International Journal of Research in Computer Science
eISSN 2249-8265 Volume 2 Issue 1 (2011) pp. 39-43

www.ijorcs.org

QUALITY OF CLUSTER INDEX BASED ON STUDY OF DECISION TREE

B.Rajasekhar[1], B. Sunil Kumar[2], Rajesh Vibhudi[3], B.V.Rama Krishna[4]

[1 2]*Assistant Professor, Jawaharlal Nehru Institute of Technology, Hyderabad*

[3]*Sri Mittapalli college of Engineering, Guntur, Hyderabad*

[4]*Associate Professor, St Mary's College of Engineering & Technology, Hyderabad*

Abstract:- *Quality of clustering is an important issue in application of clustering techniques. Most traditional cluster validity indices are geometry-based cluster quality measures. This work proposes a cluster validity index based on the decision-theoretic rough set model by considering various loss functions. Real time retail data show the usefulness of the proposed validity index for the evaluation of rough and crisp clustering. The measure is shown to help determine optimal number of clusters, as well as an important parameter called threshold in rough clustering. The experiments with a promotional campaign for the retail data illustrate the ability of the proposed measure to incorporate financial considerations in evaluating quality of a clustering scheme. This ability to deal with monetary values distinguishes the proposed decision-theoretic measure from other distance-based measures. Our proposed system validity index can also be efficient for evaluating other clustering algorithms such as fuzzy clustering.*

Keywords**:** *Clustering, Classification, Decision Tree, K-means.*

I. INTRODUCTION

Unsupervised learning clustering is one of the techniques in data mining, categorizes unlabeled objects into several clusters such that the objects belonging to the same cluster are other similar than those belonging to different clusters. Conventional clustering assigns an object to exactly one cluster. Assign an object to rough-set-based variation makes it possible. [3]. Quality of clustering is an important issue in application of clustering techniques to real-world data. A good measure of cluster quality will help in deciding various parameters used in clustering algorithms. One such parameter that is common to most clustering algorithms is the number of clusters. Many different indices of cluster validity have been proposed. In general, indices of cluster validity fall into one of three categories. Some validity indices measure partition validity to evaluate the properties of crisp structure imposed on the data by the clustering algorithm, such as Dunn indices [7] and Davies-Bould in index [2]. These validity indices are based on similarity measure of clusters whose bases are the dispersion measure of a cluster and the cluster dissimilarity measure. In the case of fuzzy clustering algorithms, some validity indices such as partition coefficient [1] and classification entropy use only the information of fuzzy membership grades to evaluate clustering results. The third category consists of validity indices that make use of not only the fuzzy membership grades but also the structure of the data. All these validity indices are essentially based on the geometric characteristics of the clusters. A decision-theoretic measure of cluster quality, decision theoretic framework has been helpful in providing a better understanding of classification models [4]. The decision theoretic rough set model considers various classes of loss functions. By adjusting loss functions, the decision-theoretic rough set model can also be extended to the multi category problem. It is possible to construct a cluster validity index by considering various loss functions based on decision theory. Such a measure has an added advantage of being applicable to rough-set-based clustering. This work describes how to develop a cluster validity index from the decision-theoretic rough set model. Based on the decision theory, the proposed rough cluster validity index is taken as a function of total risk for grouping objects using a clustering algorithm. Since crisp[5] clustering is a special case of rough clustering, index validity is applicable to both rough clustering and crisp clustering. Experiments with synthetic and real-world data show the usefulness of the proposed validity index for the evaluation of rough clustering and crisp clustering.

II. CLUSTERING TECHNIQUE

The clustering technique K-means [7] is a prototype-based, simple partitional clustering technique which attempts to find k non-overlapping clusters. These clusters are represented by their centroids (a cluster centroid is typically the mean of the points in the cluster). The clustering process of K-means is as follows. Firstly, k initial centroids are selected, where k is specified by the user and indicates the desired number of clusters.

Secondly, every point in the data is then assigned to the closest centroid, and each collection of points assigned to a centroid forms a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster. This process is repeated until no point changes clusters.

