Available online at www.pelagiaresearchlibrary.com

Pelagia Research Library

Advances in Applied Science Research, 2011, 2 (3): 314-326

Pelagia Research Library

ISSN: 0976-8610
CODEN (USA): AASRFC

SSDR: An Algorithm for Clustering Categorical Data Using Rough Set Theory

B. K. Tripathy and *Adhir Ghosh

School of Computer Science and Engineering, VIT University, Vellore, Tamil Nadu, India

---

ABSTRACT

In the present day scenario, there are large numbers of clustering algorithms available to group objects having similar characteristics. But the implementations of many of those algorithms are challenging when dealing with categorical data. While some of the algorithms available at present cannot handle categorical data the others are unable to handle uncertainty. Many of them have the stability problem and also have efficiency issues. This necessitated the development of some algorithms for clustering categorical data and which also deal with uncertainty. In 2007, an algorithm, termed MMR was proposed [3], which uses the rough set theory concepts to deal with the above problems in clustering categorical data. Later in 2009, this algorithm was further improved to develop the algorithm MMeR [2] and it could handle hybrid data. Again, very recently in 2011 MMeR is again improved to develop an algorithm called SDR [22], which can also handle hybrid data. The last two algorithms can handle both uncertainties as well as deal with categorical data at the same time but SDR has more efficiency over MMeR and MMR. In this paper, we propose a new algorithm in this sequence, which is better than all its predecessors; MMR, MMeR and SDR, and we call it SSDR (Standard deviation of Standard Deviation Roughness) algorithm. This takes both the numerical and categorical data simultaneously besides taking care of uncertainty. Also, this algorithm gives better performance while tested on well known datasets.

Keywords- Clustering, MMeR, MMR, SDR, SSDR, uncertainty.

INTRODUCTION

The basic objective of clustering is to group data or objects having the similar characteristics in the same cluster and having dissimilarity with other 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 [8]. The basic goal in cluster analysis is to discover natural groupings of objects [11]. Clustering techniques are used in many areas such as manufacturing,

medicine, nuclear science, radar scanning and research and also in development. For example, Wu et al. [21] developed a clustering algorithm specifically designed for handling the complexity of gene data. Jiang et al. [13] analyze a variety of cluster techniques, which can be applied for gene expression data. Wong et al. [16] presented an approach used to segment tissues in a nuclear medical imaging method known as positron emission tomography (PET). Haimov et al. [20] used cluster analysis to segment radar signals in scanning land and marine objects. Finally Mathieu and Gibson [19] used the 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 the above mentioned algorithms is that they mostly deal with numerical data sets that are those databases having attributes with numeric domains .The basic reason for dealing with numerical attributes is that these are very easy to handle and also it is easy to define similarity on them. 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) can be examined [21].

Other algorithms, those can handle categorical data have been proposed including work by Huang[3], Gibson et al. [4], Guha et al. [13] and Dempster et al. [1]. While these algorithms or methods are very helpful to form the clusters from categorical data they have the disadvantage that they cannot deal with uncertainty. However, in real world applications it has been found that there is often no sharp boundary between clusters. Recently some work has been done by Huang [8] and Kim et al. [14] where they have developed some clustering algorithms using fuzzy sets, which can handle categorical data. But, these algorithms suffer from the stability problem as they do not provide satisfactory values due to the multiple runs of the algorithms.

Therefore, there is a need for a robust algorithm that can handle uncertainty and categorical data together. In this sequence S. Parmar et al [3] in 2007, B.K.Tripathy et al [2] in 2009 and [22] in 2011 proposed three algorithms which can deal with both uncertainty and categorical attributes together. But the efficiency and stability come into play when Purity ratio is measured. The purity ratios of MMR, MMeR and SDR are in the increasing order.

In this paper, a new algorithm called Standard Deviation of Standard Deviation Roughness (SSDR) algorithm is proposed, which has higher purity ratio than all the previous algorithms in this series and previous to that. We establish the superiority of this algorithm over the others by testing them on a familiar data base, the zoo data set taken from the UCI repository.

MATERIALS AND METHODS

2.1 Materials

In this section we first present the literature review as the basis of the proposed work, the definitions of concepts to be used in the work and also present the notations to be used.

