# *SDR: An Algorithm for Clustering Categorical Data Using Rough Set Theory*

B.K.Tripathy
SCSE, VIT University
Vellore- INDIA,
tripathybk@rediffmail.com

Adhir Ghosh
SCSE, VIT University
Vellore-INDIA,
adhirghosh39@yahoo.com

*Abstract*— **In the present day scenario, there are a large number of clustering algorithms available, to group objects having similar characteristics. But, the implementation of most of these algorithms is challenging due to the fact that most of the datasets involve categorical data values. Again, those algorithms which are capable of handling categorical data are mostly unable to handle uncertainty and some of them are involved with the stability issues. This necessitated the development of algorithms for clustering categorical data while handling uncertainty. In an effort to solve these problems an algorithm, termed MMR [1] was proposed in 2007, which uses the basic rough set theory concepts to deal with the above problem in clustering categorical data. Later in 2009, another algorithm, termed MMeR was proposed [2], which is more efficient than MMR and also has the capability of handling heterogeneous data. In this paper, we further improve MMeR and propose an algorithm, which we call SDR (Standard Deviation Roughness) algorithm It is capable of handling heterogeneous data besides taking care of uncertainty. We establish its efficiency over many other algorithms using well known standard data sets for the purpose of testing and the purity ratio as the measure of efficiency.**

***Keywords- clustering; uncertainty; MMR; MMeR; SDR***

## I. Introduction

The basic objective of clustering is to group data or objects having similar characteristics in the same cluster and having dissimilarity into different clusters. It has been used in data mining tasks such as unsupervised classification and data summation. It is also used in segmentation of large heterogeneous data sets into smaller homogeneous subsets which is easily managed, separately modeled and analyzed [3]. The basic goal in cluster analysis is to discover natural groupings of objects [4]. Clustering techniques are used in many areas such as manufacturing, medicine, nuclear science, radar scanning and research and development. For example, Wu et al. [5] developed a clustering algorithm specifically designed for handling the complexity of gene data. Jiang et al. [6] analyze a variety of cluster techniques, which can be applied for gene expression data. Wong et al. [7] presented an approach used to segment tissues in a nuclear medical imaging method known as positron emission tomography (PET). Haimov et al. [8] used cluster analysis to segment radar signals in scanning land and marine objects. Finally Mathieu and Gibson [9] 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.

The problem with all of the above mentioned algorithms is that they mostly deal with numerical data sets; that is those databases having attributes with numeric domains .Most of the works in clustering focused on numerical attributes as they are very easy to handle and very easy to define similarity. But categorical data have multi-valued attributes. This, similarity can be defined as common objects, common values for the attributes and the association between two. In such cases horizontal co-occurrences (common value for the objects) as well as the vertical co-occurrences (common value for the attributes) need to be examined [5].

Other algorithms, those can handle categorical data have been proposed, including work by Huang[3], Gibson et al. [10], Guha et al. [6] and Dempster et al. [12]. While these algorithms are very helpful to make the clusters using categorical data, their disadvantages include the fact that they cannot deal with uncertainty. However, uncertainty has become an integral part of most of the real world applications now a days.

Therefore, there is a need for a robust algorithm that can handle uncertainty and categorical data together. In 2007 S. Parmar et al [1] proposed an algorithm which uses the basic rough set concepts in order to deal with uncertainty as well as categorical attribute values in data sets. In 2009, B.K.Tripathy et al [2] improved this algorithm in two ways. They modified the algorithm to increase its efficiency as well as provided a method to deal with both numerical and categorical data.

In this paper we propose an algorithm, called Standard Deviation Roughness (SDR), which is designed to deal with uncertainty and handling the categorical data. We use the concept of purity ratio to compare the performance of this proposed algorithm with other existing algorithms. We use standard data sets like the zoo data set, soya bean data set and the mushroom data set from the well known data repository to evaluate and compare the efficiencies of all these algorithms.

978-1-4244-9477-4/11/$26.00 ©2011 IEEE

## II. Definitions And Notations

The notion of rough sets as a model to capture impreciseness in data was introduced by Pawlak [13]. Since its inception many fruitful applications have been found in various fields. The basic assumption in rough set theory is that human knowledge depends upon their capability to classify objects. As classification of universes and equivalence relations are interchangeable notions, for mathematical reasons equivalence relations are used to define rough sets. A rough set is represented by a pair of crisp sets, called the lower approximation, which comprises of elements belonging to it and upper approximation, which comprises of elements possibly in the set with respect to the available information.