A. Clustering Crisp Method: The objective of the k-means is to assign n objects to k clusters. The process begins by randomly choosing k objects as the centroids of the k clusters. The objects are assigned to one of the k clusters based on the minimum value of the distance $d(\vec{x}_1, \vec{c}_1)$ between the object vector $\vec{x}_1$ and the cluster vector $\vec{c}_1$ the distance $d(\vec{x}_1, \vec{c}_1)$ can be the standard Euclidean distance.

Assignment of all the objects to various clusters, the new centroid vectors of the clusters are calculated as

$$\vec{c}_i = \frac{\sum_{\vec{x}_i \in \vec{c}_i} \vec{x}_i}{|\vec{c}_i|} \, , where \; 1 \le i \le k$$

Here $|\vec{c}_i|$ is the cardinality of cluster $\vec{c}_i$. The process stops when the centroids of clusters stabilize, i.e., the centroid vectors from the previous iteration are identical to those generated in the current iteration.

B. Cluster Validity: A New validity index conn_index for prototype based clustering of data sets is applicable with a wide variety of cluster characteristics clusters of different shapes, sizes, densities and even overlaps. Conn_index is based on weighted Delaunay triangulation called “connectivity matrix”.

Crisp clustering the Davies-Bouldin index and the generalized Dunn Index are some of the most commonly used indices depend on a separation measure between clusters and a measure for compactness of clusters based on distance. 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.

C. 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 $1 \le k, l \le K$, the new proposed CONN_Index will be defined with the help of Intra and Inter quantities which are considered as compactness and separation. The compactness of C_k, Intra (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 (C_k) is defined by

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

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

$$A = \sum_k^K \frac{Intra\_Conn(C_k)}{K}$$

D. Cluster Quality: Several cluster validity indices to evaluate cluster quality obtained by different clustering algorithms. An excellent summary of various validity measures [10] are two classical cluster validity indices and one used for fuzzy clusters.

1. Davies-Bouldin Index:

This index [6] is a function of the ratio of the sum of within cluster scatter to between-cluster separation. The scatter within the *i*th cluster, denoted by S_i, and the distance between cluster $\vec{c}_i$ and $\vec{c}_j$, denoted by d_{ij}, are defined as follows:

$$S_{i,q} = \left(\frac{1}{|\vec{c}_i|}\sum_{\vec{x}\in\vec{c}_i}\|\vec{x} - \vec{c}_i\|_2^q\right)^{1/q}$$

$$d_{ij,t} = \left\|\vec{c}_i - \vec{c}_j\right\|_t$$

where c_i is the center of the i^{th} cluster. c_{ij} is the number of objects in $\vec{c}_j$. Integers q and t can be selected independently such that q, t > 1. The Davies-Bouldin index for a clustering scheme (CS) is then defined as

$$DB(CS) = \frac{1}{k}\sum_{i=1}^{k} R_{i,qt} \, , where \; R_{i,qt} = \max_{1\le j\le k, j\ne 1}\{S_{i,q} + S_{j,q}/d_{ij,t}\}$$

The Davies-Bouldin index considers the average case of similarity between each cluster and the one that is most similar to it. Lower Davies-Bouldin index means a better clustering scheme.

2. Dunn Index:

Dunn proposed another cluster validity index [7]. The index corresponding to a clustering scheme (CS) is defined by

$$D(CS) = \min_{1\le j\le k}\left\{\min_{1\le j\le k, j\ne 1}\left(\frac{\delta(\vec{c}_i, \vec{c}_j)}{\max_{1\le q\le k}\Delta(\vec{c}_q)}\right)\right\}$$

$$\delta(\vec{c}_i, \vec{c}_j) = \min_{1\le i,j\le k, i\ne j}\left\|\vec{c}_i - \vec{c}_j\right\|,$$

$$\Delta(\vec{c}_i) = \max_{\vec{x}_i, \vec{x}_t \in \vec{c}_i}\|\vec{x}_i - \vec{x}_t\|$$

If a data set is well separated by a clustering scheme, the distance among the clusters, $\delta(c_i,c_j)(1\le i,j \le k)$ is usually large and the diameters of the clusters, $\Delta(c_i)$ $(1\le i \le k)$, are expected to be small. Therefore, a large value of D(CS) corresponds to a good clustering

scheme. The main drawback of the Dunn index is that the calculation is computationally expensive and the index is sensitive to noise.

