the more is the cluster compactness. If the second BMUs of all data vectors in C_k are also in C_k, then Intra_Conn(C_k)=1. The intra-cluster connectivity of all clusters (Intra_Conn) is the average compactness which is given below

$$A = \sum_{k}^{K} \frac{Intra\_Conn(C_k)}{K} \quad (18)$$

Description 2: Separation of Clusters

The inter-cluster connectivity between clusters C_k and C_l, Inter_Conn(C_k,C_l) is the ratio of the connectivity between C_k and C_l to the total connectivity of the prototypes in C_k, which have atleast one connection to a prototype in C_l.

$$Inter\_Conn(C_k, C_l) = \frac{Conn(c_k, C_l)}{\sum_{i,j}^{N}\{CONN(i,j): v_i \in V_{k,l}\}} \quad (19)$$

$$Conn(C_k, C_l) = \sum_{i,j}^{N}\{CONN(i,j) : v_i \in C_k\,,\, v_j \in C_l\} \quad (20)$$

$$V_{k,l} = \{v_i : v_i \in C_k, \exists\, v_j \in C_l : CADJ(i,j) > 0\}$$

The ratio in the above equation shows how similar the prototypes at the boundary of C_k are to the ones at the boundary of C_l. If C_k and C_l are well separated (i.e.) they have no connection at all, then Inter_Conn(C_k, C_l) = 0.The inter-connectivity of C_k to all other clusters, Inter_Conn(C_k) and the average similarity of clusters, Inter_Conn, as

$$Inter\_Conn(C_k) = \max_{l,l \leq K}\ Inter\_Conn(C_k, C_l) \quad (21)$$

$$Inter\_Conn = \sum_{k}^{K} \frac{Inter\_Conn(C_k)}{K}$$

Now, 1-Inter_Conn is the separation measure between clusters. Hence, the definition of the newly proposed cluster validity index, Conn_Index is given by the equation,

$$Conn\_Index = Intra\_Conn \times (1 - Inter\_Conn) \quad (22)$$

and Conn_Index∈ [0,1] which increases with better clustering and the value 1 indicates perfectly separated clusters.

Intricacy:

When clustering with different numbers of clustered prototypes are compared to a fully clustered SOM, only Inter_Conn should be taken into account because Intra_Conn is affected heavily by the unclustered prototypes. Similarly, when comparing the validity of a fully clustered SOM with one that has unclustered prototypes, Intra_Conn and consequently Conn_Index do not provide a meaningful measure. The data set considered in this article has no mapping of outlier data. Also, the authors have not discussed in detail about the time complexity of validation compared to other validity indices.

2.6 An Evolutionary Cluster Validation Index [18]

A new evolutionary method for cluster validation index is proposed in this article. The index learns from the generated training data set using genetic programming (GP) and the optimal number of clusters by taking the parameters of the test data set into the learned cluster validation index. The chromosomes used in Genetic Programming encodes a possible validation index as a function of the number of clusters, density measure of clusters and some other random heuristics chosen by the user. Most of the existing CVIs efficiency depends mainly on the traits of the datasets. Any CVI cannot guarantee the optimal number of clusters for all types of data sets. However, the fitness function used in the programming is employed to resolve the above said issue and is evaluated for each candidate and is defined by the difference between the actual number of clusters from training data set and the number of clusters computed by the current CVI. The authors had argued that since because the Genetic Programming is adaptive in nature, the proposed eCVI (Evolutionary based CVI) is highly reliable. eCVI follows the general framework of GP. The activities of GP are representation through encoding of strings, fitness evaluation, and application of GP operators. The algorithm will iterate these evolutionary procedures until a termination criterion is met.

Description: Intra- and Inter-cluster distance metrics

Representation in GP is significantly different from those of other evolutionary algorithms that employ linear strings. The chromosomes that are used in the evolutionary algorithm are expressed by a nonlinear form such as tree structures. The semantically rich tree structures have the chromosomes consisting of two main components namely nodes and leaves. Nodes are reserved for unary or binary functions and the leaves (terminal points) take some constants or input parameters.The leaves are set to the input parameters {N, $R_{tra}(n_c)$, $R_{ter}(n_c)$}, where N is a natural random number ranging from 1 to 9, $R_{tra}(n_c)$ is the average intra-cluster distance, $R_{ter}(n_c)$ is the average inter-cluster distance. The intra-cluster distance is computed by finding the average distance from every data point x_j in each cluster C_i to the centroid v_i of C_i. The average inter-cluster distance between data points x_k in each cluster C_i and the centroid of other clusters are computed. The equations for $R_{tra}(n_c)$ and $R_{ter}(n_c)$ are computed as given below,