By a knowledge base, we understand a relation system K= (U, R),where U is a universal set and R is a family of equivalence relations or indiscernibility relations defined over U and K is called an approximation space. Elementary sets in K are equivalence classes of R and any definable set in K is a finite union of elementary sets in K.

Let U be a universe of discourse and A be a set of attributes. With every attribute $a \in A$ we associate a set $V_a$ of its values, called the domain of a.

Let B be a nonempty subset of A and V be the set of all attribute values.

**Definition 1** (Indiscernibility relation (Ind(B))): Any subset B of A determines a binary relation Ind(B) on U, which is called an indiscernibility relation and is defined as follows:

x Ind(B) y if and only if a(x) = a(y) for every $a \in B$, where a(x) denotes the value of attribute a for element x.

It is clear that Ind(B) is an equivalence relation.

**Definition 2** (Equivalence class ($[x_i]_{Ind(B)}$)): Given Ind(B), the set of objects $x_i$ having the same values for the set of attributes in B consists of an equivalences classes, $[x_i]_{Ind(B)}$. It is also known as elementary set with respect to B.

The indiscernibility relation Ind(B) will be used next to define approximations and other basic concepts of rough set theory.

**Definition 3** (Lower approximation): Given the set of attributes B in A, set of objects X in U, the lower approximation of X is defined as the union of all the elementary sets which are contained in X. That is

$$\underline{X_B} = \cup \{ x_i \mid [x_i]_{Ind(B)} \subseteq X \}$$

**Definition 4** (upper approximation): Given the set of attributes B in A, set of objects X in U, the upper approximation of X is defined as the union of the elementary sets which have a nonempty intersection with X.That is

$$\overline{X_B} = \cup \{ x_i \mid [x_i]_{Ind(B)} \cap X \neq \phi \}$$

**Definition 5** (Roughness): The ratio of the cardinality of the lower approximation and the cardinality of the upper approximation is defined as the accuracy of estimation, which is a measure of roughness. It is presented as

$$R_B(X) = 1- \frac{|\underline{X_B}|}{|\overline{X_B}|}$$

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 6 (**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 $\{a_j\}$, that is

$$R_{a_j}(X / a_i = \alpha) = 1- \frac{|\underline{X_{a_j}(a_i = \alpha)}|}{|\overline{X_{a_j}(a_i = \alpha)}|}, \text{ where } a_i,\ a_j \in A$$

and $a_i \neq a_j$.

**Definition 7** (Mean 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 $a_i$=α, denoted by MeR ($a_i$=α) as

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

**Definition 8** (Standard deviation): Let A and X be as above. Then we define the standard deviation for the equivalence class $a_i$=α, denoted by SD($a_i$=α) as

$$SD(a_i = \alpha) = \sqrt{(1/(n-1)) \sum_{i=1}^{n-1} (R_{a_i}(X / a_i = \alpha) - \text{MeR}(\text{a}_\text{i} = \alpha))^2}$$

**Definition 9** (Distance of relevance): Given two objects B and C of categorical data with n attributes, DR for relevance of objects is defined as follows:

$$DR(B,C) = \sum_{i=1}^{n} (b_i, c_i) .$$

Here, $b_i$ and $c_i$ are values of objects B and C respectively, under the i^th attribute $a_i$. Also, we have

1. DR($b_i$ , $c_i$) = 1 if $b_i \neq c_i$
2. DR($b_i$ , $c_i$) = 0 if $b_i = c_i$
3. DR($b_i$ , $c_i$)= $\frac{|eq_{B_i} - eq_{C_i}|}{no_i}$ if $a_i$ is a numerical attribute; where '$eq_{B_i}$' is the number assigned to the equivalence class that contains $b_i$. '$eq_{C_i}$' is similarly defined and '$no_i$' is the total number of equivalence classes in numerical attribute $a_i$.

In order to compare SDR with MMeR and MMR and all other algorithms which have taken initiative to handle categorical data we developed an implementation. The traditional approach for calculating purity of a cluster is given below:

Definition 10 (Purity ratio): The purity ratio for the class ‘i’ is given by

$$Purity\,(i)=\frac{\text{The number of data occuring in both the } i^{th}\text{cluster and its corresponding class}}{\text{The number of data in the data set}}$$

$$\text{Over all Purity}=\frac{\sum_{i=1}^{\#of\ clusters} Purity(i)}{\#of\ clusters}$$

