Adaptive K-Means Clustering to Handle Heterogeneous Data Using Basic Rough Set Theory

B.K. Tripathy[1], Adhir Ghosh[1], and G.K. Panda[2]

[1] VIT University, School of Computer Science and Engineering, Vellore, India
[2] Department of CSE and IT, MITS, Rayagada, Odisha, India
{tripathybk,adhir39}@rediffmail.com, gkpmail@sify.com

Abstract. Several cluster analysis techniques have been developed till the present to group objects having similar property or similar characteristics and K-means clustering is one of the most popular statistical clustering techniques proposed by Macqueen [12] in 1967. But this algorithm is unable to handle the categorical data and unable to handle uncertainty as well. But after proposing the rough set theory by Pawlak [15], we have an alternative way of representing sets whose exact boundary cannot be described due to incomplete information. As rough set has been widely used for knowledge representation, hence it can also be applied in classification and very helpful in clustering too. In real life data mining applications we do not have the crisp boundaries for clusters. So, in 2007 and 2009 Parmar et al [14] and Tripathy et al [16] proposed two algorithms MMR and MMeR using rough set theory but these two algorithms have the stability problem due to multiple runs and higher time complexity. In this paper we are proposing a new approach of k-means algorithm using rough set which can handle heterogeneous data and uncertainty as well.

Keywords: Classification, Cluster, Crisp boundaries, Heterogeneous data, Uncertainty.

1 Introduction

Cluster analysis is an important task in data mining. It is widely used in a lot of applications, including pattern recognition, data analysis, image processing, etc. By clustering, one can discover overall pattern distributions and interesting correlations among data attributes. Basically cluster analysis is applied on large heterogeneous data sets to make it into a smaller homogeneous data subsets that can be easily managed, separately modeled and analyzed [8]. For example, Wu et al. [18] developed a clustering algorithm specifically designed to handle the complexities of gene data that can estimate the correct number of clusters and find them. Jiang et al. [9] analyzed a variety of cluster techniques for complex gene expression data. Wong et al. [17] presented an approach used to segment tissues in a nuclear medical imaging method known as positron emission tomography (PET). Mathieu and Gibson [13] used cluster analysis as a part of a decision support tool for large-scale research and development planning to identify programs to participate in and to determine resource allocation. Finally, Haimov et al. [5] used cluster analysis to segment radar signals in

N. Meghanathan et al. (Eds.): CCSIT 2012, Part I, LNICST 84, pp. 193–202, 2012.

scanning land and marine objects. But all these algorithms are very specific. There are some general clustering algorithms like K-means [12], K-modes, fuzzy centroids etc. These algorithms suffer from problems like; they don't work when we have large data sets, missing value attributes and have irregular data shapes. Also, these algorithms can handle only numerical attributes. However, there are other algorithms like those proposed by Huang et al. [8], Gibson et al. [3], Guha et al. [4], Ganti et al. [2] and Dempster et al. [1]. These algorithms are not designed to handle uncertainty in data, which is a common issue in many real life applications. Using the concept of rough sets two algorithms were developed in 2007 and 2009 by Parmar et al. [14] and Tripathy et al. [16] respectively, which can handle both uncertainty and heterogeneous data. The time complexity of these two methods is high due to lots of calculations. So, in this paper we are proposing a new algorithm using general K-means methods and rough set theory in order to get the adaptive K-means algorithms using rough set which can handle hybrid data and uncertainty as well as its complexity is relatively low.

1.1 Basic Methods for Handling Categorical Data

Dempster et al. [1] presented a partition based clustering method, called the Expectation-Maximization (EM) algorithm. EM first randomly assigns different probabilities to each class or category, for each cluster. Then it successively adjusts the probabilities for maximizing the likelihood data those are given in the each cluster. After a large number of iterations, EM terminates at a locally optimal solution. Han et al. [6] proposed a clustering algorithm to cluster related items in a market database based on an association rule hyper graph. Also, we have some other categorical algorithms including K-modes [8] which extend the K-means algorithm. One advantage of K-modes algorithm is it is useful in interpreting the results [8]. However, these algorithms suffer from the problem of not being able to deal with uncertainty.

