# A Study On Rough Clustering

Dr.K.Thangadurai[1] M.Uma[2] Dr.M.Punithavalli[3]

GJCST Computing Classification
H.3.3, I.5.3

***Abstract*-Clustering of data is an important data mining application. However, the data contained in today's databases is uncertain in nature. One of the problems with traditional partitioning clustering methods is that they partition the data into hard bound number of clusters. There have been recent advances in algorithms for clustering uncertain data, Rough set based Indiscernibility relation combined with indiscernibility graph, leads to knowledge discovery in an elegant way as it creates natural clusters in data. In this thesis, rough K-means clustering is studied and compared with the traditional K-means and weighted K-Means clustering methods for different data sets available in UCI data repository**

*Keywords*-Clusters, Boundary, Iteration, Attributes, Centroid.

## I. Introduction

Clustering is a technique to group together a set of items having similar characteristic. There are two kinds of clusters to be discovered in web usage domain they are usage clusters and page clusters. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Clustering of pages will discover groups of pages having related content. This information is useful for internet search engines and web assistance providers. Clustering can be considered the most important unsupervised learning problem, so as every other problem of this kind, it deals with finding a structure in collection of unlabeled data. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Here the simple graphical example for that

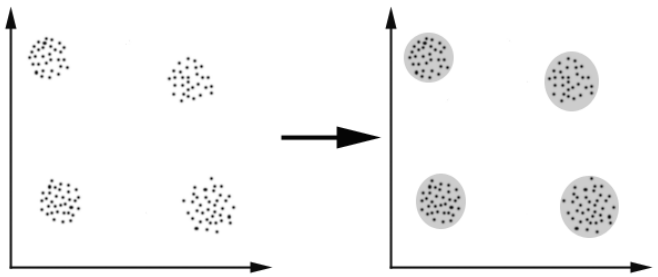

Figure 1: Cluster Analysis

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures[1].

## II. Goals of Clustering

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. There is no absolute best criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representative for homogeneous groups (data reduction), in finding natural clusters and describe their unknown properties (natural data types), in finding useful and suitable groupings (useful data class) or in finding unusual data objects (outlier detection).

### *A. The main requirements that a clustering algorithm should satisfy are*

Scalability , dealing with different types of attributes, discovering clusters with arbitrary shape, minimal requirements for domain knowledge to determine input parameters, ability to deal with noise and outliers, insensitivity to order of input records, high dimensionality, interpretability and usability[2]

### *B. Numbers of problems with clustering are*

Current clustering techniques do not address all the requirements adequately.

Dealing with large number of dimensions and large number of data items can be problematic because of time complexity.

The effectiveness of the method depends on the definition of distance.

If and obvious distance measure doesn't exist we must define it, which is not always easy, especially in multi-dimensional spaces.

The result of the clustering algorithm can be interpreted in different ways

## III. Clustering Algorithms

A large number of techniques have been proposed for forming clusters from distance matrices. The most important types are hierarchical techniques, optimization techniques and mixture models. We are going to discuss first two types here

*About-[1] Head in Computer Science, Govt. Arts College (Men), Krishnagiri, TN, India(e-mail; ktramprasad04@yahoo.com)*
*About-[2] Research Scholar, Dravidian University, Kuppam, A.P., India*
*About-[3] Director, Department of Computer Science, SRCW, Coimbatore, TN, India*

### C. Approaches to clustering

1. Centroid approaches, 2.hierarchical approaches.
*Centroid approaches:* We guess the centroids or central point in each cluster, and assign points to the cluster of their nearest centroid.
*Hierarchical approaches:* We begin assuming that each point is a cluster by itself. We repeatedly merge nearby clusters, using some measure of how close two clusters are, or how good a cluster the resulting group would be.