$$R_{tra}(u_c) = \frac{1}{n_c} \sum_{i=1}^{n_c} \left(\frac{1}{|C_i|} \sum_{x_j \in C_i} \{ d(x_j, v_i) \} \right) \quad (23)$$

$$R_{ter}(n_c) = \frac{1}{n_c} \left[\sum_{i=1}^{n_c} \frac{1}{|C_i|} \left(\sum_{j=1}^{n_c} \sum_{x_k \in C_i, i \neq j} d(x_k, v_j) \right) \right] \quad (24)$$

Each chromosome is constructed by a stochastic combination of the above elements, which represents a mathematical rule for the number of clusters.

Fitness Function

The fitness function is known to evaluate the chromosomes that are evolved with the training data set. It efficiently predicts the output of chromosomes in accordance with the input data set. The fitness value provides feedback information on the current learning with regard that which individuals have a higher chance to survive. The fitness value just now described is closely related to CVI in the sense that the chromosomes are learned for reliably modeling common features of training data sets in terms of the average intra- and the average inter-cluster distances computed under the considered number of clusters (n_c).

Consider a population $\mathcal{H} = \{h_1, h_2, \dots, h_N\}$, where N is the population size. Let f(h_k) be an output of the chromosome h_k, which is the value of eCVI for h_k. Since, eCVI tries to estimate the optimal number of clusters we can represent that

$$f(h_k) = n_c^* + \epsilon(h_k) \quad (25)$$

The fitness function FeCVI(hk) can be defined as the difference between the optimal number of clusters n_c^* and the output of chromosomes (f(hk)) which could be formulated as,

$$F_{eCVI}(h_k) = |n_c^* - f(h_k)| \quad (26)$$

Intricacy

The working of eCVI is compared with three other existing indices. eCVI index is found to perform well for any type of data set due to the adaptive nature of the fitness function used in Genetic Programming. However, the experimental data set considered are 2D and 3D data and it is not discussed anywhere about the working of eCVI for high-dimensional data set and data with outliers or noise. The time complexity of the algorithm is also not explained in detail, since the parameter has to be taken into consideration because the fitness function iterates continuously until termination condition is met.

2.7 A Validity Index based on Connectivity [21]

An index based on connectivity is proposed in this paper. This index is designed in such a way that the data sets have well separated clusters of any shape and size. The index uses the concept of relative neighborhood graph for measuring the amount of connectedness to a particular cluster. Suppose the clusters formed are denoted by C_i for i=1,…,K, where K is the number of clusters, the diameter of a particular cluster is denoted as diam(C_i) for i=1,…,K, and is defined as follows

$$diam(C_i) = \max_{x,y \in C_i}\{d_{short}(x,y)\} \quad (27)$$

d_{short} is the distance between any two points x and y and is measure along the relative neighborhood graph [25] . Find all possible paths among these two points along the RNG. Suppose there are total p paths between x and y, and the number of edges along the i^{th} path, denoted$ed_1^i, ed_2^i, \ldots, ed_{n_i}^i$, if any weights provided they are multiplied by their corresponding weights, w. Now, the shortest distance between x and y is formulated as,

$$d_{short}(p,q) = \min_{i=1}^{p} \max_{j=1}^{n_i} w(ed_j^i) \quad (28)$$

The distance between any two clusters C_i and C_j, where i,j = 1,2,…,K and $i \neq j$ and it is defined by,

$$dist(C_i, C_j) = \min_{x \in C_i \text{ and } y \in C_j}\{d_{short}(x,y)\} \quad (29)$$

The newly proposed connectivity based Cluster Validity Index, connect-index is formulated by,

$$connect = \min_{1 \le i \le K}\left\{\min_{1 \le j \le K, i \neq j}\left\{\frac{dist(C_i, C_j)}{\max_{1 \le k \le K}\{diam(C_k)\}}\right\}\right\}$$

