I.J. Intelligent Systems and Applications, **2017, 8, 25-33**
Published Online August 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2017.08.03

MMeMeR: An Algorithm for Clustering Heterogeneous Data using Rough Set Theory

B.K. Tripathy
School of Computing Science and Engineering VIT University, Vellore-632014 Tamil Nadu, India
E-mail: tripathybk@vit.ac.in

Akarsh Goyal, Rahul Chowdhury and Patra Anupam Sourav
School of Computing Science and Engineering VIT University, Vellore-632014 Tamil Nadu, India
E-mail: {akarsh.goyal15, chowdhuryrahul5, anupam.sourav}@gmail.com

Received: 13 February 2017; Accepted: 17 March 2017; Published: 08 August 2017

***Abstract*—**In recent times enumerable number of clustering algorithms have been developed whose main function is to make sets of objects having almost the same features. But due to the presence of categorical data values, these algorithms face a challenge in their implementation. Also some algorithms which are able to take care of categorical data are not able to process uncertainty in the values and so have stability issues. Thus handling categorical data along with uncertainty has been made necessary owing to such difficulties. So, in 2007 MMR algorithm was developed which was based on basic rough set theory. MMeR was proposed in 2009 which surpassed the results of MMR in taking care of categorical data and it could also handle heterogeneous values as well. SDR and SSDR were postulated in 2011 which were able to handle hybrid data. These two showed more accuracy when compared to MMR and MMeR. In this paper, we further make improvements and conceptualize an algorithm, which we call MMeMeR or Min-Mean-Mean-Roughness. It takes care of uncertainty and also handles heterogeneous data. Standard data sets have been used to gauge its effectiveness over the other methods.

***Index Terms*—**Categorical data, clustering, uncertainty, MMR, MMeR, SDR, SSDR.

I. Introduction

One of the major components of the Data Mining approach is to extract meaningful information from data sets. This is done by extracting the relevant information and converting them into knowledge for later use. There are so many methods like classification, clustering, rule mining, reduct formation etc. which are used to meet this. It is our aim in this paper to deal with the process of clustering.

Clustering [6] is a process in which the elements in a data set are decomposed into a number of groups on the basis of their similar characteristics. Nowadays, in some of the data sets no decision class is provided due to which we cannot classify them. Also sometimes the objects do not have any comprehendible relations between themselves. So, clustering proves to be useful in these cases. They minimize the similarity between the clusters and maximize the similarity inside the clusters.

The type of input data can be numeric or categorical or even hybrid. Several algorithms exist in literature for clustering categorical data. The number of algorithms developed for categorical data is relatively less. Only a limited number of values can be taken when a dataset is categorical. But generally the data sets are found to contain attributes which can be numeric, nominal or ordinal. Nominal, ordinal and interval-scaled values compose a categorical data set.

Modern day data sets have uncertainty inherent in them. One needs uncertainty based models to cluster such data sets. We find several uncertainty based models like the fuzzy set, intuitionistic fuzzy set, rough set and their hybrid models like the fuzzy rough sets and intuitionistic fuzzy models. Algorithms have been established using these models. But, multiple runs of the algorithms have stability issues. Hence, there arises a need for a robust process is felt when handling both the categorical data and uncertainty in it. In an attempt to handle this problem Parmar et al. in 2007 proposed an algorithm called the Min Min Roughness (MMR) algorithm which primarily deals with categorical data and uses a splitting attribute in the process. It was improved to propose an algorithm called the Min Mean Roughness (MMeR) algorithm by Kumar et al in 2009. The advantages of MMeR over MMR are that it can handle hybrid data and also a problem in determining the splitting attribute is resolved so that more accuracy could be obtained in classifying datasets. This was further extended to propose the Minimum of Standard deviation Roughness (SDR) algorithm by Tripathy et al in 2011 and Standard Deviation of Standard deviation Roughness (SSDR) algorithms. It has been observed that SDR outdoes MMeR and SSDR has almost the same capability as SDR. However the efficacy and permanence are seen as major points of concern when Purity ratio is measured. The ascending order of the accuracies is MMR, MMeR, SDR and SSDR [19] [17].

In this paper, we have tried to provide an algorithm which is an improvement over all the algorithms of the family mentioned in the above paragraph, called the Min-Mean-Mean-Roughness (MMeMeR). It has better accuracy than these algorithms as is evident from the experimental analysis and results over different datasets presented in the result and analysis section. The list of datasets considered for this purpose is tabulated below along with their descriptions. All these datasets are taken from The UCI repository and are standard datasets.

