MMeNR: Neighborhood Rough Set Theory Based Algorithm for Clustering Heterogeneous Data

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

Akarsh Goyal
School of Computer Science and Engineering
VIT University
Vellore-632014, Tamil Nadu, India
akarsh.goyal15@gmail.com

Rahul Chowdhury
School of Computer Science and Engineering
VIT University
Vellore-632014, Tamil Nadu, India
chowdhuryrahul5@gmail.com

Sharmila Banu K.
School of Computer Science and Engineering
VIT University
Vellore-632014, Tamil Nadu, India
sharmilabanu.k@vit.ac.in

***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 [1] algorithm was developed which was based on basic rough set theory. MMeR [2] was proposed in 2009 which surpassed the results of MMR in taking care of categorical data. It has the capability of handling heterogeneous data but only to a limited extent because it is based on classical rough set model. In this paper, we generalize the MMeR algorithm with neighbourhood relations and make it a neighbourhood rough set model which we call MMeNR (Min Mean Neighborhood Roughness). It takes care of the heterogeneous data and also the uncertainty associated with it. Standard data sets have been used to gauge its effectiveness over the other methods.**

Keywords- Neighborhood; MMeNR; Clustering; Uncertainty; MMR; MMeR;

I. INTRODUCTION

Making sets of data which are analogous to each other is the prime notion of clustering. Also those which are not similar are grouped separately. The technique is used for many data analysis chores. Clustering is used to make small subsets which can be easily managed, analysed and taken care of by segmenting large hybrid data sets [3]. Groupings which come naturally to the objects are found out using clustering [4]. Many areas make use of clustering techniques. For instance, gene data complexity handling method was made by Wu et al. [5] using clustering. Clustering techniques which can be used for the analysis of gene expression data were developed by Jiang et al. [6]. Positron emission tomography (PET) method was given by Wong et al. [7]. In this nuclear medical imaging was used to segment the tissues. In 1989 the segmentation of radar signals while scanning land and marine objects was done using cluster analysis by Haimov et al. [8]. 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.

These techniques mostly handle only numerical databases. Hence these cannot be used for data sets which have domains that are categorical. Earlier works in the field of clustering used to develop algorithms which could only take care of numerical data as it was very easy to formulate similarity functions between them. However when it comes to categorical data it becomes difficult as they have features which are multi-valued. The correspondence is in the form of values which are same in a given attribute and also objects which are similar. Because of this we have to see both in the rows as well as the columns for the similarity [5].

Dempster et al. [12], Huang [3], Guha et al. [11], and Gibson et al. [10] gave methods which take care of categorical features. But they are not capable enough to

978-1-5090-5297-4/17/$31.00 ©2017 IEEE

handle uncertainty. Due to this these algorithms have stability issues which renders them ineffective for real world databases in which uncertainty is very common.

Hence there is an increasing need of a technique which could process both the categorical and numerical features and also have stability. Parmar proposed in [1] the MMR or Min-Min-Roughness algorithm. In 2009 Tripathy et al [2] [17] [22] extended the MMR to MMeR algorithm and got a much higher clustering accuracy. Both the algorithms could handle uncertainty on the categorical data sets using Pawlak basic rough set theory [13] [14]. In comparison to MMR the MMeR algorithm had higher accuracy in doing the same process. Also MMeR [18] [19] introduced a clustering notion to handle numerical attributes as well. But then also these techniques cannot be used for robust clustering of datasets which are heterogeneous. Only nominal attributes can be handed by Pawlak's rough set model. These two issues arise due to the fact that the equivalence relations used by the model divides the universe of discourse into elemental concepts which can be used to process only a limited range of attributes.

These problems can be mitigated if we apply neighbourhood concept to hybrid data sets. Decision classes can be approximated with the help of neighbourhood granules generated by neighborhood relations [15] [16]. Min-Mean Neighborhood Roughness (MMeNR) is the concept we theorize in this paper based on the above properties of the neighborhoods. To compare the accuracy of different procedures we will use purity ratio. UCI repository data sets like the teacher assistant evaluation data and acute inflammations data have been used to validate our findings.