The larger values of connect indicates good partitioning and hence the appropriate number of clusters is determined by maximizing connect over the different values of K. Assuming, $connect_i$, denoting the connect-index value for the number of clusters, K=I, the appropriate numbers of clusters K^*, is given by the equation,

$$K^* = argopt\{\max_{i=1,\ldots K_{max}} connect_i\} \quad (30)$$

K_{max} is the maximum number of clusters. If the cluster is completely connected then the shortest distance between any two points would be very small and thus the diameter of that particular cluster would be small too. Since, connect-index tries to minimize the maximum diameter amongst all clusters, this in turn tries to minimize the diameter of every clusters. This indicates that when all clusters are well connected, their diameters are small and the denominator of the connect-index gets a smaller value. The numerator of the connect-index is the minimum separation between any two clusters which is measured as the minimum shortest distance between any two points belonging to two different clusters.

Intricacy

The proposed cluster validity index had been tested for various artificial and real time data sets. The connect-index had been proved that it is better than only one of the existing validity indices and was not compared with some of the others. A clear mathematical proof is not given for such an index which could be a limitation when writing a lemma. In addition to the above, the data set may contain outliers since the data set taken under consideration is the medical data set. There was no discussion on whether the proposed connectivity-based index performs well on data sets even with noisy data.

2.8 A Subtractive Based Subspace Clustering Algorithm on High Dimensional Data [25]

The dimension of the data is always a curse in the process of Data Exploration. Clustering technique and Cluster Validation indices is a major task in high dimension and very high dimensional data set. Automatic determination of the number of clusters is done by utilizing a subtractive based clustering algorithm and advanced cluster validity index for such clustering technique. The quality of clustering should lead to minimal intra-cluster compactness while the inter-cluster separation is as maximized as possible. The most existing CVIs are based on distances between data objects. The distance utilizes the feature vector of data objects in the case of high dimensional data. All the functions in conventional CVIs make use of all the dimensions. However, in high dimensional spaces, most distance functions are not very useful due to the curse of dimensionality and the clusters always exist in different sub spaces. Therefore, new CVI for validating high-dimensional clustering is proposed.

Description 1: Compactness in High-dimensional spaces

In view of in high dimensional spaces data points may cluster differently in varying subspaces transformed by w [29]. Let D_i represent a dimension set, which is composed of correlated dimensions in which data objects in the i^{th} cluster, and $distD_i(x_i, x_j)$ is a distance function reflecting only these dimensions. The distance function is given by the equation below,

$$dist_{D_i}(x_i, x_j) = \sqrt{\sum_{i=1}^{d} w_{il}(x_{il} - x_{jl})^2} \qquad (31)$$

The above definition of distance between x_i and x_j provides a new direction for compactness and separability in high dimensional spaces. For data objects within a cluster, equ (31) has relatively a smaller value and for two clusters to be well separated, equ (31) should have relatively a larger value. Though, the two clusters share the same dimension set, $distD_i(c_i, c_j)$ also has a relatively high value because they should be well separated in the space defined by the common cluster dimension set, where ci and cj are the centroid of i^{th} and j^{th} clusters respectively. Newly designed validity indices are given below,

$$H\_V_k = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m dist_{D_i}(x_j, c_i)^2 \quad + \frac{1}{c}\sum_{i=1}^{c} dist_{D_i}^2(c_i, \bar{c})}{\min\limits_{i=1,..,c}\left\{\min\limits_{k=1,..,c, i\neq k} dist_{D_i}(c_i, c_k)^2\right\}}$$

$$= \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m\left(\sum_{l=1}^{d} w_{il}(x_{jl}-c_{il})^2\right)+\frac{1}{c}\sum_{i=1}^{c}\left(\sum_{l=1}^{d}\left(\sqrt{w_{il}c_{il}}\sqrt{w_{0l}\bar{c}_l}\right)^2\right)}{\min_{i\neq k}\left(\sum_{l=1}^{d} w_{il}(c_{il}-c_{kl})^2\right)} \qquad (32)$$