1.2 Handling Uncertainty

One of the first algorithms to deal with uncertainty is fuzzy K-means [11]. In this algorithm, each pattern or object is allowed to have membership functions to all clusters rather than having a distinct membership to exactly one cluster. Krishnapuram and Keller [10] propose a probabilistic approach to clustering in which the membership of a feature vector in a class has nothing to do with its membership in other classes and modified clustering methods are used to generate membership distributions. Krishnapuram et al. [11] have presented several fuzzy and probabilistic algorithms to detect linear and quadratic shell clusters. It may be noted that the initial work in handling uncertainty was based on numerical data. As they are unable to handle uncertainty in categorical data we cannot apply those algorithms in our real life applications as most of them depend on categorical data. Rough set theory has been used to develop clustering algorithms which handle uncertainty as well as deal with both categorical and numerical data [14, 16]. We shall discuss these approaches in the next section.

In real life situations, we find data with uncertainty, which may be numerical or categorical and so we need algorithms in order to deal such situations. Our effort in this paper adds one more algorithm in this direction, which is relatively faster than most of the other existing algorithms in this direction.

2 Rough Sets on Information Systems

Two of the most fruitful methods in dealing with uncertainty in data are the notion of Fuzzy Sets, introduced by Zadeh [19] and the notion of Rough Sets, introduced by Pawlak [15], which complement each other instead of being rivals. We formally introduce the notion of basic rough sets on information systems below.

Let U be a universe and X is a subset of U. Let **A** be the set of all the attributes of objects in U and **B** is a non-empty set of **A**. (U, A) is called an information system.

Definition 1 (Indiscernibility relation)
Given two tuples x, y ∈ *U* we say that x and y are indiscernible by the set of attributes **B** in **A** if and only if a(x) = a(y) for every a ∈ **B**. This relation is an equivalence relation on *U* and decomposes into disjoint equivalence classes and is denoted by Ind(**B**). For any x ∈ U, the equivalence class of x with respect to the set of attributes in **B** is denoted by $[x]_{Ind(\mathbf{B})}$.

Definition 2 (Approximation)
For any subset B of A and a set of objects X in U, the lower approximation of X with respect to B and the upper approximation of X with respect to B are defined as

$$\underline{X_{\mathbf{B}}} = \bigcup\{x / [x]_{Ind(B)} \subseteq X\} \tag{1}$$

$$\overline{X_B} = \bigcup\{x / [x]_{Ind(B)} \cap X \neq \varnothing\} \tag{2}$$

Definition 3 (Roughness)

The accuracy of estimation, is denoted by $R_B(X)$ and is defined by

$$R_B(X) = 1 - \left(\left| \underline{X_B} \right| \Big/ \left| \overline{X_B} \right| \right) \tag{3}$$

If $R_B(X) = 0$, X is crisp with respect to B, in other words, X is precise with respect to B. If $R_B(X) < 1$, X is rough with respect to B, That is, B is vague with respect to X.

Definition 4 (Relative roughness)
Given $a_i \in$ A, X is a subset of objects having one specifics value α of attribute a_i, $\underline{X_{a_j}(a_i = a)}$ and $\overline{X_{a_j}(a_i = a)}$ refer to the lower and upper approximation of X with respect to $\{a_j\}$, then $R_{a_j}(X)$ is defined as the roughness of X with respect to $\{a_j\}$, i.e.,

$$R_{a_j}(X/a_i=\alpha)=1-\left(\left|\underline{X_{a_j}(a_i=\alpha)}\right|\Big/\left|\overline{X_{a_j}(a_i=\alpha)}\right|\right) \quad \text{where } a_i, a_j \in A \text{ and } a_i \neq a_j. \tag{4}$$