## III. Proposed Algorithm

In this section we present our algorithm which we call SDR. The notations and definitions of concepts have been discussed in the previous section.

```
Procedure SDR(U, k)
Begin
Set current number of cluster CNC = 1
Set ParentNode = U
Loop1:
If CNC < k and CNC ≠1 then
ParentNode = Proc ParentNode (CNC)
End if
// Clustering the ParentNode
For each $a_i$ ∈A ( i = 1 to n, where n is the number of attributes in A)
Determine $[X_m]_{Ind(a_i)}$ (m = 1 to number of objects)
For each $a_j$ ∈A (j = 1 to n, where n is the number of the attributes in A, j≠i)
Calculate $Rough_{a_j}$ ($a_i$)
Next
MeR ($a_i$=α) $=(\sum_{j=1, j\neq i}^{n} R_{a_j}(X/a_i=\alpha))/(n-1)$.
Next
Apply standard deviation
$SD(a_i=\alpha)=\sqrt{(1/(n-1))\sum_{i=1}^{n-1}(R_{a_i}(X/a_i=\alpha)-\text{MeR}(a_i=\alpha))^2}$
Next
Set SDR =Min min {SD ($a_i$=$\alpha_1$),….SD ($a_i$=$\alpha_{k_j}$ ),
Where $k_j$ is the number of equivalence classes in Dom ($a_i$).
Determine splitting attribute $a_i$ corresponding to the Standard deviation-Roughness
Do binary split on the splitting attribute $a_i$
CNC = the number of leaf nodes
Go to Loop1:
End
Proc ParentNode (CNC)
Begin
Set i = 1
Do until i < CNC
If Avg-distance of cluster i is calculated
Goto label
else
n = Count (Set of Elements in Cluster i).
Avg-distance (i) = 2*($\sum_{j=1}^{n-1}\sum_{k=j+1}^{n}$ ( Distance of relevance between objects $a_j$ and $a_k$ ))/(n*(n -1))
label :
increment i
Loop
Determine Max (Avg-distance (i))
Return (Set of Elements in cluster i) corresponding to Max (Avg-distance (i))
End
```

## IV. Empirical Evaluation

We have implemented the algorithms using JAVA language and the results obtained there in are summarised below. Due to scarcity of space we are unable to produce the details of implementation. As mentioned earlier, the metric for measuring the efficiency of the algorithms is the concept of ‘purity ratio’ and as mentioned earlier, the higher the purity ratio the efficiency is higher. We take the three data sets of ‘Zoo Data Set’, ‘Soya bean Data Set’ and the ‘Mushroom Data Set’ from the UCI repository for the evaluation and comparison.

### *A. Experiment 1 (Zoo Data Set)*

The Zoo data set is comprised of 101 objects, where each data point represents the information of an animal in terms of 18 categorical attributes. All the animals in this data set are classified into 7 classes. So, the stopping criterion of SDR algorithm is same as MMR and MMeR algorithms and is set at seven clusters. TABLE I summarizes the formation of clusters on Zoo data set when it is applied on SDR algorithm. The column C1 to C7 represents the number of classes and other two columns represent the cluster number and purity ratio.

TABLE I. PURITY RATIO OF ZOO DATA SET

| Cluster Number | C1 | C2 | C3 | C4 | C5 | C6 | C7 | Purity |
|---|---|---|---|---|---|---|---|---|
| 1 | 19 | 3 | 0 | 0 | 0 | 0 | 2 | 0.7916 |
| 2 | 0 | 0 | 0 | 13 | 0 | 0 | 0 | 1 |
| 3 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 1 |
| 4 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 1 |
| 5 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 1 |
| 6 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 1 |
| 7 | 22 | 9 | 0 | 0 | 0 | 0 | 8 | 0.5 641 |
| **Overall Purity** | | | | | | | | **0.9079** |

So, the overall purity is = $\frac{\sum_{i=1}^{\#of\ clusters} Purity(i)}{\#of\ clusters}$ =0.9079. Thus the overall purity of the clusters is 90.79%. Kim et al. [14] have applied the fuzzy centroid method to the Zoo data set and compared the results with k-modes and fuzzy k-modes and Tripathy et al. [2] compared their algorithm with MMR and fuzzy based clustering algorithms. So, in this paper we can compare our algorithm SDR with all other five algorithms to check the efficiency and purity ratio which is summarizes in the TABLE II.

