LITERATURE SURVEY ON ENHANCING CLUSTER QUALITY

Mrs.S.THILAGAMANI
Asst.Professor/B.Tech(IT)
M.KUMARASAMY COLLEGE OF ENGINERING
Tamilnadu,India-639 113

Dr.N.SHANTHI
Professor/B.tech(IT)
K.S.R COLLEGE OF TECHNOLOGY
Tamilnadu,India-639 113

Abstract--In this paper, we extensively study about the important aspect of various Clustering techniques, *the cluster quality*. The goodness of clustering is measured in terms of cluster validity indices where the results of clustering are validated every time to give the maximum efficiency. The quality of clusters is measured in a decision-theoretic rough set oriented approach rather than the traditional geometry-based measures. Experiments are carried out with synthetic, standard and real world data for evaluating rough and crisp clustering. Also a new advancement in estimating the number of clusters in the analysis of gene expression data is studied. Here we follow a scheme called *System Evolution* to estimate the number of clusters based on Partitioning around medoids algorithm.

Index Terms-- Cluster validity, decision theory, rough-set based clustering, Cluster analysis, system evolution.

INTRODUCTION:

Cluster analysis or **clustering** is the assignment of a set of observations into subsets (called *clusters*) so that observations in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bio informatics.

Cluster validity is the measurement of goodness of a clustering relative to others created by Other clustering algorithms, or by the
same algorithms using different parameter values.

Cluster validation is very important issue in clustering analysis because the result of clustering needs to be validated in most applications. In most clustering algorithms, the number of clusters is set as user parameter. There are a lot of approaches to find the best number of clusters. Some validity indices partition validity to evaluate the properties of crisp structure imposed on the data. This includes the well known Dunn indices and Davies-Bouldin index. Partitional clustering algorithms divide up a data set into clusters or classes, where similar data objects are assigned to the same cluster whereas dissimilar data objects should belong to different clusters. In real applications there is very often no sharp boundary between clusters so that fuzzy clustering is often better suited for the data. Membership degrees between zero and one are used in fuzzy clustering instead of crisp assignments of the data to clusters. The most prominent fuzzy clustering algorithm is the fuzzy c-means, a fuzzification of k-Means. Another class of clustering algorithm called the decision theoretic rough set model which provides a better perspective of classification models. It considers various classes of loss functions. A loss function represents the loss (cost in money or loss in utility in some other sense) associated with an estimate being "wrong" (different from either a desired or a true value) as a function of a measure of the degree of wrongness. An important goal with large-scale gene expression studies is to find biologically important subsets of genes and samples. Clustering algorithms have been widely applied to this problem. These can be classified into partitioning and hierarchical clustering algorithms. Examples of hierarchical algorithms include agglomerative clustering and Partitioning algorithms include K-Means, Self-Organizing Maps, and Partitioning Around Medoids (PAM). With both types of algorithms, we are interested in the number of clusters. In a hierarchical tree, this corresponds with the lowest level at which the clusters are still significant. A comprehensive survey of methods for estimating the number of clusters is given in Milligan & Cooper (1985).

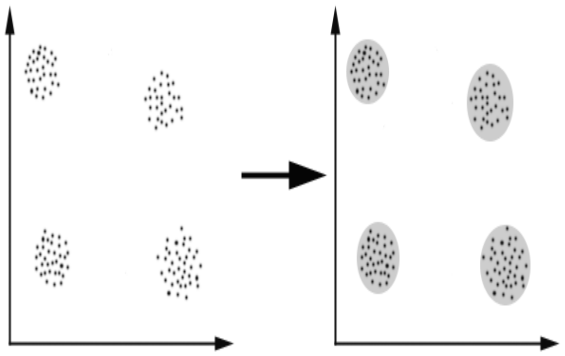

Figure 1:Clustering

Crisp Clustering algorithm:

Every object is assigned exactly one clusters. The k-medoid cluster algorithm can be described in four steps:

1. k predefined clusters
2. selecting of k representative objects and clustering the remaining objects
3. improving the set of representative objects and hence clustering
4. crisp assignment of object x_i to cluster K_j (crisp mapping)

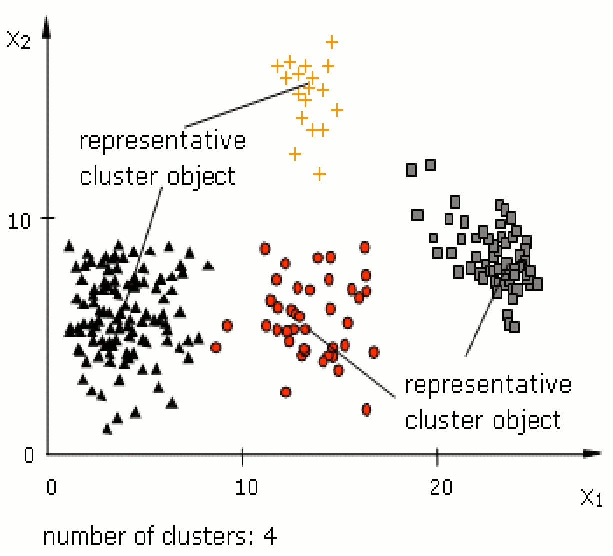

Figure 2: Crisp clusters.