Table 1. Datasets Description

Data Set	Soybean	Zoo	Mushroom
Features	Multivariate	Multivariate	Multivariate
Value Type	Categorical	Categorical, Integer	Categorical
Operations	Classification	Classification	Classification
Number of Objects	47	101	8124
Number of Columns	35	17	22
Missing	No	No	Yes
Decision Classes	D1,D2,D3,D4	1-7	Poisonous, Edible

II. Related Work

A lot of work has been done in the field of data mining and data clustering. New methods have been proposed frequently as there is no fixed method for clustering data. Let us first present the history of clustering.

Expectation-Maximization (EM) algorithm was given by Dempster et al. [1]. For different classes, EM assigned differing probabilities. At last the solution was a locally optimal one. In 1982 Pawlak came out with the concept of rough sets in [12]. In this, the uncertainty of the clusters was taken care of. In 1989 the segmentation of radar signals while scanning land and marine objects was done using cluster analysis by Haimov et al. [4]. The indiscernibility of objects in a group was shown by Pawlak [13]. The concept of k-mode was introduced by Z. Huang in [5] which paved the way for a whole paradigm in itself. A spectral graph reduction algorithm for categorical data called STIRR was formulated by Gibson et al. [2]. This method is an iterative one which mapped non-linear dynamic systems to categorical data. ROCK hierarchical algorithm was postulated by Guha et al. [3]. The proximity between entities can be gauged by it. In [10] Mazlack gave explanations for coherent partitioning of the dataset to make cohesive clusters with less dissonance among them. Through [6] the natural set making of things in clustering is shown. Positron emission tomography (PET) method was given by Wong et al. [27]. In this nuclear medical imaging was used to segment the tissues. Clustering approaches which can be applied for gene expression data were given by Jiang et al. [7]. In [20] [21] a proper analysis of fuzzy K-modes was done. Kim et al [8] postulated the fuzzy centroid method which is an extension of fuzzy k-modes, where the hard-type centroid used in the fuzzy K-modes algorithm is modified. Mathieu et al. [9] identified the programs to participate in and determined the resource allocation by using cluster analysis. High scale research and development planning were a part of the decision enhancement module. A clustering algorithm specifically made to take care of the complexity of gene data was formulated by Wu et al. [28] in 2004. Parmar proposed in [11] the MMR or Min-Min-Roughness algorithm. In 2009 Tripathy et al [23] extended the MMR to MMeR algorithm and got a much higher clustering accuracy. MMeR was further extended to develop SDR and SSDR in [18] and [19] respectively which were based on standard deviation computation. All these algorithms along with the MADE procedure were compiled in a single paper by Tripathy et al in [17] and they were compared on their purity. The details of all the algorithms have been discussed in the forthcoming sections of the document.

III. Definition and Notation

The impreciseness in data was captured through the inception of rough sets by Pawlak in [12] [13]. The understanding of a human depends upon its capability to group different entities. This is the basic notion of the rough set theory. Many equivalence relations are defined over the rough sets. This is due to the fact that the classification [16] of the universe and equivalence relations are such notions which can be interchanged with each other. The lower approximation of a rough set comprises the entities contained in it and the upper approximation are the entities which may be available in it depending upon the information.

K= (U, R) is a knowledge base. The universal set is U and R is a group of equivalence relations. The indiscernibility relations is a family of relations postulated upon U and K.

U is also the universe of discourse and A is the group of features. Any attribute can be denoted by a $\in$ A. A set of values V_a is the domain of a.

V is the superset of all values and B is the subset of A which is nonempty.

Definition 1 (Indiscernibility relation): The indiscernibility relation is given as Ind(B). It is postulated as:

For every $a \in B$ if a(x) = a(y) then x Ind(B) y

Here the value of attribute ‘a’ for entity x is given by a(x).

Definition 2 (Equivalence class): All the objects that have the same value for a given attribute compose an equivalence class.

Definition 3 (Lower approximation): Lower approximation gives the subset of all the objects which satisfy a given equivalence class.

$$\underline{X_B} = \cup\{ x_i \mid [x_i]_{Ind(B)} \subseteq X\} \quad (1)$$

Definition 4 (upper approximation): The compound of all the groups which give an intersection with X which is not null is known as upper approximation.

$$\overline{X_B} = \bigcup\{ x_i \mid [x_i]_{Ind(B)} \cap X \neq \varphi\} \quad (2)$$

Definition 5 (Roughness): Roughness is defined as unity minus the ratio of the number of objects contained in the lower approximation divided by the upper approximation.

$$R_B(X) = 1 - \frac{|\underline{X_B}|}{|\overline{X_B}|} \quad (3)$$

X is crisp if $R_B(X) = 0$. This is with respect to B. If $R_B(X) < 1$ B is vague with respect to X.