### D. Hierarchical Clustering Algorithms

A hierarchical algorithm yields a dendogram, representing the nested grouping of patterns and similarity levels at which groupings change. The dendogram can be broken at different levels to yield different clustering of the data. Most hierarchical clustering algoritms are variantas of the single-link, complete-link, and minimum-variance algorithms[3]. The single-link and complete-link algorithms are most popular. These two algorithms differ in the way of characterize the similarity between a parir of cluster.
In the single link method, the distance between two clusters is the minimum of the distances between all pairs of patterns drawn from the two clusters. In the complete link algorithm, the distance between two clusters is the minimum of all pair wise distance between patterns in the two clusters. The clusters obtained by the complete link algorithm are more compact then those obtained by the single link algorithm

## IV. PARTITIONAL ALGORITHMS

A partitional clustering algorithm obtains a single partition of the data instead of a clustering structure, such a dendogram produced by a hierarchical technique. Parititional methods have advantages in applications involving large data sets for which the construction of a dendogram is computationally prohibitive. A problem accomanying the use os partitional algorithm is the choice of the number of desired output clusters. The partitional technique usually produce clusters by optimizing a citerion function defined either locally or globally.

### A. Clustering Techniques

Let X be a data set, that is, X = {$x_i$=1……N}
Now let be the partition, $\Re$, of X into m sets, $C_j$ , j=1…m. These sets are called clusters and need to satisfy the following conditions:
• $C_i \neq \emptyset$ , i = 1... m
• $\bigcup_{i=1}^{m} C_i$ =X
• $C_i \cap C_j = \emptyset$, i $\neq$ j, i,j=1,….,m
It is important to say that the objects (vectors) contained in a cluster $C_i$ are more similar to each other and less similar to the objects (vectors) contained in the other clusters. The intention in the clustering algorithms is to join (or separate) the most similar (or dissimilar) objects of a data set X, it is necessary to apply a function that can make a quantitative measure among vectors [8].

Partitional algorithm is typically run multipel times with different starting states, and the best configuration obtained from all of the runs issued as the output clustering.

### B. Types of partitional Algorithms

- Squared Error Algorithms
- Graph-Theoretic Clustering
- Mixture-Resolving
- Mode-Seeking Algorithms

*K-Means Algorithm:*The K-means method aims to minimize the sum of squared distances between all points and the cluster centre. This procedure consists of the following steps, as described by Tou and Gonzalez.
1. Choose K initial cluster centre z1 (1), z2 (1)…zk (1).
2. At the k-th iterative step, distribute the samples {x} among the K clusters using the relation

$$x \in C_j(k) if \ \| x - z_j(k) \| < \| x - z_i(k) \|$$

For all i=1, 2…K; I ≠ j; where Cj (k) denotes the set of samples whose cluster centre is zj (k).
3. Compute the new cluster centre zj (k+1), j=1, 2…K such that the sum of the squared distances from all points in Cj (k) to the new cluster centre is minimized. The measure which minimizes this is simply the sample mean of Cj (k). Therefore, the new cluster centre is given by

$$z_j(k+1) = \frac{1}{N_j} \sum_{x \in C_j(k)} x$$

j=1, 2… K
Where $N_j$ is the number of samples in $C_j$ (k)
4. If $z_j$ (k+1) =$z_j$ (k) for j=1, 2…K then the algorithm has converged and the procedure is terminated.
5. Otherwise go to step 2

### C. Drawbacks of K-Means algorithm

The final clusters do not represent a global optimization result but only the local one, and complete different final clusters can arise from difference in the initial randomly chosen cluster centers.
We have to know how many clusters we will have at the first.

### D. Working Principle

The K-Means algorithm working principles are clearly explained in the following algorithm steps.

*Algorithm:*

1) ***Initialize the number of clusters k.***
2) ***Randomly selecting the centroids in the given data set ($c_1 c_2 \ldots c_k$)***
3) ***Compute the distance between the centroids and objects using the Euclidean Distance equation.***
   a. $d_{ij} = \left|\left|x_i - c_k\right|\right|^2$
4) ***Update the centroids.***
5) ***Stop the process when the new centroids are nearer to old one.***
   ***Otherwise, go to step-3.***

### E. Weighted K-Means Algorithm

Weighted K-Means algorithm is one of the clustering algorithms, based on the K-Means algorithm calculating with weights. A natural extension of the K-Means problem allows us to include some more information, namely, a set of weights associated with the data points. These might represent a measure of importance, a frequency count, or some other information. This algorithm is same as normal K-Means algorithm just adding the weights. Weighted K-Means attempts to decompose a set of objects into a set of disjoint clusters, taking into consideration the fact that the numerical attributes of objects in the set often do not come from independent identical normal distribution.