Fuzzy Cluster algorithms:

The objects are assigned with a gradual membership to the clusters. The minimization of the objective function is yield with the following procedure

1. k predefined clusters
2. initializing membership values
3. iterative improving the membership values of the objects
4. non-linear optimization problem with constraints for a objective function
5. fuzzy mapping of objects x_i to clusters K_j
6. grade of membership is defined by membership function

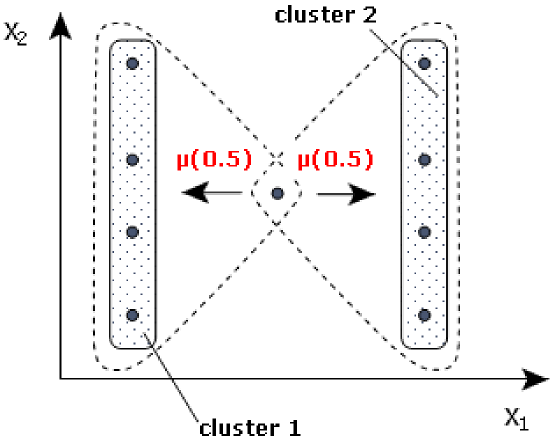

Figure 3:Fuzzy clusters.

Rough set based clustering:

The necessity of Crisp clustering algorithms is that each object should precisely belong to one cluster. This is too *restrictive* in many applications. Fuzzy set representation like Fuzzy C-means make it possible for an object to belong to multiple clusters with a degree of membership. This may be too *descriptive* for interpreting clustering results. Rough set based clustering provides a solution that is less *restrictive* than traditional clustering and less *descriptive* than the fuzzy clustering.

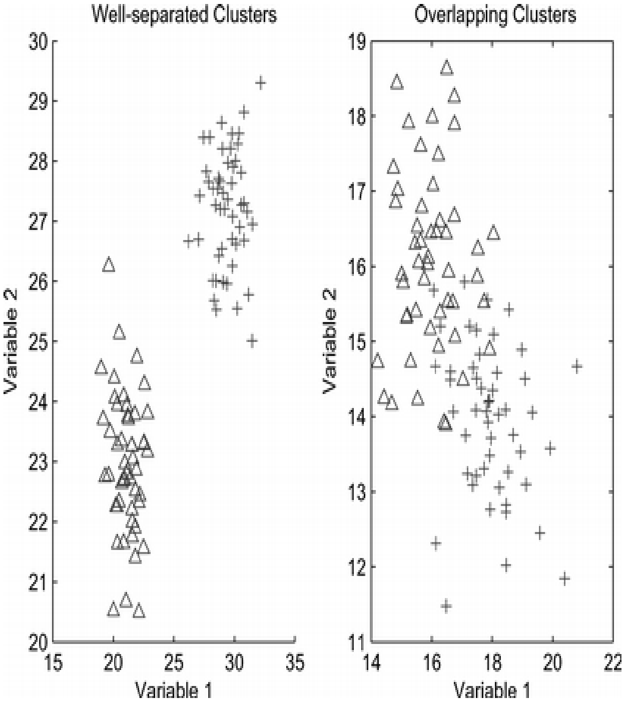

Figure 4:Rough set Clustering.

CLUSTER QUALITY:

- **Dunn's Validity Index:**

This technique (Dunn, 1974) is based on the idea of identifying the cluster sets that are compact and well separated. For any partition of clusters, where c_i represent the *i*-cluster of such partition, the Dunn's validation index, *D*, could be calculated with the following formula:

$$D = \min_{1\le i\le n}\left\{\min_{\substack{1\le j\le n\\ i\ne j}}\left\{\frac{d(c_i,c_j)}{\max_{1\le k\le n}\left(d'(c_k)\right)}\right\}\right\},$$

where $d(c_i,c_j)$ – distance between clusters c_i, and c_j (intercluster distance); $d'(c_k)\}$ – intracluster distance of cluster c_k, *n* – number of clusters. The minimum is calculating for number of clusters defined by the mentioned partition. The main goal of the measure is to maximise the intercluster distances and minimise the intracluster distances. Therefore, the number of cluster that maximise *D* is taken as the optimal number of the clusters.

- **Davies-Bouldin Validity Index:**

This index (Davies and Bouldin, 1979) is a function of the ratio of the sum of within-cluster scatter to between-cluster separation

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\ne j}\left\{\frac{S_n(Q_i)+S_n(Q_j)}{S(Q_i,Q_j)}\right\},$$

where n - number of clusters, S_n - average distance of all objects from the cluster to their cluster centre, $S(Q_i,Q_j)$ - distance between clusters centres. Hence the ratio is small if the clusters are compact and far from each other. Consequently, Davies-Bouldin index will have a small value for a good clustering.