Definition 5 (Mean relative roughness)
Let **A** have n attributes and $a_i \in A$. X be the subset of objects having a specific value α of the attribute a_i. Then we define the mean roughness for the equivalence class $a_i = \alpha$, denoted by MeR ($a_i = \alpha$) as

$$MeR(a_i=\alpha)=\left(\sum_{\substack{j=1\\ j\neq i}}^{n} R_{a_j}(X/a_i=\alpha)\right)/(n-1) \tag{5}$$

Definition 6 (Relative distance)
Given two objects P and Q of categorical data with n attributes, the relative distance of P and Q is denoted by RD(P, Q) and is defined as follows:

$$RD(P,Q)=\frac{1}{n}\sqrt{\sum_{i=1}^{n} RD(p_i,q_i)^2} \tag{6}$$

Here, p_i and q_i are values of P and Q respectively, under the i^{th} attribute a_i. We have, If cluster C_j has single object or in 0^{th} iteration, then

$$\left.\begin{aligned} RD(p_i,q_i)&=0, \text{ if } p_i=q_i\\ RD(p_i,q_i)&=1, \text{ if } p_i\neq q_i \end{aligned}\right\} \tag{7}$$

Else

$$RD(p_i,q_i)=\left|\text{avg centroid of } p_i \text{ in } C_j - \text{number of occurences in corresponding objects in } q_i\right| \tag{8}$$

When P and Q are numerical valued attributes, we have,
If the cluster C_j has single object or in 0^{th} iteration, then

$$\left.\begin{aligned} RD(p_i,q_i)&=0, \text{ if } p_i=q_i\\ RD(p_i,q_i)&=\left|\text{values of } p_i \text{ in } C_j - \text{values of } q_i\right|, \text{ if } p_i\neq q_i \end{aligned}\right\} \tag{9}$$

Else

$RD(p_i,q_i)$ is given by (8).

3 Generalized K-Means Method

K-means clustering is one of the most popular statistical techniques [7, 12]. Here we take, $X = \{x_1, x_2 \ldots x_n\}$. We now present the generalized K-means method.

GKM : A generalized algorithm of *K*-means clustering

GKM 1 : Give initial cluster centers $v_1, \ldots, v_k$. Let the cluster represented by v_i be G_i or G (v_i).

GKM 2 : Reallocate each object x to the nearest center v_i. $i = \min_{1\leq j\leq k} d(x, v_j)$

GKM 3 : After all objects are reallocated, update the cluster center.

$$v_i = \min_{v} \sum_{x_k \in G_i} d(x_k, v) \tag{10}$$

GKM 4 : Check for the convergent criterion. If not convergent, go to **GKM 2.**
End of GKM

The convergence criterion of K-means is when all the centroids of each cluster stabilize.

Incorporation of rough sets into K-means clustering requires the addition of the concept of lower and upper bounds. Calculation of the centroids of clusters from conventional K-Means needs to be modified to include the effects of lower as well as upper bounds. The modified centroid calculations for rough sets are presented in section 3.1.

3.1 Adaptation of K-Means to Rough Set Theory

As rough set needs the calculation of equivalence classes and lower and upper approximations, we need to calculate the upper and lower approximation of each cluster to update its centroid. The new centroid is as follows:

If Avg $\underline{R}X$ (of i^{th} attribute of cluster C_j) = $\varnothing$, then

$$v_i(C_k) = \frac{\frac{1}{\#(distinct\ objects)} \cdot \sum_{for\ each\ distinct\ object} \left\{ \sum_{\substack{j=1 \\ j \neq i}}^{n} \left(R_{a_j}(X / a_i = \alpha) \right) / (n-1) \right\}}{\frac{1}{\#(distinct\ objects)} \left(\sum_{For\ each\ distinct\ object} \overline{R}X \right)} \tag{11}$$