2.1.1 Literature Review

In this section we present the literature of existing categorical clustering algorithms. Dempster et al. [1] presents a partitional clustering method, called the Expectation-Maximization (EM) algorithm. EM first randomly assigns different probabilities to each class or category, for each cluster. These probabilities are then successively adjusted to maximize the likelihood of the data

given the specified number of clusters. Since the EM algorithm computes the classification probabilities, each observation belongs to each cluster with a certain probability. The actual assignment of observations to a cluster is determined based on the largest classification probability. After a large number of iterations, EM terminates at a locally optimal solution. Han et al. [26] propose a clustering algorithm to cluster related items in a market database based on an association rule hypergraph. A hypergraph is used as a model for relatedness. The approach targets binary transactional data. It assumes item sets that define clusters are disjoint and there is no overlap amongst them. However, this assumption may not hold in practice as transactions in different clusters may have a few common items. K-modes [8] extend K-means and introduce a new dissimilarity measure for categorical data. The dissimilarity measure between two objects is calculated as the number of attributes whose values do not match. The K-modes algorithm then replaces the means of clusters with modes, using a frequency based method to update the modes in the clustering process to minimize the clustering cost function. One advantage of K-modes is it is useful in interpreting the results [8]. However, K-modes generate local optimal solutions based on the initial modes and the order of objects in the data set. K-modes must be run multiple times with different starting values of modes to test the stability of the clustering solution. Ralambondrainy [15] proposes a method to convert multiple categories attributes into binary attributes using 0 and 1 to represent either a category absence or presence, and to treat the binary attributes as numeric in the K-means algorithm. Huang [8] also proposes the K-prototypes algorithm, which allows clustering of objects described by a combination of numeric and categorical data. CACTUS (Clustering Categorical Data Using Summaries) [23] is a summarization based algorithm. In CACTUS, the authors cluster for categorical data by generalizing the definition of a cluster for numerical attributes. Summary information constructed from the data set is assumed to be sufficient for discovering well-defined clusters. CACTUS finds clusters in subsets of all attributes and thus performs a subspace clustering of the data. Guha et al. [6] propose a hierarchical clustering method termed ROCK (Robust Clustering using Links), which can measure the similarity or proximity between a pair of objects. Using ROCK, the number of ‘‘links’’ are computed as the number of common neighbors between two objects. An agglomerative hierarchical clustering algorithm is then applied: first, the algorithm assigns each object to a separate cluster, clusters are then merged repeatedly according to the closeness between clusters, where the closeness is defined as the sum of the number of ‘‘links’’ between all pairs of objects. Gibson et al. [4] propose an algorithm called STIRR (Sieving Through Iterated Relational Reinforcement), a generalized spectral graph partitioning method for categorical data. STIRR is an iterative approach, which maps categorical data to non-linear dynamic systems. If the dynamic system converges, the categorical data can be clustered. Clustering naturally lends itself to combinatorial formulation. However, STIRR requires a non-trivial post-processing step to identify sets of closely related attribute values [23]. Additionally, certain classes of clusters are not discovered by STIRR [23]. Moreover, Zhang et al. [24] argue that STIRR cannot guarantee convergence and therefore propose a revised dynamic system algorithm that assures convergence. He et al. [7] propose an algorithm called Squeezer, which is a one-pass algorithm. Squeezer puts the first-tuple in a cluster and then the subsequent-tuples are either put into an existing cluster or rejected to form a new cluster based on a given similarity function. He et al. [25] explore categorical data clustering (CDC) and link clustering (LC) problems and propose a LCBCDC (Link Clustering Based Categorical Data Clustering), and compare the results with Squeezer and K-mode. In reviewing these algorithms, some of the methods such as STIRR and EM algorithms cannot guarantee the convergence while others have scalability issues. In addition, all of the algorithms have one common assumption: each object can be classified into only one cluster and all objects have the same degree of confidence when grouped into a cluster [5]. However, in real world applications, it is difficult to draw clear

boundaries between the clusters. Therefore, the uncertainty of the objects belonging to the cluster needs to be considered.