III. DECISION TREE

A. Decision Tree

A decision tree depicts riles for Classifying data into groups. Splits entire data set into some number of pieces and then another rule may be applied to a piece, different rules to different pieces forming a second generation of pieces. The tree depicts the first split into pieces as branches emanating from a root and subsequent splits as branches emanating from nodes on older branches. The leaves of the tree are the final groups, the unsplit nodes. For some perverse reason, trees are always drawn upside down, like an organizational chart. For a tree to be useful, the data in a leaf must be similar with respect to some target measure, so that the tree represents the segregation of a mixture of data into purified groups.

Consider an example of data collected on people in a city park in the vicinity of a hotdog and ice cream stand. The owner of the concession stand wants to know what predisposes people to buy ice cream. Among all the people observed, forty percent buy ice cream. This is represented in the root node of the tree at the [9] top of the diagram. The first rule splits the data according to the weather. Unless it is sunny and hot, only five percent buy ice cream. This is represented in the leaf on the left branch. On sunny and hot days, sixty percent buy ice cream. The tree represents this population as an internal node that is further split into two branches, one of which is split again.

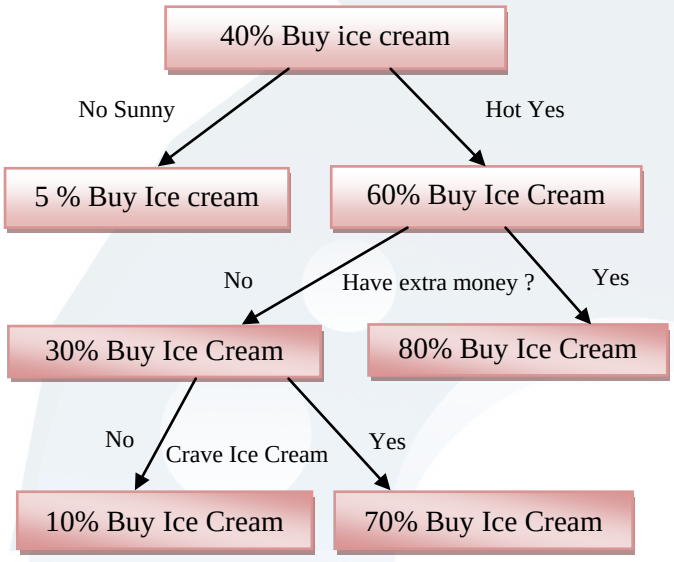

Figure 1 Example of Decision Tree

B. Yao's model for Decision Tree:

The model consists of two parties to holding values xЄX and yЄY respectively who can communicate with each other, would like to compute a function f: X*Y-> {0,1} at (x,y) with minimal amount of interaction between them. Interaction is some measure of communication between the two parties and it is usually the total number of bits exchanged between the parties. The classification of objects according to approximation operators in rough set theory can be easily fitted into the Bayesian decision-theoretic framework. Let $\Omega=\{A,A^c\}$ denote the set of states indicating that an object is in A and not in A, respectively. Let $A=\{a_1,a_2,a_3\}$ be the set of actions, where a_1,a_2,a_3 represent the three actions in classifying an object, deciding POS(A), deciding NEG(A), and deciding BND(A), respectively.

C. Implementation of the CRISP-DM:

CRISP-DM is based on the process flow showed in Figure 1. The model proposes the following steps:

1. *Business Understanding* – to understand the rules and business objectives of the company.
2. *Understanding Data* – to collect and describe data.
3. *Data Preparation* – to prepare data for import into the software.
4. *Modelling* – to select the modelling technique to be used.
5. *Evaluation* – to evaluate the process to see if the technique solves the problem of modelling and creation of rules.
6. *Deployment* – to deploy the system and train its users.