The weighted k-means algorithm uses weight vector to decrease the affects of irrelevant attributes and reflect the semantic information of objects. Weighted K-Means algorithms are iterative and use hill-climbing to find an optimal solution (clustering), and thus usually converge to a local minimum.

In the Weighted K-Means algorithm, the weights can be classified into two types.

*Dynamic Weights:* In the dynamic weights, the weights are changed during the program.

*Static Weights:* In the static weights, the weights are not changed during the program.

The Weighted K-Means algorithm is used to clustering the objects. Using this algorithm we can also calculating the weights dynamically and clustering the data in the dataset.

*Working Principle*

The Weighted K-Means algorithm working procedure is same as the procedure for K –Means algorithm but the only weight is included in the weighted k means algorithm. The working procedure is given in the following algorithm steps.

*Input: a set of n data points and the number of clusters (K)*

*Output: centroids of the K clusters*

1. *Initialize the number of clusters k.*
2. *Randomly selecting the centroids* $(c_1 c_2 \ldots c_k)$ *in the data set.*
3. *Choosing the Static weight* $W$ *,which is range from 0 to 2.5 or (5.0)*
4. *Find the distance between the centroids using the Euclidean Distance equation.*

$$d_{ij} = \left|\left|w_{.} * (x_{i} - c_k)\right|\right|^2$$

5. *Update the centroids using this equation.*
6. *Stop the process when the new centroids are nearer to old one. Otherwise, go to step-4.*

### F. Rough Set Clustering Algorithm

Rough sets were introduced by Zdzislaw Pawlak [6][7] to provide a systemic framework for studying imprecise and insufficient knowledge. Rough sets are used to develop efficient heuristics searching for relevant tolerance relations that allow extracting objects in data. An attribute-oriented rough sets technique reduces the computational complexity of learning processes and eliminates the unimportant or irrelevant attributes so that the knowledge discovery in database or in experimental data sets can be efficiently learned. Using rough sets, has been shown to be effective for revealing relationships within imprecise data, discovering dependencies among objects and attributes, evaluating the classificatory importance of attributes, removing data re-abundances, and generating decision rules [5]. Some classes, or categories, of objects in an information system cannot be distinguished in term of available attributes. They can only be roughly, or approximately, defined. The idea of rough sets is based on equivalence relations which partition a data set into equivalence classes, and consists of the approximation of a set by a pair of sets, called lower and upper approximations. The lower approximation of a given sets of attributes, can be classified as certainly belonging to the concept. The upper approximation of a set contains all objects that cannot be classified categorically as not belonging to the concept. A rough set also is defined as an approximation of a set, defined as a pair of sets: the upper and lower approximation of a set [7].

### G. Rough K-Means Algorithm

Step 0: Initialization. Randomly assign each data object to exactly one lower approximation. By definition (Property 2) the data object also belongs to the upper approximation of the same cluster.

Step 1: Calculation of the new means. The means are calculated as follows:

$$m_k = \begin{cases} w_l \sum_{X_k \in C_k} \frac{X_n}{|C_k|} + w_B \sum_{X_k \in C_k^B} \frac{X_n}{|C_k^B|} & for\ C_k^B \neq \emptyset. \\ w_l \sum_{X_k \in C_k} \frac{X_n}{|C_k|} & Otherwise. \end{cases}$$

where the parameters wl and wb define the importance of the lower approximation and boundary area of the cluster. The expression |Ck| indicates the numbers of data objects in lower approximation of the cluster and |CBk | = |Ck −Ck| is the number of data objects in the boundary areas.

Step 2: Assign the data objects to the approximations. (i) For a given data object Xn determine its closest mean mh:

$$d_{n,h}^{min} = d(X_n, m_k) = min_{k=1\ldots k}\, d(X_n, m_k)$$

Assign Xn to the upper approximation of the cluster h:Xn ∈ Ch.