Definition 6 (Relative roughness**)**: **T**he lower and upper approximation of X with respect to $\{a_j\}$ is given as $\underline{X_{a_j}(a_i = \alpha)}$ and $\overline{X_{a_j}(a_i = \alpha)}$

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

where $a_i, a_j \in A$ and $a_i \neq a_j$. (4)

$R_{a_j}(X)$ is the roughness with respect to $\{a_j\}$.

Definition 7 (Mean roughness): The mean roughness for the equivalence class $a_i = \alpha$ is given as

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

Definition 8 (Mean mean roughness): Here we define the mean of mean roughness defined in definition 7. Let $\beta, \gamma, \delta, \chi \ldots$ and so on be the values other than α of the attribute a_i. So, in this we take the mean of all roughness obtained for all these values with respect to the other attributes which is done as

$$MeMeR(a_i) = (MeR(a_i = \alpha) + MeR(a_i = \beta) + MeR(a_i = \gamma) + MeR(a_i = \delta) + MeR(a_i = \chi) + \ldots)/|V_{a_i}| \quad (6)$$

Definition 9 (Distance of relevance): DR for relevance [3] of things is:

$$DR(B,C) = \sum_{i=1}^{n} (b_i, c_i) \quad (7)$$

Here B and C are objects and b_i and c_i are their values respectively, under the i^{th} attribute a_i. In addition, 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 there is a numerical attribute; where 'eq_{B_i}' is the number assigned to the equivalence class that contains b_i. 'eq_{C_i}' is is the number assigned to the equivalence class that contains c_i and the number of equivalence classes in numerical attribute a_i is 'no_i'.

To compare the accuracies of different algorithms the approach is given below:

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

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

$$Over\ all\ Purity = \frac{\sum_{i=1}^{no.\ of\ clusters} Purity(i)}{no.\ of\ clusters} \quad (8)$$

IV. Proposed Algorithm

The whole process of MMeMeR has been discussed in this section.

1. Process MMeMeR(U, k)
2. Start
3. CNC = 1 (CNC is the current number of clusters).
4. U is the ParentNode.
5. Loop1:
6. If (CNC ≠1 and CNC < k)
7. ParentNode = Proc ParentNode (CNC)
8. End if
   // the ParentNode clustering begins
9. For every ai compute the equivalence classes.
10. Calculate $Rough_{a_j}(a_i)$ for each aj ∈ A (where j is not equal to i).
11. Apply (5) to get mean roughness.

$$MeMeR(a_i) = \frac{\{MeR(a_i = \alpha) + MeR(a_i = \beta) + MeR(a_i = \gamma) + + MeR(a_i = \delta) + \ldots\}}{|V_{a_i}|}$$

12. Set

$$MMeMeR(a_i) = Min\{MeMeR(a_i = \alpha) + MeMeR(a_i = \beta) + MeMeR(a_i = \gamma) + \ldots upto\ a_i\}$$

13. Let a_i be the splitting attribute

14. Perform binary split on a_i
15. This could be done by taking the equivalence class whose roughness value is nearer to the roughness of the splitting attribute a_i.
16. The number of leaf nodes is equal to CNC.
17. Go to Loop 1.
18. Stop
19. Process ParentNode (CNC)
20. Start
21. Let i = 1 to CNC-1
22. If the Avg-distance of cluster i is available
23. Goto label
24. Else
25. n = Count (Cluster i elements).
26. Avg Dist (i) = 2* (DR)/(n*(n-1)).
27. label :
28. i++
29. Loop
30. Find Max (Avg-distance (i))
31. Send back (Entities in cluster i) which have Max (Avg-distance (i))
32. Stop

V. Explaination and Experimental Part

Here, first we start by explaining most of the important parts of the algorithm. We elucidate them so as to make a clear understanding of why we have adopted those methods and why they will lead to logically good accuracy of clustering.