- **Silhouette Validation Method:**

The Silhouette validation technique calculates the silhouette width for each sample, average silhouette width for each cluster and overall average silhouette width for a total data set. Using this approach each cluster could be represented by so-called silhouette, which is based on the comparison of its tightness and separation. The average silhouette width could be applied for evaluation of clustering validity and also could be used to decide how good the number of selected clusters is.

To construct the silhouettes S(i) the following formula is used:

$$S(i) = \frac{(b(i)-a(i))}{\max\{a(i),b(i)\}},$$

where *a(i)* –average dissimilarity of *i*-object to all other objects in the same cluster; *b(i)* – minimum of average dissimilarity *of i*-object to all objects in other cluster (in the closest cluster).

DECISION THEORETIC ROUGH SET MODEL:

Ever since the introduction of rough set theory by Pawlak in 1982, many proposals have been made to include probabilistic approaches into the theory. They include, for example, rough set based probabilistic classification, 0.5 probabilistic rough set model, decision-theoretic rough set models, variable precision rough set models, rough membership functions, parameterized rough set models, and Bayesian rough set models. The outcome of these studies increases our appreciative of the rough set theory and its domain of applications.

The decision-theoretic rough set models and the variable precision rough set models were proposed in the early 1990’s. The two models are formulated differently in order to generalize the 0.5 probabilistic rough set model. In fact, they produce the same rough set approximations. Their main differences lie in their respective treatment of the required parameters used in defining the lower and upper probabilistic approximations.

The decision-theoretic models scientifically calculate the parameters based on a loss function through the Bayesian decision procedure. The physical meaning of the loss function can be interpreted based on more practical notions of costs and risks. In contrast, the variable precision models regard the parameters as primitive notions and a user must supply those parameters. A lack of a systematic method for parameter assessment has led researchers to use many ad hoc methods based on trial and error.

The results and thoughts of the decision-theoretic model, based on the well established and semantically sound Bayesian decision procedure, have been successfully applied to many fields, such as data analysis and data mining, information retrieval, feature selection, web-based support systems, and intelligent agents. Some authors have generalized the decision-theoretic model to multiple regions.

Estimating the number of clusters Using System Evolution:

The method of System evolution is applied when there are small clusters near large clusters or slight overlapping between clusters. It analyses the cluster structures and estimate NC based on whether it is stable that two potential clusters are separated or merged. The method categorically analyses the separability of two closest clusters among k potential clusters called twin clusters. By analyzing the separability of twin clusters and the process that a dataset is divided into k clusters from small to big NC.

CONCLUSION:

Cluster quality based on decision theoretic rough set model includes a loss function to calculate the quality index. Such a depiction would be more useful in business-oriented data mining applications. The advantage of using Rough Set clustering over conventional methodologies is that it is less restrictive than crisp clustering and less descriptive than fuzzy clustering. Rough clustering also allows the user to set a parameter called threshold to help determine optimal number of clusters.

References:

[1] J. Sander, M. Ester, H. Kriegel, and X. Xu, “Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications,” Data Mining and Knowledge Discovery, vol. 2,no. 2, pp. 169-194, 1998.

[2] J. Han and M. Kamber, Data Mining: Concepts and Techniques.Morgan Kaufmann, 2000.

[3] K.Y. Yip, D.W. Cheung, and M.K. Ng, “HARP: A Practical Projected Clustering Algorithm,” IEEE Trans. Knowledge and Data Eng., 2004.

[4] Xiao-Feng Wang and De-Shuang Huang, Senior Member, IEEE,” A Novel Density-Based Clustering Framework by Using Level Set Method IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 11, NOVEMBER 2009.

[5] Pawan Lingras, Member, IEEE, Min Chen, and Duoqian Miao,” Rough Cluster Quality Index Based on Decision Theory” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 7, JULY 2009

[6] Kaijun Wang, Jie Zheng, Junying Zhang, *Member, IEEE*, and Jiyang Don,” Estimating the Number of Clusters via System Evolution for Cluster Analysis of Gene Expression Data”, IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 13, NO. 5, SEPTEMBER 2009.

[7] Andrew K.C. Wong, Fellow, IEEE, and Gary C.L. Li,” Simultaneous Pattern and Data Clustering for Pattern Cluster Analysis”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 7, JULY 2008.

[8] Mi-Yen Yeh, Bi-Ru Dai, and Ming-Syan Chen, Fellow, IEEE,” Clustering over Multiple Evolving Streams by Events and Correlations”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 10, OCTOBER 2007

[9] Christian S. Jensen, Member, IEEE, Dan Lin, Student Member, IEEE, and Beng Chin Ooi, Member, IEEE,” Continuous Clustering of Moving Objects”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 9, SEPTEMBER 2007.

[10] Eugenio Cesario, Giuseppe Manco, and Riccardo Ortale,” Top-Down Parameter-Free Clustering of High-Dimensional Categorical Data”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 12, DECEMBER 2007

[11] J. Han and M. Kamber, Data Mining: Concepts and Techniques.Morgan Kaufmann, 2000.

[12] J. Sander, M. Ester, H. Kriegel, and X. Xu, “Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications,” Data Mining and Knowledge Discovery, vol. 2,no. 2, pp. 169-194, 1998.