$$H\_V_{Hong} = coh(c) + \alpha . Dis(c) \qquad (33)$$

$$= \sum_{i=1}^{c}\left(\sum_{j=1}^{n}(u_{ij})^m\left(\sum_{l=1}^{d} w_i\left(x_{jl}-c_{il}\right)^2/n_i\right)\right) + \frac{\left(\sum_{i=1}^{c_{max}}\left(\sum_{l=1}^{d}\left(\sqrt{w_{il}c_{il}}-\sqrt{w_{0l}\bar{c}_l}\right)^2\right)\right)/c_{max}}{\sum_{i=1}^{c_{max}}\left(\sum_{j=1}^{n}(u_{ij})^m\left(\sum_{l=1}^{d} w_i(x_{jl}-c_{il})^2\right)/n_i^{1/2}\right)} \cdot \frac{1}{\min_{i\neq k}\left(\sum_{l=1}^{d} w_{il}(c_{il}-c_{kl})^2\right)}$$

$$where\ \bar{c} = \frac{1}{n}\sum_{j=1}^{n} x_j \quad , and \quad n_i = \sum_{j=1}^{n} u_{ij}$$

Intricacy:

The experimental results on both synthetic and real data sets have shown that the new algorithm outperformed other conventional cluster validity indices. However, a clear mathematical proof is not provided for the new proposed indices. The time complexity of clustering for high-dimensional data spaces generally increases, and so with the cluster validation process. The time complexity of such index is not discussed. The experimental data set is of high-dimension but there is no provision of handling data with outliers in the proposed indices.

3. Concluding Remarks

For each of the category of literatures, a comprehensive review of previous analysis has been presented. Most of the existing literatures that concentrate on the study and proof of CVIs are based on intra-cluster and inter-cluster distance metrics only. The authors have concluded that any type of CVI cannot guarantee the best number of clusters because the evaluation metrics are computed based on geometrical distances, which in turn has a high degree of masking the discriminatory capacity especially when the input data sets are of very high dimension and highly embedded with noise, considered as outliers. Hence, the performance of partitioning all types of data sets, whether high dimension or with outliers, shall work better when considering the geometry of clusters too. Therefore, the authors have found that there is a tremendous scope in research to develop CVIs to include the geometrical shape of the clusters formed for multiple dimensions and even mixed type data sets.

References

1. James C. Bezdek, *Fellow, IEEE*, and Nikhil R. Pal, (1998) “Some New Indexes of Cluster Validity”, IEEE transactions on Systems, Man, and Cybernetics—Part b: Cybernetics, vol. 28, no. 3.

2. http://en.wikipedia.org/wiki/Convex_hull.

3. J. C. Bezdek, W. Q. Li, Y. Attikiouzel, and M. Windham,(1997) “A geometric approach to cluster validity for normal mixtures,” Journal on Soft Computing – A Fusion of Foundations, Methodologies and Applications, vol. 1, no.4, 166–179.

4. A. Jain and R. Dubes, (1998) “Algorithms for Clustering Data”, Englewood Cliffs, NJ: Prentice Hall.

5. Maria Halkidi Michalis Vazirgiannis, (2001) “Clustering Validity Assessment: Finding the optimal partitioning of a data set”, First IEEE International Conference on Data Mining (ICDM'01).

6. Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan,(1998) "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications". Proceedings of ACM SIGMOD, vol. 27, Issue 2.

7. Alexander Hinneburg, Daniel Keim, (1998) "An Efficient Approach to Clustering in Large Multimedia Databases with Noise". Proceeding of KDD '98.

8. Ujjwal Maulik, Sanghamitra Bandyopadhyay, (2002) “Performance Evaluation of Some Clustering Algorithms and Validity Indices”, IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol. 24, No. 12.

9. L.O. Hall, I.B. Ozyurt, and J. C. Bezdek, (1999) “Clustering with a Genetically Optimized Approach,” IEEE Transactions on Evolutionary Computation, vol. 3, no. 2,103-112.

10. R.B. Calinski and J. Harabasz, (1974) “A Dendrite Method for Cluster Analysis,” Communication in Statistics – Simulation and Computation, Vol. 3, Issue 1, 1-27.

11. Minho Kim, R.S. Ramakrishna, (2005) "New indices for cluster validity assessment", Elsevier Journal on Pattern Recognition Letters 26, 2353–2363.

12. Berry, M.J.A., Linoff, G., (1997) "Data Mining Techniques: For Marketing, Sales, and Customer Support", John Wiley & Sons, Berlin.

13. Kadim Tas¸demir and Erzs´ebet Mer´enyi, (2007) "A new cluster validity index for prototype based clustering algorithms based on inter- and intra-cluster density", In Proceedings of International Joint Conference on Neural Networks, 2007 (IJCNN 2007), Orlando, FL.