(ii) Determine the means mt that are also close to Xn—they are not farther away from Xn than d( Xn,mh)where is a given threshold:

$$T = \{t: d(X_n, m_k) - d(X_n, m_h) \leq \varepsilon \cap h \neq k\}$$

If T= ∅ (Xn is also close to at least one other mean mt besides mh)
Then Xn ∈ Ct , ∀t ∈ T .
• Else Xn ∈ Ch.
Step 3: If the algorithms continue with Step 1.
Else STOP.

### *H. Experimental Results And Discussion*

The experimental analysis is carried out in this chapter by considering three different data sets from UCI data depository and the algorithms are validated through XIE – BIEN index

### *I. Xie-Beni Validity Index*

In this thesis, the Xie-Beni index has been chosen as the cluster validity measure because it has been shown to be able to detect the correct number of clusters in several experiments. Xie-Beni validity is the combination of two functions. The first calculates the compactness of data in the same cluster and the second computes the separateness of data in different clusters. Let S represent the overall validity index, π be the compactness and s be the separation of the rough k-partition of the data set. The Xie-Beni validity can now be expressed as:

$$\pi = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n} \mu_{ij}^{2} \parallel x - z_i \parallel^2}{n}$$

Where
And $s= (d_{min})^2$
dmin is the minimum distance between cluster centres, given by
dmin= minij ||zi-zj||
Where n is the number of users, k is the number of clusters, and Zi is the cluster centre of cluster Ci, wl is taken as 0.7 for the elements that are placed in lower approximation, wu is taken 0.3 for the elements that are placed in Upper approximation, µij is taken as 0.3 for the elements that are placed in boundary region. µij be the membership value of the user in boundary region. Smaller values of π indicate that the clusters are more compact and larger values of s indicate the clusters are well separated. Thus a smaller S reflects that the clusters have greater separation from each other and are more compact. In this thesis, Xie-Beni validity index is used to validate the clusters obtained after applying the clustering algorithms

## V. Conclusion

The K-Means, Weighted K-Means and Rough K-Means clustering algorithms have been studied and implemented. All the three algorithms are analyzed using the validity measure of Xie - Bien Index for three different UCI data sets. It is observed that Rough K-Means algorithm is performing well comparatively

## VI. References


1) Agrawal R, Imielinski T and Swami A. "Mining association rules between sets of items in large databases", In *Proc. 1993 Int. Conf. Management of Data (SIGMOD-93)*, 207-216. May 1993
2) Agrawal R, Mannila H, Srikant R, Toivonen H and Verkamo AI. " Fast discovery of association rules.", In: Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R. (Eds.) *Advances in Knowledge Discovery and Data Mining*, 307-328, 1996.
3) Bhattacharyya S, Pictet O, Zumbach G. "Representational semantics for genetic programming based learning in high-frequency financial data.", *Genetic Programming 1998: Proc. 3rd Annual Conf.*, 11-16. Morgan Kaufmann, 1998.
4) Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, USA, 2001.
5) Kusiak M, "Rough set theory: *A Data Mining tool for semiconductor manufacturing",* IEEE Transactions on Electronics Packaging Manufacturing 24 (1) (2001) 44-50
6) Lingras P and West C, *"Interval set clustering of web users with rough K-means",* Journal of Intelligent Information Systems 23 (1) (2004) 5-16.
7) Lingras P, Yan R and M. Hogo, "*Rough set based clustering: evolutionary, neural, and statistical approaches",* Proceedings of the First Indian International Conference on Artificial Intelligence (2003) 1074-1087.
8) Lingras, P. "*Rough Set Clustering for Web Mining"*, Proceedings of 2002 IEEE International Conference on Fuzzy Systems. 2002.
9) .Milligan G.W and Cooper M.C., "*An examination of procedures for determining the number of clusters in a data set",* Psychometrika, vol. 50, pp. 159-179, 1985.
10) Monmarche N. Slimane M, and Venturini G. Antclass, "Discovery of cluster in numeric data by an hybridization of an ant colony with the k-means algorithm", Technical Report 213, Ecole d' Ingenieurs en Informatique pour l'Industrie (E3i), Universite de Tours, Jan. 1999.