One of the first attempts to handle uncertainty is fuzzy K-means [9]. 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 [18] 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. [17] present several fuzzy and probabilistic algorithms to detect linear and quadratic shell clusters. Note the initial work in handling uncertainty was based on numerical data. Huang [8] proposes a fuzzy K-modes algorithm with a new procedure to generate the fuzzy partition matrix from categorical data within the framework of the fuzzy K-means algorithm. The method finds fuzzy cluster modes when a simple matching dissimilarity measure is used for categorical objects. By assigning confidence to objects in different clusters, the core and boundary objects of the clusters can be decided. This helps in providing more useful information for dealing with boundary objects. More recently, Kim et al. [14] have extended the fuzzy K-modes algorithm by using fuzzy centroid to represent the clusters of categorical data instead of the hard-type centroid used in the fuzzy K-modes algorithm. The use of fuzzy centroid makes it possible to fully exploit the power of fuzzy sets in representing the uncertainty in the classification of categorical data. However, fuzzy K-modes and fuzzy centroid algorithms suffer from the same problem as K-modes, that is they require multiple runs with different starting values of modes to test the stability of the clustering solution. In addition, these algorithms have to adjust one control parameter for membership fuzziness to obtain better solutions. This necessitates the effort for multiple runs of these algorithms to determine an acceptable value of this parameter. Therefore, there is a need for a categorical data clustering method, having the ability to handle uncertainty in the clustering process while providing stable results. One methodology with potential for handling uncertainty is Rough Set Theory (RST) which has received considerable attention in the computational intelligence literature since its development by Pawlak in the 1980s. Unlike fuzzy set based approaches, rough sets have no requirement on domain expertise to assign the fuzzy membership. Still, it may provide satisfactory results for rough clustering. The objective of this proposed algorithm is to develop a rough set based approach for categorical data clustering. The approach, termed Standard deviation of Standard deviation roughness (SSDR), is presented and its performance is evaluated on large scale data sets.

2.1.2 Basics of rough sets

Most of our traditional tools for formal modeling, reasoning and computing are deterministic and precise in character. Real situations are very often not deterministic and they cannot be described precisely. For a complete description of a real system often one would require by far more detailed data than a human being could ever recognize simultaneously, process and understand. This observation led to the extension of the basic concept of sets so as to model imprecise data which can enhance their modeling power. The fundamental concept of sets has been extended in many directions in the recent past. The notion of Fuzzy Sets, introduced by Zadeh [10] deals with the approximate membership and the notion of Rough Sets, introduced by Pawlak [12] captures indiscernibility of the elements in a set. These two theories have been found to complement each other instead of being rivals. The idea of rough set consists of approximation of a set by a pair of sets, called the lower and upper approximations of the set. The basic assumption in rough set is that, knowledge depends upon the classification capabilities of human beings. Since every classification (or partition) of a universe and the concept of equivalence

relation are interchangeable notions, the definition of rough sets depends upon equivalence relations as its mathematical foundations [12].

Let U ($\neq\varnothing$) be a finite set of objects, called the universe and R be an equivalence relation over U. By U / R we denote the family of all equivalence classes of R (or classification of U) referred to as *categories* or *concepts* of R and $[x]_R$ denotes a category in R containing an element $x\in U$. By a Knowledge base, we understand a relation system k= (U, R), where U is as above and R is a family of equivalence relations over U.

For any subset P ($\neq\varnothing$) $\subseteq$ R, the intersection of all equivalence relations in P is denoted by IND (P) and is called the *indiscernibility relation over* P. The equivalence classes of IND (P) are called P- *basic knowledge* about U in K. For any $Q\in R$, Q is called a Q-elementary knowledge about U in K and equivalence classes of Q are called Q-*elementary concepts of knowledge* R. The family of P-basic categories for all $\varnothing \neq P \subseteq R$ will be called the *family of basic categories* in knowledge base K. By IND (K), we denote the family of all equivalence relations defined in k. Symbolically, IND (K) = {IND (P): $\varnothing \neq P \subseteq R$}.

For any $X \subseteq U$ and an equivalence relation $R \in$ IND (K), we associate two subsets, $\underline{R}X = \bigcup\{Y \in U / R : Y \subseteq X\}$ and $\overline{R}X = \bigcup\{Y \in U / R : Y \cap X \neq \varnothing\}$, called the R-*lower* and R-*upper approximations* of X respectively. The R-*boundary* of X is denoted by $BN_R(X)$ and is given by $BN_R(X) = \overline{R}X - \underline{R}X$. The elements of $\underline{R}X$ are those elements of U which can be certainly classified as elements of X employing knowledge of R. The borderline region is the undecidable area of the universe. We say X is *rough* with respect to R if and only if $\underline{R}X \neq \overline{R}X$, equivalently $BN_R(X) \neq \varnothing$. X is said to be R- *definable* if and only if $\underline{R}X = \overline{R}X$, or $BN_R(X) = \varnothing$. So, a set is rough with respect to R if and only if it is not R-definable.