14. K.L Wu, and M.S. Yang, "A cluster validity index for fuzzy clustering,(2005) "ElSevier Journal on Pattern Recognition Letters, vol. 26, Issue 9, 1275–1291.

15. U. Maulik, and S. Bandyopadhyay, (2002) "Performance evaluation of some clustering algorithms and validity indices," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 12.

16. X.L. Xie, and G. Beni, (1991) "A validity measure for fuzzy clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.13, no.8, pp.841–847.

17. Chang Wook Ahn and R.S. Ramakrishna, (2002) "A Genetic Algorithm for Shortest Path Routing Problem and the Sizing of Populations", IEEE Transactions on Evolutionary Computation, Vol.6, No.6.

18. Sanghoun Oh, Chang Wook Ahn, Moongu Jeon, (2008), "An Evolutionary Cluster Validation Index", Proceedings of 3rd International Conference on Bio- Inspired Computing: Theories and Applications, BICTA 2008, 83-88.

19. Nam Hyun Park, Chang Wook Ahn, and R.S. Ramakrishna, (2005) "Adaptive Clustering Technique Using Genetic Algorithms", IEICE Transactions on Information and System, Vol.E88-D. No.12.

20. C.-H Chou, M.-C. Su, and E. Lai, (2006) "A new cluster validity measure and its application to image Compression Sergios Theodoridis", Pattern Recogniton (Third Edition), Academic Press, Inc. Orlando, FL, USA.

21. Sriparna Saha and Sanghamitra Bandyopadhyay, (2009) "A Validity Index Based on Connectivity", Seventh International Conference on Advances in Pattern Recognition.

22. C. H. Chou, M. C. Su, and E. Lai, (2004) "A new cluster validity measure and its application to image compression," ACM Journal on Pattern Analysis and Applications, vol. 7, Issue 2, 205–220.

23. S. Bandyopadhyay and S. Saha, (2008) "A point symmetry based clustering technique for automatic evolution of clusters," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 11, 1–17.

24. S. Saha and S. Bandyopadhyay, (2008) "Application of a new symmetry based cluster validity index for satellite image segmentation," IEEE Geoscience and Remote Sensing Letters, vol. 5, no. 2, 166–170.

25. Deng Ying,Yang Shuangyuan ,and Liu Han, (2009) "A Subtractive Based Subspace Clustering Algorithm on High Dimensional Data", Proceedings of the 1st International Conference on Information Science and Engineering (ICISE2009).

26. H. Sun and M. Sun, (2006) "Trail-and-error approach for determining the number of clusters"[J]. ICMLC 2005, LNAI 3930, vol. 3930, 229 – 238.

27. Lifei Chen, Qingshan Jiang, Shengrui Wang, (2008) "Cluster validation for subspace clustering on high dimensional data" [C], Proceeding of the 2008 IEEE Asia Pacific Conference on Circuits and Systems, Macao:China.

28. L.Jing, M.K.Ng and J.Z.Huang, (2007) "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data"[J]. IEEE Transactions on Knowledge and Data Engineering, vol. 19, no.8, 1-16.

29. C.Domeniconi, D.Gunopulos, et al. (2007) "Locally adaptive metrics for clustering high dimensional data", ACM Journal on Data Mining and Knowledge Discovery, vol 14, Issue 1.

30. Zhiling Hong, Qingshan Jiang,Huailin Dong and Shengrui Wang. (2008) "A new cluster validity index for fuzzy clustering", Elsevier Journal on Information Sciences, vol. 178, Issue 4.

31. S.M. Pan and K.-S. Cheng, (2007) "Evolution-based tabu search approach to automatic clustering," IEEE Transactions on Systems, Man, and Cybernetics. C, Appl. Rev., vol. 37, no. 5, 827–838.

32. E. Hruschka, R. J. G. B. Campello, A. A. Freitas, and A. C. Ponce Leon F. de Carvalho, (2009) "A survey of evolutionary algorithms for clustering," IEEE Transactions on Systems, Man, and Cybernetics. C, Appl. Rev., vol. 39, no. 2, 133–155.

33. U.Maulik, (2008) "Hierarchical pattern discovery in graphs," IEEE Transactions on Systems, Man, and Cybernetics C, Appl. Rev., vol. 38, no. 6, 867–872 .