In the beginning we take the entire dataset and make the domain of all the equivalent classes according to the distinct values of the attribute. Then we calculate the roughness of that value of the attribute by using (5). In this step we use mean of all the roughness for the particular value of the attribute with respect to the other attributes. This has been done because on going for mean we are taking into account the overall power of that value to make crisp partitions in the dataset. This gives us the roughness achieved for all distinct values of all the attributes in the dataset.

In the next step we again take the mean of the roughness [20] [15] of all the values of the attribute present as done in (6) and do it for all the attributes. If instead we would have taken minimum of all roughness of all the distinct values then the selected value, let it be α, would have been able to make crisp divisions of the dataset. But then there may be some attributes for which the roughness with respect to α would have been on a higher side but due to the fact that we are taking overall minimum so that value of roughness is not showing in the result. And there may be some other value β of the same attribute that would be giving a low roughness with respect to other attributes. So taking mean in this step proves to be much better in most datasets.

After taking two means we take the minimum of the roughness of all the attributes and get the overall roughness of the entire dataset. In lieu of it if we would have gone for anything which needs computations and combinations like mean or standard deviation then the dissonance of the clusters formed would have been high. According to Mazlack [10] we have to increase the resonance of the partitions. So by taking minimum in this step we also ensure that the attribute selected is giving the lowest roughness. So, the split-ups formed from it will be very crisp.

Also, after selecting the splitting attribute we use the method defined in [23] for performing binary splitting on the attribute and not the method defined by Parmar in [11]. This is because in [11] the value which has less minimum roughness is simply taken to form a new cluster and the left out data objects are sent again to repeat the procedure. But in [23] the splitting is done and then the intra data object distance is calculated for the two factions formed. The group which has less average distance is retained and the other group is sent back for recursion. Also, when we take other recursions the group once separated out is also compared with fresh groups because it could also have greater average distance. So this procedure takes into account all scenarios as compared to Parmar's paper. Also it follows from [10] that we should try to reduce intra-item dissonance so as to achieve cohesion in the partitions formed. So the distance method is very much relevant here.

Moreover, we are considering that equivalent class in the attribute which is closest to the mean as one cluster and the left out data objects in the other cluster. This is because it follows from our second step where we take the mean of all the values as done in [6]. So, value closest to mean will give a lesser deviation and will prove to give more coherent splitting.

Further, when we perform the whole procedure we always get the two-valued attribute as the splitting attribute. This follows from [10] where a reduction heuristic was hypothesized which first took two-valued attributes only due to the fact that they create more balanced partitions. Also, if we consider a two-valued attribute a_i and take its roughness with respect to a higher-valued attribute then we see that it has more chances of getting a high lower approximation [17] and hence a higher crispness. So, in almost all datasets the splitting attribute selected is bi-valued.

Now, we give an example to explain the above algorithm. A fictitious mammal data set has been made:

Table 2. Mammal Set

MAMMAL NAME	HEIGHT	TYPE	SKIN	LIFE SPAN
M1	Short	Lion	Yellow	36
M2	Mid	Lion	Yellow	27
M3	Long	Panda	Grey	20
M4	Short	Panda	Grey	41
M5	Mid	Whale	Blue	39
M6	Long	Whale	Blue	16
M7	Long	Whale	Blue	18

Let the number of classes be 3. This means that k is 3. After that U is the ParentNode. First CNC = 1.

In the first time the number of clusters generated will be 1. Directly we will go for relative roughness with respect to the other attributes. a_i is 'HEIGHT' when i=1. Let X be the objects which have the same value given an attribute. This feature has three values 'Short', 'Mid' and 'Long'; so when we take α = 'Short' first we get X= {M1, M4} and taking j=2 we get a_j='Type'. So, the set of the equivalence classes of a_j is {(M1, M2), M3, M4, (M5, M6, M7)} and $\underline{X_{a_j}(a_i=\alpha)}$ = {ϕ} which is the lower approximation and $\overline{X_{a_j}(a_i=\alpha)}$ = {M1, M2, M4} is the upper approximation. The roughness of a_i (when a_i ='Height' and α='Short') is given by

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

After this when we change j (when j = 3, 4) and keeping the value of a_i (a_i = 'height') same and α (α = 'Short') we have to find the roughness of a_i relative to the attributes 'SKIN' (when j = 3) and 'LIFE SPAN' (when j = 4) and is given by

$$R_{X_j}(X/a_i=\alpha)=1-\frac{\left|X_{a_j}(a_i=\alpha)\right|}{\left|X_{a_j}(a_i=\alpha)\right|}=1-\frac{0}{5}=1$$

when j=3 and a_j='SKIN'

$$R_{X_j}(X/a_i=\alpha)=1-\frac{\left|X_{a_j}(a_i=\alpha)\right|}{\left|X_{a_j}(a_i=\alpha)\right|}=1-\frac{2}{2}=0$$

when j=4 and a_j='LIFE SPAN'

Now by taking the mean of the values we get 0.67. So this is the roughness when α='Short'. This is stored in a variable separately.