2.1.3 Definitions

Definition 2.1.3.1 (Indiscernibility relation (Ind (B))): Ind (B) is a relation on U. Given two objects $x_i, x_j \in U$, they are indiscernible by the set of attributes B in A, if and only if a (x_i) = a (x_j) for every a∈B. That is, ($x_i, x_j \in$ Ind (B) if and only if $\forall$ a∈B where B⊆A, a (x_i) = a (x_j).

Definition 2.1.3.2 (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.

Definition 2.1.3.3 (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 2.1.3.4 (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

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

Definition 2.1.3. 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 2.1.3.6 (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

$$R_{a_j}(X / a_i = \alpha) = 1 - \frac{|\underline{X}_{a_j}(a_i = \alpha)|}{|\overline{X}_{a_j}(a_i = \alpha)|}$$, where $a_i, a_j \in A$ and $a_i \neq a_j$.

Definition 2.1.3.7 (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 $a_i = \alpha$, denoted by MeR $(a_i = \alpha)$ as

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

Definition 2.1.3.8 (Standard deviation) : After calculating the mean of each $a_i \in$ A, we will apply the standard deviation to each a_i by the formula

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

Definition 2.1.3.9 (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.

Definition 2.1.3.10 (Purity ratio) : 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.

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

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

METHODS

In this section we present the main algorithm of the paper and the experimental part deals with an example.

2.2.1 Proposed Algorithm

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

1. Procedure SSDR(U, k)
2. Begin
3. Set current number of cluster CNC = 1
4. Set ParentNode = U
5. Loop1:
6. If CNC < k and CNC ≠1 then
7. ParentNode = Proc ParentNode (CNC)
8. End if

// Clustering the ParentNode

9. For each $a_i \in A$ (i = 1 to n, where n is the number of attributes in A)
10. Determine $[X_m]_{Ind(a_i)}$ (m = 1 to number of objects)
11. For each $a_j \in A$ (j = 1 to n, where n is the number of the attributes in A, j≠i)
12. Calculate $Rough_{a_j}(a_i)$
13. Next
14. MeR ($a_i=\alpha$) = $(\sum_{\substack{j=1 \\ j\neq i}}^{n} R_{a_j}(X / a_i = \alpha)) / (n-1)$.
15. Next
16. Apply standard deviation

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

17. Next
18. Set SDR =SD {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).
19. Determine splitting attribute a_i corresponding to the Standard deviation-Roughness
20. Do binary split on the splitting attribute a_i

21. CNC = the number of leaf nodes
22. Go to Loop1:
23. End
24. Proc ParentNode (CNC)
25. Begin
26. Set i = 1
27. Do until i < CNC
28. If Avg-distance of cluster i is calculated
29. Goto label
30. else
31. n = Count (Set of Elements in Cluster i).
32. 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))
33. label :
34. increment i
35. Loop
36. Determine Max (Avg-distance (i))
37. Return (Set of Elements in cluster i) corresponding to Max (Avg-distance (i))
38. End

Experimental Part

In this section we present the experimental hybrid table which the characterization of various animals in terms of size, animality, color and age. In later section we will show the efficiency of this algorithm. The experimental table is as follows:

Table 1

ANIMAL NAME	SIZE	ANIMALITY	COLOUR	AGE
A1	Small	Bear	Black	25
A2	Medium	Bear	Black	16
A3	Large	Dog	Brown	9
A4	Small	Cat	Black	30
A5	Medium	Horse	Black	28
A6	Large	Horse	Black	5
A7	Large	Horse	Brown	7

Let us consider the value of k is 3 that is k=3 which mean the number of clusters will be 3. Initially the value of CNC is 1 and the value of the ParentNode is U which indicates, the initial value of ParentNode is whole table. So, we need to apply our algorithm three times to get the desired clusters.