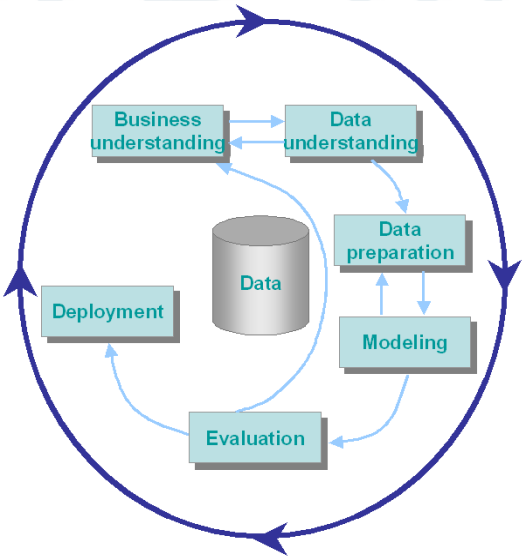

Figure 2 Example of Crisp data mining

IV. PROPOSED SYSTEM

Unsupervised classification method when the only data available are unlabeled. It is need to know the number of clusters. A cluster validity measure can provide us some information about the appropriate number of clusters. Our solution possible to construct a cluster validity measure by considering various loss functions based on decision theory.

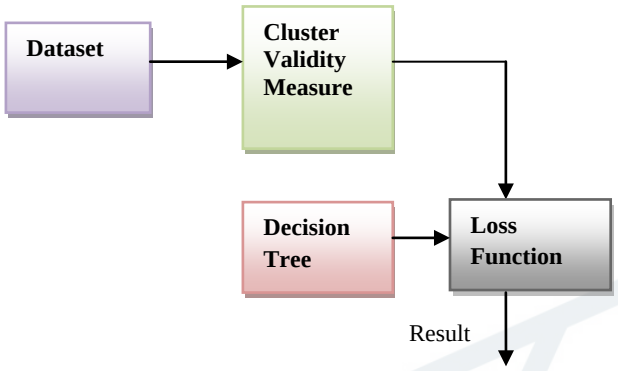

Figure 3 is proposed system ([6])

We choose K-means clustering because 1)it is data driven method relatively few assumptions on the distributions of the underlying data and 2) greedy search strategy of K-means guarantees at least a local minimum of the criterion function, thereby accelerating the convergence of clusters on large datasets.

A. Cluster Quality on Decision Theory:

Unsupervised learning method is the techniques we apply only data available are unlabeled, algorithms need to know the number of clusters. Cluster validity measures are Davies-Bouldin can help us assess whether a clustering method accurately presents the structure of the data set. There are several cluster indices to evaluate crisp and fuzzy clusteruing. Decision framework has been helpful in providing a better understanding of the classification model. Decision rough set model considers various classes of loss functions, the extension of the decision rough set model to multicategory is possible to construct a cluster validity measure by considering various loss functions based on decision theory. Within a given set of objects there may be clusters such that objects in the same cluster are more similar than those in different clusters. Clustering is to find the right groups or clusters for the given set of objects. To find right cluster we need exponential time comparisons has been proved to be NP-hard. For defining framework we assume partitions a set of objects $X=\{x_{1....}x_n\}$ into clusters $CS=\{c_1 \ldots c_k\}$, the k-means algorithm approximate the actual clustering. It is possible that each object may not necessarily belong to only one cluster. However there will be corresponding to each cluster within the clustering scheme, the centroid of the hypothetical core will be used Cluster core. Let core (c_i) be the core of the cluster c_i, which is used to calculate the centroid of the cluster. Any x_{1C} core (c_i)

cannot belong to other clusters. Therefore, core (c_i) can be considered the best representation of c_i to a certain extent.

B. Comparison of Clustering and Classification:

Clustering work well for finding unlabeled clusters in small to large data points K-means algorithm is its favorable execution time and the user has to know in advance how many clusters are to be searched, k-means is data driven is efficient for smaller data sets and anomaly detection. Instead of taking the mean value of the objects in a cluster as a reference point, a Medoid can be used, which is the most centrally located object in a cluster. Clustering requires the distance between every pair of objects only once and uses the distance at every stage of iteration.