Similarly we do for α='Mid' and 'Long'. After all this we will get three mean values for each different α. Later we take the mean of different values of α and store this answer somewhere.

This is continued for each a_i (for a_i='TYPE', 'SKIN' and 'LIFE SPAN') and the obtained values will be stored. The obtained means are taken for calculation in the next step. For finding the roughness value of the whole data we take the minimum of all the attribute roughness values. Mostly the value of the roughness will not match with that of the mean values of the equivalence classes. So we take the one which is the closest and on the basis of it we will divide the dataset into two separate clusters.

After that we need to find the distance of relevance between the entities present in a given cluster. This is shown below. For example, let us take two tuples M4 and M6 which are as follows

Table 3. Sample Set

MAMMAL NAME	HEIGHT	TYPE	SKIN	LIFE SPAN
M4	Short	Panda	Grey	41
M6	Long	Whale	Blue	16

Here B=M4 and C=M6. The DR (B, C) is defined as

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

$$= \text{DR } (b_{height}, c_{neight}) + \text{DR } (b_{type}, c_{type}) + \text{DR } (b_{skin}, c_{skin}) + \text{DR } (b_{life\,span}, c_{life\,span})$$

So, $\text{DR } (b_{height}, c_{neight}) = 1$ as $b_{height} \neq c_{height}$

$\text{DR } (b_{type}, c_{type}) = 1$ as $b_{type} \neq c_{type}$

$\text{DR } (b_{skin}, c_{skin}) = 0$ as $b_{skin} = c_{skin}$.

As 'LIFE SPAN' is a numerical attribute for DR ($b_{life\ span}$, $c_{life\ span}$) we need to have some different method. To do this we have to find the mean of the no. of equivalence classes of categorical attributes. Hence the average of size of equivalence class is (3+4+2)/3 = 3. In this case we have got an integer value but sometimes we may get a fraction also. Then we have to round it off.

Later, sorting needs to be done for the attribute LIFE SPAN. As a result we get {16, 18, 20, 27, 36, 39, 41}. Three sets have been formed below which are the distributions.

Set 1 = {16, 18}

Set 2 = {20, 27}

Set 3 = {36, 39, 41}

Next we calculate DR ($b_{life\ span,}$ $c_{life\ span}$). In our case $b_{life\ span}$ = 41 and $c_{life\ span}$ = 16. Hence, now 41 is substituted with 3 and 16 with 1.

So,

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

Finally, $\text{DR (B,C)} = \text{DR } (b_{height}, c_{neight}) + \text{DR } (b_{type}, c_{type}) + \text{DR } (b_{skin}, c_{skin}) + \text{DR } (b_{life\,span}, c_{life\,span})$

= 1+1+0+0.67

= 2.67

So, this is how the distance between objects present in cluster 1 and cluster 2 is calculated. We send back the one having more mean distance for iteration.

Hence unless we get the right number of clusters, which is three here, we have to apply this algorithm.

VI. Result and Analysis

We have implemented the algorithms and tested it by taking the three datasets mentioned in section 3. The results obtained are summarized in this section. The measure of the efficacy of the algorithms is the concept of 'purity ratio'. SSDR and SDR gave the same purity for all datasets. So SSDR has not been considered for comparison.

A. Experiment 1 (Soybean Data Set)

The dataset contains 47 entities. 35 attributes are there. 4 decision classes in the form of diseases for the soybean plant are present. After 4 clusters have been made, we will stop this program. The disease classes are D1, D2, D3, and D4. Purity and cluster number are the other two attributes.