Computational Part

So, initially CNC < k and CNC≠1 is false. So it will calculate the average distance of the parent node, but initially only one table we have so there is no need to calculate the average distance, directly we will calculate the roughness of each attribute relative to the rest of the attributes which is known as relative roughness. So, when i=1, the value of a_i is 'SIZE' that is a_i = size. This attribute has three distinct values 'Small', 'Medium' and 'Large' so considering α = 'Small'

first we get X={A1, A4} (where X is a subset of objects having one specific value α of attribute a_i) and considering j=2(as i ≠ j) we get a_j='Animality'. So the equivalence classes of a_j is {(A1, A2), A3, A4, (A5, A6, A7)} and the lower approximation of $X_{a_j}(a_i = \alpha)$ is given by $\underline{X_{a_j}(a_i = \alpha)}$ = {□} and the upper approximation of $X_{a_j}(a_i = \alpha)$ is given by $\overline{X_{a_j}(a_i = \alpha)}$ = {A1, A2, A4}. So, the roughness of a_i (when a_i ='SIZE' and α='Small') is given by

$$R_{X_j}(X / a_i = \alpha) = 1 - \frac{|X_{a_j}(a_i = \alpha)|}{|\overline{X_{a_j}(a_i = \alpha)}|} = 1 - \frac{0}{3} = 1$$

Now, by changing the value of j (when j=3, 4,) and keeping constant the value of a_i (a_i='size') and α (α='Small') we need to find the roughness of a_i relative to the attributes 'COLOR' (when j=3) and 'AGE' (when j=4) and is given by

$$R_{X_j}(X / a_i = \alpha) = 1 - \frac{|X_{a_j}(a_i = \alpha)|}{|\overline{X_{a_j}(a_i = \alpha)}|} = 1 - \frac{0}{5} = 1 \text{ when j=3 and } a_j\text{='COLOR'}$$

$$R_{X_j}(X / a_i = \alpha) = 1 - \frac{|X_{a_j}(a_i = \alpha)|}{|\overline{X_{a_j}(a_i = \alpha)}|} = 1 - \frac{2}{2} = 0 \text{ when j=4 and } a_j\text{='AGE'}$$

Now, to get the standard deviation of a_i (a_i='size') when α='Small' we need to find the mean of these values and is given by $\frac{1+1+0}{3} = \frac{2}{3}$. And applying standard deviation formula we get the value 0.4714 and will be stored in a variable.

This similar process will be continued by changing the value of α (for α='Medium' and 'Large') and keeping constant the value of a_i. And lastly we will get three standard deviation values for each different α. And again we will store those values in a variable. After calculating the SD (standard deviation) of each α we will take the minimum value of those different values of α and will store it in another variable.

The above procedure will be continued for each a_i (for a_i='ANIMALITY', 'COLOR' and 'SIZE' when i=2, 3 and 4) and the corresponding values will be stored in the variable. After completing the above step we will take those minimum values for next calculation. We will apply SD (standard deviation) to those minimum values to get the Splitting attributes. If the value of SD does not match with the minimum values then will we take the nearest minimum vale as the splitting attribute and will do the binary splitting that is we will divide this table into two clusters.

Let after splitting we have got two cluster **c1** and **c2** and **c1** contains 2 elements and **c2** contains 5 elements. So now we need to calculate the average distance to choose the clustering table for further calculation. This can be done by applying distance of relevance formula.

Let us see how we calculate DR (distance of Relevance). For example let us take two tuple A4 and A6 which is as follows

Table 2

ANIMAL NAME	SIZE	ANIMALITY	COLOR	AGE
A4	Small	Cat	Black	30
A6	Large	Horse	Black	5

Here B=A4 and C=A6 and DR (B, C) is defined as

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

$$= DR(b_{size}, c_{size}) + DR(b_{animality}, c_{animality}) + DR(b_{color}, c_{color}) + DR(b_{age}, c_{age})$$

So, $DR(b_{size}, c_{size}) = 0$ as $b_{size} \neq c_{size}$

$DR(b_{animality}, c_{animality}) = 0$ as $b_{animality} \neq c_{animality}$

$DR(b_{color}, c_{color}) = 1$ as $b_{color} = c_{color}$

But for $DR(b_{age}, c_{age})$ we need to follow some different method as 'AGE' is the numerical attribute. To calculate the DR of a numerical attribute we need to exclude that numerical attribute from that table and need to find the average equivalence class of all attributes. So, in this case we need to exclude the attribute 'AGE' first and then we have to find the average equivalence class.