Else

$$v_i(C_k) = \frac{\frac{1}{\#(distinct\ objects)} \cdot \sum_{for\ each\ distinct\ object} \left\{ \sum_{\substack{j=1 \\ j \neq i}}^{n} \left(R_{a_j}(X / a_i = \alpha) \right) / (n-1) \right\}}{\left| \left\{ \frac{1}{\#(distinct\ objects)} \left(\sum_{For\ each\ distinct\ object} \underline{R}X \right) \right\} - \left\{ \frac{1}{\#(distinct\ objects)} \left(\sum_{For\ each\ distinct\ object} \overline{R}X \right) \right\} \right|} \tag{12}$$

The new centroid v_i of i^{th} attribute of cluster C_K is determined by the above equation. If the average of lower approximation is equal to null then the average mean roughness is divided by average upper approximation otherwise it is divided by absolute average difference of the lower and upper approximation. The mean roughness of attribute a_i is determined by the predefined value of α with respect to all other attributes in the cluster. Whether the object belongs to the lower approximation of the cluster or upper approximation of the cluster can be checked by the formula

$$\textbf{If } \left\{ \left\{ \frac{1}{n} \sum_{j=1}^{n} \left(\frac{1}{l} \sum_{i=1}^{l} \underline{R_j}(X_i) \right) \right\} - RD(V,X) \right\} < \left\{ \left\{ \frac{1}{n} \sum_{j=1}^{n} \left(\frac{1}{l} \sum_{i=1}^{l} \overline{R_j}(X_i) \right) \right\} - RD(V,X) \right\} \tag{13}$$

Then, the object X belongs to the lower approximation of the corresponding cluster. Here, l = #(distinct objects).

Else The object belongs to the upper approximation of the cluster. In case of a tie the object belongs to both the approximations.

4 Proposed Algorithm

In this section we propose our algorithm which is known as "Adaptive K-Means Clustering to Handle Heterogeneous Data" and is as follows:

Procedure RBKM (U,k)

1. **Begin**
2. Set current number of cluster CNC = k;
3. Set ParentNode =U;
4. Assign randomly selected objects to each cluster C_k;
5. **Label 1:**
   Reallocate each object x to the nearest cluster C_k;

$$K = \min \frac{1}{n} \sqrt{\sum_{i=1}^{n} RD(v_i, x)^2}$$

   //Updation of cluster centroids;
6. Update cluster();
   //Check for the convergent criterion;
7. If all the newly updated centroid value ≠ previous centroids value
8. Goto Label 1
9. **End**

Update cluster ()

1. Begin
2. For each $a_i \in A$ (i = 1 to n, where n is the number of attributes in A and j≠i)

   Calculate $Rough_{a_j}(a_i)$;

$$MeR(a_i = \alpha) = \left(\sum_{\substack{j=1 \\ j \neq i}}^{n} R_{a_j} \left(X / a_i = \alpha \right) \right) / (n - 1)$$

3. Next
   //Mean (MeR (a_i = α)) for each α;
4. Find the lower approximation of each a_i;
5. Make the average of lower approximation of each a_i for different α value;
6. Find the upper approximation of each a_i;
7. Make the average of upper approximation of each a_i for different α value;
8. If average of lower approximation = null, then

$$v_i(C_k) = \frac{\frac{1}{total\ distinct\ objects} \cdot \sum_{for\ each\ distinct\ object} \left\{ \sum_{\substack{j=1 \\ j \neq i}}^{n} \left(R_{a_j}(X / a_i = o_i) \right) / (n-1) \right\}}{\frac{1}{total\ distinct\ object} \left(\sum_{For\ each\ distinct\ object} \overline{RX} \right)}$$

9. Else

$$v_i(C_k) = \frac{\frac{1}{total\ distinct\ objects} \cdot \sum_{for\ each\ distinct\ object} \left\{ \sum_{\substack{j=1 \\ j \neq i}}^{n} \left(R_{a_j}(X / a_i = o_i) \right) / (n-1) \right\}}{\left| \left\{ \frac{1}{total\ distinct\ objects} \left(\sum_{For\ each\ distinct\ object} \overline{RX} \right) \right\} - \left\{ \frac{1}{total\ distinct\ objects} \left(\sum_{For\ each\ distinct\ object} \overline{RX} \right) \right\} \right|}$$