Table 4. Analysis of Soybean Dataset

Cluster Number	D1	D2	D3	D4	Purity
1	0	10	0	0	1
2	10	0	0	0	1
3	0	0	5	17	0.77
4	0	0	5	0	1
Overall Purity					**0.9425**

$$\text{So, } Over\ all\ Purity = \frac{\sum_{i=1}^{no.of\ clusters} Purity(i)}{no.of\ clusters} = 0.9425$$

0.9425 or 94.25% is the purity. This data set was used by other researchers like Kim et al and Tripathy et al. So we can compare our findings with theirs as follows:

Table 5. Comparison of Soybean Dataset with other algorithms

Kmodes	Fuzzy Kmodes	Fuzzy centroids	MMR	MMeR	SDR	MMeMeR
0.69	0.77	0.97	0.83	0.83	0.93	0.9425

'Fuzzy centroid' has higher purity ratio than all other algorithms and MMeMeR algorithm comes slightly behind it. This difference is very low in comparison to the difference of purity ratio between the two algorithms of MMeMeR and 'Fuzzy Centroid' in experiment 2 below. But we can see that MMeMeR is much better clearly than SDR [18], MMeR and MMR which are based on rough set theory. Also we have to consider that the soybean data set is very small when compared with the data sets used below, where MMeMeR has clearly gone past the other methods in terms of the purity ratio.

B. Experiment 2 (Zoo Data Set)

101 entities are there in the zoo database. 18 categorical attributes are present. 7 decision classes are present. Hence seven clusters have to be formed. Table 6 gives a summary of the clusters in Zoo data set when MMeMeR is applied.

Table 6. Analysis of Zoo Dataset

Cluster Number	C1	C2	C3	C4	C5	C6	C7	Purity
1	0	20	0	0	0	0	0	1
2	2	0	0	0	0	0	0	1
3	39	0	3	0	4	0	0	0.829
4	0	0	1	13	0	0	0	0.93
5	0	0	0	0	0	0	1	1
6	0	0	0	0	0	7	9	0.563
7	0	0	0	0	0	1	0	1
Overall Purity								**0.902**

$$Over\ all\ Purity = \frac{\sum_{i=1}^{no.of\ clusters} Purity(i)}{no.of\ clusters} = 0.902$$

90.2% is the purity given on applying it. Tripathy et al. made a comparison of their method with MMR[11] and algorithms which take into account the fuzziness. The fuzzy centroid technique by Kim et al. also implemented their technique with the help of this data set and made a comparison with k-modes [5] and fuzzy k-modes[21] [22]. So in Table 7 all the comparisons have been provided -

Table 7. Comparison of Zoo Dataset with other algorithms

Kmodes	Fuzzy Kmodes	Fuzzy centroids	MMR	MMeR	SDR	MMeMeR
0.60	0.64	0.75	0.787	0.902	0.907	0.902

From Table 7 we can see that SDR is very slightly better than MMeMeR. We test the mushroom dataset next.

C. Experiment 3 (Mushroom Data Set)

Mushroom data set is a very large data set and our algorithm MMeMeR has been applied to it. 8124 objects are there and 22 categorical attributes are present in it. There are two classes in which the data points are classified. In MMeMeR algorithm also we have taken the stopping criterion as 20 clusters. This helps us in comparing the purity ratio of our algorithm with these algorithms. So, our algorithm also forms 20 clusters. The clusters formed are given below -

Table 8. Analysis of Mushroom Dataset

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	3	0.971
10	96	96	0.5
11	192	0	1
12	0	73	1
13	1024	0	1
14	72	0	1
15	512	0	1
16	0	256	1
17	20	1636	0.988
18	28	0	1
19	192	0	1
20	0	24	1
Overall Purity			**0.973**

$$Over\ all\ Purity = \frac{\sum_{i=1}^{no.\ of\ clusters} Purity(i)}{no.\ of\ clusters} = 0.973$$

The purity ratio comes out to be almost 1. We compare its purity ratio with those of others like MMR, MMeR and SDR algorithms. The comparison is as follows -

Table 9. Comparison of Mushroom Dataset

Data Set	MMR	MMeR	SDR	MMeMeR
Mushroom	0.84	0.9641	0.9723	0.973

MMeMeR has the greatest accuracy in comparison to MMR, MMeR and SDR in this case. So on this data set also MMeMeR shows its superiority and efficacy against the likes of MMR, MMeR and SDR methods.

From the experimental analysis performed above, we can say that MMeMeR is one of the best clustering algorithms present in the domain of hybrid data clustering. Though MMeMeR slightly lags behind fuzzy centroid only for the soybean dataset and SDR is a little better than it when zoo dataset is considered, we come to know that all-round MMeMeR is better than all the other processes handling categorical or numerical data.