So, the average equivalence class is (3+4+2)/3 = 3. In this case we have got a integer value but we can get a fraction also then we need to take either its floor value or its roof value.

Now we need to sort the attribute value of the attribute 'AGE'. After sorting in ascending order we get {5, 7, 9, 16, 25, 28, 30}. Now we will distribute these numbers into three sets which is as follows

Set 1 = {5, 7}
Set 2 = {9, 16}
Set 3 = {25, 28, 30}
Now we will calculate $DR(b_{age}, c_{age})$. In our case $b_{age} = 30$ and $c_{age} = 5$. So, we will put 3 and 1 in place of 30 and 5 as 30 belongs to the set 3 and 5 belongs to the set 1.

$$\text{So, } DR(b_{age}, c_{age}) = \frac{|3-1|}{total\_number\_of\_sets} = \frac{2}{3}$$

$$\text{Finally, } DR(B, C) = DR(b_{size}, c_{size}) + DR(b_{animality}, c_{animality}) + DR(b_{color}, c_{color}) + DR(b_{age}, c_{age})$$

$$= 0 + 0 + 1 + \frac{2}{3}$$

$$= 1.666667$$

So, in this way we will calculate the average distance of **C1** and **C2** and the cluster having the larger average distance we will take that particular cluster as the input for further calculation.

So, in this fashion we will apply this algorithm until we get the desired number of cluster. In our case we will stop when we will get **C3** because in our case the total number of clusters is 3.

RESULTS AND DISCUSSION

In this section we present the original result that is tested on ZOO dataset which was also taken by MMR, MMeR and SDR algorithm.

The ZOO data has 18 attributes and out them 15 are Boolean attribute, 2 are numeric and 1 is animal name and it has 101 objects. The total objects are divided into seven classes so; we need to stop when we will get seven clusters. After taking the ZOO dataset as the input we have got the following output which is as follows:

Table 3

Cluster Number	Class I	Class II	Class III	Class IV	Class V	Class VI	Class VII	Purity Ratio
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	0.5641
Overall Purity								**0 .9079**

3.2.1 Comparison of SSDR with MMeR, MMR, SDR and Algorithms based on FUZZY Set Theory

Till the development of MMR, the only algorithms which aimed at handling uncertainty in the clustering process were based upon fuzzy set theory[26].These algorithms based on fuzzy set theory include fuzzy K-modes, fuzzy centroids. The K-modes algorithm replaces the means of the clusters (K-means) with modes and uses a frequency based method to update the modes in the clustering process to minimize the clustering cost function. Fuzzy K-modes generates a fuzzy partition matrix from categorical data. By assigning a confidence to objects in different clusters, the core and boundary objects of the clusters are determined for clustering purposes. The fuzzy centroids algorithm uses the concept of fuzzy set theory to derive fuzzy centroids to create clusters of objects which have categorical attributes. But in MMR, MMeR and in SDR they have used rough sets concept to build those algorithms but as compared to efficiency MMeR is more efficient than MMR and less efficient than SDR but SSDR is much more efficient than other.

3.2.2 Emperical Analysis

The earlier algorithms for classification with uncertainty like K-modes, Fuzzy K-modes and Fuzzy centroid on one hand and MMR, MMeR and SDR on the other hand were applied to ZOO data sets. Table 4 below provides the comparison of purity for these algorithms on this datasets. It is observed that SSDR has a better purity than all other algorithms when applied on zoo data set.

As mentioned earlier, all the fuzzy set based algorithms face a challenging problem that is the problem of stability. These algorithms require great effort to adjust the parameter, which is used to control the fuzziness of membership of each data point. At each value of this parameter, the algorithms need to be run multiple times to achieve a stable solution.

MMR, MMeR and SDR on the other hand have no such problem. SSDR continues to have the advantages of MMR, MMeR and SDR over the other algorithms as mentioned above. But it has higher purity than MMR, MMeR and SDR which establishes its superiority over MMR, MMeR and SDR.