10. End

5 Empirical Analysis

The study data was obtained is a random dataset and has only 15 objects in it. This dataset has eight attributes including the object identifier as "Row". From the rest of seven attributes six are categorical and the other one is numerical. We choose the total number of cluster as four and initially we will pick the object number 3, 7, 10 and 13 as the initial centroids and will assign all the rest of the objects by the measure of relative distance

$$Min\left(\frac{1}{n} \sqrt{\sum_{i=1}^{n} RD(v_i, x)^2} \right)$$

So, after 0^{th} iteration the cluster structure will look like as follows:

Table 1. Cluster I

Row	A1	A2	A3	A4	A5	A6	SIZE
3	Small	Yellow	Soft	Fuzzy	Plush	Positive	2
4	Medium	Blue	Moderate	Fuzzy	Plastic	Negative	3
6	Big	Green	Hard	Smooth	Wood	Positive	17
8	Small	Yellow	Soft	Indefinite	Plastic	Positive	7
9	Big	Green	Hard	Smooth	Wood	Neutral	10

Table 2. Cluster II

Row	A1	A2	A3	A4	A5	A6	SIZE
7	Small	Yellow	Hard	Indefinite	Metal	Positive	14
1	Big	Blue	Hard	Indefinite	Plastic	Negative	4
5	Small	Yellow	Soft	Indefinite	Plastic	Neutral	21
11	Small	Yellow	Soft	smooth	Wood	Neutral	18
14	Small	Green	Hard	Metal	Wood	Neutral	7
15	Large	Yellow	hard	Metal	Plush	Negative	8

Table 3. Cluster III

Row	A1	A2	A3	A4	A5	A6	SIZE
10	Medium	Green	Moderate	Smooth	Plastic	Neutral	19
2	Medium	Red	Moderate	Smooth	Wood	Neutral	5

Table 4. Cluster IV

ROW	A1	A2	A3	A4	A5	A6	SIZE
13	Small	Red	Moderate	Indefinite	Wood	Neutral	5
12	Medium	Red	Moderate	Indefinite	Plastic	Positive	22

After this step we need to update the center of each cluster using the above proposed algorithm. Let us consider Table 1 for updating purpose. First we will calculate the mean roughness of each α (Big, Small and Medium) of attribute a_i as A1 with respect to the all other attributers and hence here total α is 3. To calculate roughness we need to calculate lower and upper approximations of a_i for each α and then need to find out the average to update the cluster center. This process will continue for each of the attributes with respect to the other attributes. After calculating the average roughness we will update the each cluster by the given equation (12) and (13). After all the calculations we get the final four clusters which are as follows:

Table 5. Final cluster I

Row	A1	A2	A3	A4	A5	A6	SIZE
4	Medium	Blue	Moderate	Fuzzy	Plastic	Negative	3
12	Medium	Red	Moderate	Indefinite	Plastic	Positive	22
15	Large	Yellow	hard	Metal	Plush	Negative	8
7	Small	Yellow	Hard	Indefinite	Metal	Positive	14

Table 6. Final cluster II

Row	A1	A2	A3	A4	A5	A6	SIZE
11	Small	Yellow	Soft	smooth	Wood	Neutral	18
14	Small	Green	Hard	Metal	Wood	Neutral	7
10	Medium	Green	Moderate	Smooth	Plastic	Neutral	19
3	Small	Yellow	Soft	Fuzzy	Plush	Positive	2
8	Small	Yellow	Soft	Indefinite	Plastic	Positive	7

Table 7. Final cluster III

Row	A1	A2	A3	A4	A5	A6	SIZE
5	Small	Yellow	Soft	Indefinite	Plastic	Neutral	21
13	Small	Red	Moderate	Indefinite	Wood	Neutral	5

Table 8. Final cluster IV

ROW	A1	A2	A3	A4	A5	A6	SIZE
1	Big	Blue	Hard	Indefinite	Plastic	Negative	4
6	Big	Green	Hard	Smooth	Wood	Positive	17
9	Big	Green	Hard	Smooth	Wood	Neutral	10
2	Medium	Red	Moderate	Smooth	Wood	Neutral	5

These are the final clusters those we got after the convergence criterion. For the experimental purpose we have used the small data set but it can be applied to the large data bases also to make the heterogeneous dataset into smaller homogeneous set.