Let us explain a few points when we compare the SDR and MMeMeR algorithms. It is known that SDR [18] will only express deviation from mean roughness for a particular attribute or equivalence class. But, if roughness in itself is large then it is likely to generate unbalanced partitions and so the purity will be less. But MMeMeR takes mean value directly and so the value with higher roughness will not get selected. Only mediocre values will be selected even if there are some equivalence classes with high roughness. And if all the roughness values [24] [25] [26] are high only then it will directly indicate that the overall roughness of the dataset is high, which is not the case with SDR. So MMeMeR is better than SDR as well.

VII. Conclusion

In the real world databases categorical data have become very obligatory. But only a few good techniques are there to cluster these datasets. So keeping this in mind we formulate a notion called MMeMeR, which is more efficient than most of the earlier algorithms which have been made in this direction. The uncertainty in data is handled using rough set theory. Firstly, a process has been laid out which can be used to simultaneously cluster categorical and numerical attributes and the distance of relevance method is also given by us which gives better results as compared to MMR, where the number of objects are only seen for clustering. Also the accuracy is better than MMeR. Also we have made a logical and coherent analysis of why taking mean or minimum at every step will give better accuracy. So while SDR will give better results in a few datasets where the spread

factor comes in handy while making partitions but MMeMeR algorithm proves more useful for most as it considers the whole cumulative distribution and the average for making crisp divisions. Also the knowledge provided by it proves very useful because the objects at the edge of a dataset are more captivating than those which can be clustered with certainty. Hence, the Min-Mean-Mean-Roughness has proven to be an important enrichment to clustering approaches, particularly in the direction of soft computing methods.

VIII. Scope for Future Work

Good clusters can be formed if the distance of relevance formula is chosen more appropriately. Initial cluster makes a lot of difference for the whole process to follow. Thus a method to select the initial cluster will pave a great path for the entire procedure ahead. Fuzzy cognitions can be introduced to get a good splitting attribute. Rough-fuzzy concepts will then be formed. The notion discussed here could be applied for the detection of any outliers as well. Also, one can try to establish a true relationship among various arguments or parameters in the proposed algorithm.

References

[1] Dempster, A., Laird, N. and Rubin, D., "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, Vol.39 (1), (1977), pp. 1–38. J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

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

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

[4] Haimov, S., Michalev,M. and Savchenko, A. and Yordanov, O., "Classification of radar signatures by autoregressive model fitting and cluster analysis", IEEE Transactions on Geo Science and Remote Sensing Vol.8 (1), (1989), pp. 606–610.

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

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

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

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

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

[10] Mazlack, L.J., He, A. and Zhu, Y., "A rough set approach in choosing partitioning attributes", Proceedings of the ISCA 13th International Conference (CAINE-2000), (2000).

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

[12] Pawlak, Z., "Rough Sets", Int. Jour of Computer and information Sciences, Vol.11, (1982), pp.341- 356.

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

[14] Sharmila, B.K. and Tripathy, B.K., "Clustering Mixed Data using Neighborhood Rough Sets", International Journal of Advanced Intelligence Paradigms, September (2016).

[15] Sharmila, B.K. and Tripathy, B.K., "Exploring incidence-prevalence patterns in spatial epidemiology via neighborhood rough sets", International Journal of Healthcare Information Systems and Informatics, Vol. 12(1), (2017), pp. 30-43.

[16] Swarnalatha, P. and Tripathy, B.K., "A Centroid Model for the Depth Assessment of Images using Rough Fuzzy Set Techniques", IJISA, vol.4 (3), (2012), pp.20-26.

[17] Tripathy, B.K. and Ghosh, A., "Data Clustering Algorithms Using Rough Sets", Handbook of Research on Computational Intelligence for Engineering, Science, and Business, (2012), p.297.

[18] Tripathy, B.K. and Ghosh, A., "SDR: An algorithm for clustering categorical data using rough set theory", Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE, Trivandrum, (2011), pp. 867-872.

[19] Tripathy, B.K. and Ghosh, A., "SSDR: An Algorithm for Clustering Categorical Data Using Rough Set Theory", Advances in Applied Science Research, Vol.2 (3), (2011), pp. 314-326.

[20] Tripathy, B.K., Goyal, A. and Patra, A.S., " A Comparative Analysis of Rough Intuitionistic Fuzzy K-mode for Clustering Categorical Data", Research Journal of Pharmaceutical, Biological and Chemical Sciences, Vol. 7(5), (2016), pp. 2787-2802.

[21] Tripathy, B.K., Goyal, A. and Patra, A.S., "Clustering Categorical Data Using Intuitionistic Fuzzy K-mode", International Journal of Pharmacy and Technology, Vol. 8 (3), September (2016), pp. 16688-16701.