Comparing to [8] clustering, classification algorithms performs efficient for complex datasets, noise and outlier detection such as algorithm designers have had much success with equal width method, equal depth method approaches to building class descriptions. It is chosen decision tree learners made popular by ID3, C4.5 and CART, because they are relatively fast and typically they produce competitive classifiers. In fact, the decision tree generator C4.5, a successor to ID3, has become a standard factor for comparison in machine learning research, because it produces good classifiers quickly. For non numeric datasets, the growth of the run time of ID3 (and C4.5) is linear in all examples.

The practical run time complexity of C4.5 has been determined empirically to be worse than O (e2) on some datasets. One possible explanation is based on the observation of Oates and Jensen (1998) that the size of C4.5 trees increases linearly with the number of examples. One of the factors of a in C4.5's run-time complexity corresponds to the tree depth, which cannot be larger than the number of attributes. Tree depth is related to tree size, and thereby to the number of examples. When compared with C4.5, the run time complexity of CART is satisfactory.

V. CONCLUSION

A cluster quality index based on decision theory, proposal uses a loss function to construct the quality index. Therefore, the cluster quality is evaluated by considering the total risk of classifying all the objects. Such a decision-theoretic representation of cluster quality may be more useful in business-oriented data mining than traditional geometry-based cluster quality measures. In addition to evaluating crisp clustering, the proposal is an evaluation measure for rough clustering. This is the first measure that takes into account special features of rough clustering that allow for an object to belong to more than one cluster. The measure is shown to be useful in determining important aspects of a clustering exercise such as determining the appropriate number of clusters and size of boundary region. The application of the measure to synthetic data with known number of clusters and boundary region provides credence to the proposal.

A real advantage of the decision-theoretic cluster validity measure is its ability to include monetary

considerations in evaluating a clustering scheme. Use of the measure to derive an appropriate clustering scheme for a promotional campaign in a retail store highlighted its unique ability to include cost and benefit considerations in commercial data mining. We can also extend it to evaluating other clustering algorithms such as fuzzy clustering. Such a cluster validity measure can be useful in further theoretical development in clustering.

VI. REFERENCES

[1] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, 1981. doi:10.1007/978-1-4757-0450-1_5

[2] D.L. Davies and D.W. Bouldin, "A Cluster Separation Measure," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 1, no 2, pp. 224-227, Apr. 1979. doi:10.1109/TPAMI.1979.4766909

[3] J.C. Dunn, "Well Separated Clusters and Optimal Fuzzy Partitions," J. Cybernetics, vol. 4, pp. 95-104, 1974. doi:10.1080/01969727408546059

[4] S. Hirano and S. Tsumoto, "On Constructing Clusters from Non- Euclidean Dissimilarity Matrix by Using Rough Clustering," Proc. Japanese Soc. for Artificial Intelligence (JSAI) Workshops, pp. 5-16, 2005.

[5] T.B. Ho and N.B. Nguyen, "Nonhierarchical Document Clustering by a Tolerance Rough Set Model," Int'l J. Intelligent Systems, vol. 17, no. 2, pp. 199-212, 2002. doi:10.1002/int.10016

[6] Rough Cluster Quality Index Based on Decision Theory Pawan Lingras, Member, IEEE, Min Chen, and Duoqian Miao IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 7, JULY 2009. doi:10.1109/TKDE.2008.236

[7] W. Pedrycz and J. Waletzky, "Fuzzy Clustering with Partial Supervision," IEEE Trans. Systems, Man, and Cybernetics, vol. 27, no. 5, pp. 787-795, Sept. 1997. doi:10.1109/3477.623232

[8] Partition Algorithms– A Study and Emergence of Mining Projected Clusters in High-Dimensional Dataset-International Journal of Computer Science and Telecommunications [Volume 2, Issue 4, July 2011]

[9] Jensen, D. D. and Cohen, P. R (1999), "Multiple Comparisons in Induction Algorithms," Machine Learning (to appear). Excellent discussion of bias inherent in selecting an input. Explore http://www.cs.umass.edu/~jensen/papers.

How to cite

B.Rajasekhar, B. Sunil Kumar, Rajesh Vibhudi, B.V.Rama Krishna, "Quality of Cluster Index Based on Study of Decision Tree ". *International Journal of Research in Computer Science*, 2 (1): pp. 39-43, December 2011. doi:10.7815/ijorcs.21.2011.013