6 Conclusions and Further Enhancement

In this paper we described modifications of K-means algorithm based on the concept of lower and upper bounds. This algorithm can handle databases with missing attribute values, hybrid type of values and having uncertainty. We have found it as an efficient method from empirical analysis. But, actual implementation and testing by using standard databases is likely to establish its position with respect to other related algorithms. Further enhancement can be done by providing better measure of relative distance (RD) and other measures of central tendency like standard deviation instead mean while computing relative roughness. Hybrid techniques like combinations of rough and fuzzy may improve the performance of this algorithm also.

References

1. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39(1), 1–38 (1977)
2. Ganti, V., Gehrke, J., Ramakrishnan, R.: CACTUS – clustering categorical data using summaries. In: Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 73–83 (1999)
3. Gibson, D., Kleinberg, J., Raghavan, P.: Clustering categorical data: an approach based on dynamical systems. The Very Large Data Bases Journal 8(3-4), 222–236 (2000)
4. Guha, S., Rastogi, R., Shim, K.: ROCK: a robust clustering algorithm for categorical attributes. Information Systems 25(5), 345–366 (2000)
5. Haimov, S., Michalev, M., Savchenko, A., Yordanov, O.: Classification of radar signatures by autoregressive model fitting and cluster analysis. IEEE Transactions on Geo Science and Remote Sensing 8(1), 606–610 (1989)
6. Han, E., Karypis, G., Kumar, V., Mobasher, B.: Clustering based on association rule hypergraphs. In: Workshop on Research Issues on Data Mining and Knowledge Discovery, pp. 9–13 (1997)
7. Hartigan, J.A., Wong, M.A.: Algorithm AS136: A K-Means Clustering Algorithm. Applied Statistics 28, 100–108 (1979)
8. Huang, Z.: Extensions to the k-means algorithm for clustering large data sets with categorical value. Data Mining and Knowledge Discovery 2(3), 283–304 (1998)

9. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: a survey. IEEE Transactions on Knowledge and Data Engineering 16(11), 1370–1386 (2004)
10. Krishnapuram, R., Keller, J.: A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems 1(2), 98–110 (1993)
11. Krishnapuram, R., Frigui, H., Nasraoui, O.: Fuzzy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation. IEEE Transactions on Fuzzy Systems 3(1), 29–60 (1995)
12. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press (1967)
13. Mathieu, R., Gibson, J.: A Methodology for large scale R&D planning based on cluster analysis. IEEE Transactions on Engineering Management 40(3), 283–292 (2004)
14. Parmar, D., Teresa, W., Jennifer, B.: MMR: An algorithm for clustering categorical data using Rough Set Theory. Data & Knowledge Engineering, 879–893 (2007)
15. Pawlak, Z.: Rough Sets- Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Norwell (1992)
16. Tripathy, B.K., Kumar, P.: MMeR: An algorithm for clustering Heterogeneous data using rough Set Theory. International Journal of Rapid Manufacturing (special issue on Data Mining) 1(2), 189–207 (2009)
17. Wong, K., Feng, D., Meikle, S., Fulham, M.: Segmentation of dynamic pet images using cluster analysis. IEEE Transactions on Nuclear Science 49(1), 200–207 (2002)
18. Wu, S., Liew, A., Yang, M.: Cluster analysis of gene expression data based on self-splitting and merging competitive learning. IEEE Transactions on Information Technology in Bio Medicine 8(1), 5–15 (2004)
19. Zadeh, L.A.: Fuzzy Sets. Information and Control 11, 338–353 (1965)