[22] Tripathy, B.K., Khandelwal, S., and Satapathy, M.K., "A Bag Theoretic Approach towards the Count of an Intuitionistic Fuzzy Set", IJISA, vol.7 (5), (2015), pp.16-23.

[23] Tripathy, B.K. and Kumar, M S, "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, Issue No.2, (2009), pp.189-207.

[24] Tripathy, B.K. and Nagaraju, M., "On Some Topological Properties of Pessimistic Multigranular Rough Sets", IJISA, vol.4 (8), 2012, pp.10-17.

[25] Tripathy, B.K. and Parida, S.C., "Covering Based Optimistic Multigranular Approximate Rough Equalities and their Properties", International Journal of Intelligent Systems and Applications (IJISA), Vol.8 (6), (2016), pp.70-79.

[26] Tripathy, B. K., Rawat, R., Vani, D., and Parida, S.C., "Approximate Reasoning through Multigranular Approximate Rough Equalities", IJISA, vol.6 (8), (2014), pp.69-76.

[27] Wong, K., Feng, D. and Meikle, S. and Fulham, M., "Segmentation of dynamic pet images using cluster

analysis", IEEE Transactions on Nuclear Science Vol.49 (1), (2002), pp. 200–207.

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

Authors' Profiles

B.K. Tripathy has received 03 gold medals for topping the list of candidates at graduation and post-graduation level of Berhampur University. He was a Professor and Head of the department of Computer Science of Berhampur University till 2007. Dr. Tripathy is now working as a Senior Professor in School of Computing Science and Engineering, VIT University, Vellore, India. He has received research/academic fellowships from UGC, DST, SERC and DOE of Govt. of India for various academic pursuits. Dr. Tripathy has published more than 400 technical papers in different international journals, proceedings of reputed international conferences and edited research volumes. He has produced 26 PhDs, 13 MPhils and 4 M.S (By research) under his supervision. Dr. Tripathy has published two text books on Soft Computing and Computer Graphics. He was selected as honorary member of the American Mathematical Society from 1992-1994 for his distinguished contribution as a reviewer of American Mathematical Review. Dr. Tripathy has served as the member of Advisory board or Technical Programme Committee member of several International conferences inside India and abroad. Also, he has edited two research volumes for IGI publications and is editing three more research volumes. He is a life/senior member of IEEE, ACM, IRSS, CSI, ACEEE, OMS and IMS. Dr. Tripathy is an editorial board member/reviewer of more than 60 journals. He has guest edited some research journals. Dr. Tripathy has Technical grants for research projects from various funding agencies like UGC, DST and DRDO. His research interest includes Fuzzy Sets and Systems, Rough Sets and Knowledge Engineering, Data Clustering, Social Network Analysis, Soft Computing, Granular Computing, Content Based Learning, Neighbourhood Systems, Soft Set Theory, Social Internet of Things, Big Data Analytics, Theory of Multisets and List theory.

Akarsh Goyal, is a student, pursuing a Bachelors of Technology in Computer Science and Engineering at VIT University, Vellore, Tamil Nadu, India. He is an avid reader and computer enthusiast. He has published a few scientific papers in area of IoT, data mining, software engineering and marketing. His areas of interest are data mining, machine learning, intelligent systems, IoT, and android based applications. Akarsh has a passion for design which is reflected by his participations in hackathons. He has a particular aptitude to projects with real life application and scope using computer applications.

Rahul Chowdhury is a student, pursuing a Bachelors of Technology in Computer Science and Engineering at VIT University, Vellore, Tamil Nadu, India. He has written papers in text mining, rough sets and big data. Competitive Coding amazes him. He likes to work on real-time systems and machine learning based design models. He has been a part of various hackathons and working on application designs is his forte.

Patra Anupam Sourav, is a student, School Of Computer Science and Engineering, VIT University, Vellore. He is a regular competitive programmer and has experience in algorithm design and solving complex problems. His areas of interest are data mining, machine learning and big data analytic.

How to cite this paper: B.K. Tripathy, Akarsh Goyal, Rahul Chowdhury, Patra Anupam Sourav,"MMeMeR: An Algorithm for Clustering Heterogeneous Data using Rough Set Theory", International Journal of Intelligent Systems and Applications(IJISA), Vol.9, No.8, pp.25-33, 2017. DOI: 10.5815/ijisa.2017.08.03