Table 4

DATA SET	K-modes	Fuzzy K-modes	Fuzzy centroids	MMR	MMeR	SDR	SSDR
ZOO	0.6	0.64	0.75	0.787	0.902	0.907	0.907*

*In this case we have got the same Purity ratio as compared to SDR but as standard deviation has better central tendency over mean or minimum it will give better result for other data sets. Manually it has been checked for a small data set that it is giving much better result than MMR, MMeR and SDR

CONCLUSION

In this paper, we proposed a new algorithm called SSDR, which is more efficient than most of the earlier algorithms including MMR, MMeR and SDR, 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] A. Dempster, N. Laird, D. Rubin, *Journal of the Royal Statistical Society* 39 (1) (**1977**) 1–38.
[2] B.K.Tripathy and M S Prakash Kumar Ch.: *International Journal of Rapid Manufacturing* (special issue on Data Mining) (Switzerland),vol.1, no.2, (**2009**), pp.189-207.
[3] D Parmar, Teresa Wu, Jennifer B, *Data & Knowledge Engineering* (**2007**)
[4] D. Gibson, J. Kleinberg, P. Raghavan, *The Very Large Data Bases Journal* 8 (3–4) (**2000**) 222–236.
[5] M. Halkidi, Y. Batistakis, M. Vazirgiannis, *Journal of Intelligent Information Systems* 17 (2–3) (**2001**) 107–145.
[6] S. Guha, R. Rastogi, K. Shim, *Information Systems* 25 (5) (**2000**) 345–366.
[7] Z. He, X. Xu, S. Deng, *Journal of Computer Science & Technology* 17 (5) (**2002**) 611–624.
[8] Z. Huang, *Data Mining and Knowledge Discovery* 2 (3) (**1998**) 283–304.
[9] E. Ruspini, *Information Control* 15 (1) (**1969**) 22–32.
[10] L.A. Zadeh, *Information and Control*, 11 (**1965**), pp.338-353.
[11] R. Johnson, W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, New York, **2002**.
[12] Zdzislaw Pawlak, Rough Sets- Theoretical Aspects of Reasoning About Data. Norwell: Kluwar Academic Publishers, (**1992**).
[13] D. Jiang, C. Tang, A. Zhang *IEEE Transactions on Knowledge and Data Engineering* 16 (11) (**2004**) 1370–1386.
[14] D. Kim, K. Lee, D. Lee, *Pattern Recognition Letters* 25 (11) (**2004**) 1263–1271.Mkm.
[15] H. Ralambondrainy, *Pattern Recognition Letters* 16 (11) (**1995**) 1147–1157.

[16] K. Wong, D. Feng, S. Meikle, M. Fulham, *IEEE Transactions on Nuclear Science* 49 (1) (**2002**) 200–207.
[17] R. Krishnapuram, H. Frigui, O. Nasraoui, *IEEE Transactions on Fuzzy Systems* 3 (1) (**1995**) 29–60.
[18] R. Krishnapuram, J. Keller, *IEEE Transactions on Fuzzy Systems* 1 (2) (**1993**) 98–110.
[19] R. Mathieu, J. Gibson, *IEEE Transactions on Engineering Management* 40 (3) (**2004**) 283–292.
[20] S. Haimov, M. Michalev, A. Savchenko, O. Yordanov, *IEEE Transactions on Geo Science and Remote Sensing* 8 (1) (**1989**) 606–610.
[21] S. Wu, A. Liew, H. Yan, M. Yang, *IEEE Transactions on Information Technology in BioMedicine* 8 (1) (**2004**) 5–15.
[22] Tripathy, B.K. and A.Ghosh: SDR: An Algorithm for Clustering Categorical Data Using Rough Set Theory, Communicated to the International IEEE conference to be held in Kerala, (**2011**).
[23] V., Ganti, J. Gehrke, R. Ramakrishnan, CACTUS – clustering categorical data using summaries, in: Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (**1999**), pp. 73–83.
[24] Y. Zhang, A. Fu, C. Cai, P. Heng, Clustering categorical data, in: Proceedings of the 16th International Conference on Data Engineering, (**2000**), pp. 305–324.
[25] Z. He, X. Xu, S. Deng, A link clustering based approach for clustering categorical data, Proceedings of the WAIM Conference, (**2004**). <http://xxx.sf.nchc.org.tw/ftp/cs/papers/0412/0412019.pdf>.
[26] E. Han, G. Karypis, V. Kumar, B. Mobasher, Clustering based on association rule hypergraphs, in: Workshop on Research Issues on Data Mining and Knowledge Discovery, (**1997**), pp. 9–13.