TABLE II. COMPARISON OF ZOO DATA SET WITTH OTHER ALGORITHMS

| Data Set | K-modes | Fuzzy k-modes | Fuzzy centroids | MMR | MMeR | SDR |
|---|---|---|---|---|---|---|
| Zoo | 0.60 | 0.64 | 0.75 | 0.787 | 0.902 | 0.907 |

As from the TABLE II, we can see that SDR algorithm has the highest purity ratio compared to the all other existing clustering algorithms which shows its superiority and efficiency with respect to all algorithms.

### *B. Experiment 2 (Soybean Data Set)*

Soybean data set contains 47 objects on the diseases of soybean. Each object can be classified as one of the four diseases namely, Diaporthe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot and Phytophthora Rot and is described by 35 categorical data. As this data set has four types of diseases, we should stop this running SDR algorithm after the formation of four clusters. Table III shows the formation of clusters as follows where the column names D1 to D4 represent the classes of disease and other two columns represent the cluster number and purity ratio.

TABLE III. PURITY RATIO OF SOYBEAN DATA SET

| Cluster Number | D1 | D2 | D3 | D4 | Purity |
|---|---|---|---|---|---|
| 1 | 0 | 10 | 0 | 0 | 1 |
| 2 | 10 | 0 | 0 | 0 | 1 |
| 3 | 0 | 0 | 7 | 18 | 0.72 |
| 4 | 0 | 0 | 2 | 0 | 1 |
| **Overall Purity** | | | | | **0.93** |

So, the overall purity is = $\frac{\sum_{i=1}^{\#of\ clusters} Purity(i)}{\#of\ clusters}$ =0.93

The overall purity ratio is 0.93 or 93%. Kim et al. [14] and Tripathy et al. [2] also compared this data set with other algorithms. So, we can compare our algorithm with this data set with other algorithms to check its superiority and efficiency. The comparison is as follows:

TABLE IV. COMPARISON OF SOYBEAN DATA SET WITH OTHER ALGORITHMS

| Data Set | K-modes | Fuzzy k-modes | Fuzzy centroids | MMR | MMeR | SDR |
|---|---|---|---|---|---|---|
| Soybean | 0.69 | 0.77 | 0.97 | 0.83 | 0.83 | 0.93 |

From this table we can see that the algorithm SDR has higher purity ratio than all other algorithms except the 'Fuzzy centroid' algorithm, which has a slightly better purity ratio. In this case MMR and MMeR have the same purity ratio but fuzzy centroid has the most significant purity ratio and is 0.97 which is almost very close to 1. Therefore in this data set Fuzzy centroids show its higher efficiency than all other existing categorical clustering algorithms. But in comparison to the difference of purity ratio between the two algorithms of SDR and 'Fuzzy Centroid' in experiment 1 above, this difference is very low. However, once again SDR has established its superiority over MMR, MMeR, which are the algorithms based on rough set theory. In fact, it is worth noting that 'Soya bean data set' is a relatively small data set in comparison to the other two data sets we consider in this paper, where SDR has established its superiority over all other existing methods including the 'Fuzzy centroid method'.

### *C. Experiment 3 (Mushroom Data Set)*

Our algorithm (SDR) has also been applied to the large data set (Mushroom). This Mushroom data set has 8124 number of objects where each data point represents the information of a mushroom in terms of 22 categorical data. All the data in this set are classified into 2 classes namely Edible or non poisonous and poisonous. When

MMeR and MMR algorithms are applied to this data set then it formed 20 clusters. So in SDR algorithm, the stopping criterion is 20 i.e. to compare the purity ratio with these algorithms our algorithm should stop itself after forming the 20 clusters. TABLE V shows the cluster formation of the Mushroom data set which is as follows.

TABLE V. PURITY RATIO OF MUSHROOM DATA SET

| Cluster Number | Poisonous | Non-Poisonous | Purity |
|---|---|---|---|
| 1 | 0 | 8 | 1 |
| 2 | 0 | 1296 | 1 |
| 3 | 0 | 24 | 1 |
| 4 | 0 | 144 | 1 |
| 5 | 0 | 72 | 1 |
| 6 | 336 | 0 | 1 |
| 7 | 192 | 0 | 1 |
| 8 | 1728 | 0 | 1 |
| 9 | 100 | 4 | 0.9615 |
| 10 | 96 | 96 | 0.5 |
| 11 | 192 | 0 | 1 |
| 12 | 0 | 72 | 1 |
| 13 | 1024 | 0 | 1 |
| 14 | 72 | 0 | 1 |
| 15 | 512 | 0 | 1 |
| 16 | 0 | 256 | 1 |
| 17 | 24 | 1636 | 0.9855 |
| 18 | 24 | 0 | 1 |
| 19 | 192 | 0 | 1 |
| 20 | 0 | 24 | 1 |
| **Overall Purity** | | | **0.9723** |

So, the overall purity is $= \dfrac{\sum_{i=1}^{\# of\ clusters} Purity(i)}{\# of\ clusters} = 0.9723$

The overall purity ratio is 0.9723 or 97.23%, which is close to 1. To get the superiority and efficiency of this algorithm we need to compare the purity ratio of this cluster with MMR and MMeR. The comparison is as follows.

TABLE VI. COMPARISON OF MUSHROOM DATA SET WITH OTHER ALGORITHMS

| Data Set | MMR | MMeR | SDR |
|---|---|---|---|
| Mushroom | 0.84 | 0.9641 | 0.9723 |

From the above table it is clear that SDR has the highest purity as compared to MMR and MMeR and is 97.23 %. So in this data set also SDR shows its superiority and efficiency over MMR and MMeR algorithms.

From the all example given in this section we can conclude that SDR is the most efficient clustering algorithm we have in the domain of categorical clustering. Though fuzzy centroid gives a slight better result than SDR, we find that the overall performance of SDR is superior to all the algorithms dealing with numeric or categorical data and handling uncertainty in data is inherent due to its basis depending upon rough set theory.

## V. CONCLUSION

In this paper, we proposed a new algorithm called SDR, which is more efficient than most of the earlier algorithms including MMR and MMeR, which are recent algorithms developed in this direction. It handles uncertain data using rough set theory. Firstly, we have provided a method where both numerical and categorical data can be handled and secondly, by providing the distance of relevance we are getting much better results than MMR where they are choosing the table to be clustered, according to the number of objects. The comparison of purity ratio shows its superiority over MMeR. Future enhancements of this algorithm may be possible by considering hybrid techniques like rough-fuzzy clustering or fuzzy-rough clustering.

## REFERENCES

[1] D Parmar, Teresa Wu, Jennifer B, "MMR: An algorithm for clustering categorical data using Rough Set Theory", Data & Knowledge Engineering, 63 (2007), pp.879 - 893.

[2] B.K.Tripathy and M S Prakash Kumar Ch.: MMeR: An algorithm for clustering Heterogeneous data using rough Set Theory, International Journal of Rapid Manufacturing (special issue on Data Mining) (Switzerland),vol.1, no.2, (2009), pp.189-207.

[3] Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery 2 (3) (1998), pp. 283–304.

[4] R. Johnson, W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, New York, (2002).

[5] S. Wu, A. Liew, H. Yan, M. Yang, Cluster analysis of gene expression data based on self-splitting and merging competitive learning, IEEE Transactions on Information Technology in BioMedicine 8 (1) (2004), pp. 5–15.

[6] D. Jiang, C. Tang, A. Zhang, Cluster analysis for gene expression data: a survey, IEEE Transactions on Knowledge and Data Engineering 16 (11) (2004), pp. 1370–1386.

[7] K. Wong, D. Feng, S. Meikle, M. Fulham, Segmentation of dynamic pet images using cluster analysis, IEEE Transactions on Nuclear Science 49 (1) (2002), pp. 200–207.

[8] S. Haimov, M. Michalev, A. Savchenko, O. Yordanov, Classification of radar signatures by autoregressive model fitting

and cluster analysis, IEEE Transactions on Geo Science and Remote Sensing 8 (1) (1989), pp. 606–610.

[9] R. Mathieu, J. Gibson, A Methodology for large scale R&D planning based on cluster analysis, IEEE Transactions on Engineering Management 40 (3) (2004), pp. 283–292.

[10] D. Gibson, J. Kleinberg, P. Raghavan, Clustering categorical data: an approach based on dynamical systems, The Very Large Data Bases Journal 8 (3–4) (2000), pp. 222–236.

[11] S. Guha, R. Rastogi, K. Shim, ROCK: a robust clustering algorithm for categorical attributes, Information Systems 25 (5) (2000), pp. 345–366.

[12] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society 39 (1) (1977), pp. 1–38.

[13] Zdzislaw Pawlak, Rough Sets, Int. Jour of Computer and information Sciences, vol.11, (1982), pp.341- 356.

[14] Zdzislaw Pawlak, Rough Sets- Theoretical Aspects of Reasoning About Data. Norwell: Kluwar Academic Publishers, (1992).

[15] D. Kim, K. Lee, D. Lee, Fuzzy clustering of categorical data using fuzzy centroids, Pattern Recognition Letters 25 (11) (2004), pp. 1263–1271.