II. Definitions And Notations

The addendum of Pawlak's rough set model is the neighborhood rough set model. Hu et al [15] developed it. It reduces the number of attributes by finding the reducts. Also another function it possesses is that it is used for selecting subsets of features which are hybrid in nature. The main issues which the rough set model has are that of making subsets which have very similar objects and also to approximate these objects. This subset making is also known as granulation and this process leads to elemental granules. Equivalence classes which are mutually exclusive to each other is generated by the basic rough set model proposed by Pawlak. This can only be applied to a limited range of attributes which are mostly nominal in nature. Neighborhoods [23] [24] can combat this fall and play an important character when spaces are numerical. Decision classes can be approximated by neighborhood granules. These granules are generated by the application of neighborhood relations on the universe. This also leads to the formulation of a uniform framework.

$IS = \langle U, A \rangle$ is the information system which is used for supervised learning of the data set. Here U is the universe. It contains {x_1, x_2,..., x_n}, and is not null. {a_1, a_2,..., a_m} is the group of features which compose A. The decision characteristic is D and C is the conditional characteristic. A is also given as $A = C \cup D$.

Definition 1: Let x_i be an object which belongs to U. And B be a subset of C. Let distance function be Δ.

$\delta_B(x_i) = \{x_j \mid x_j \in U, \Delta^B(x_i, x_j) \le \delta\}$, is the neighborhood then.

Definition 2: If all x_1, x_2 and x_3 belongs to U. then -

(1) Distance between x_1, and x_2 is greater than 0 usually.
(2) Distance between them is null if $x_1 = x_2$.
(3) Distance between x_1 and x_2 is equal to distance between x_2 and x_1.
(4) Distance from x_1 to x_2 added with distance from x_3 to x_2 is greater than distance from x_1 to x_3.

Definition 3: B_1 is a numerical attribute given by $B_1 \subseteq A$ and B_2 is a categorical attribute [20] [21] given by $B_2 \subseteq A$.The neighborhood granule of sample x induced by B_1, B_2 and $B_1 \cup B_2$ are defined as

(1) For numerical attributes -

$\delta_{B_1}(x) = \{x_i \mid \Delta_{B_1}(x, x_i) \le \delta, x_i \in U\};$

(2) For categorical attributes -

$\delta_{B_2}(x) = \{x_i \mid \Delta_{B_2}(x, x_i) = 0, x_i \in U\};$

(3) For hybrid attributes -

$\delta_{B_1 \cup B_2}(x) = \{x_i \mid \Delta_{B_1}(x, x_i) \le \delta \wedge \Delta_{B_2}(x, x_i) = 0, x_i \in U\},$

where $\wedge$ means "and" operator.

Therefore Definition 3 is applicable to numerical, categorical data and their mixture.

Definition 4: The neighborhood approximation space is given as $\langle U, N \rangle$. X is a subset of U. So -

$$\underline{N}X = \{x_i \mid \delta(x_i) \subseteq X, x_i \in U\},$$

$$\overline{N}X = \{x_i \mid \delta(x_i) \cap X \ne \varepsilon, x_i \in U\}.$$

These two are the lower and upper approximations respectively. $\underline{N}X \subseteq X \subseteq \overline{N}X$.

$BNX = \overline{N}X - \underline{N}X.$, is the region which contains the boundary in this space.

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

$$NR_B(X) = 1 - \frac{\underline{NX}}{\overline{NX}}$$

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

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

$$NR_{a_j}(X/a_i=\alpha)=1-\frac{\left|\underline{NX}_{a_j}(a_i=\alpha)\right|}{\left|\overline{NX}_{a_j}(a_i=\alpha)\right|},$$ where a_i, a_j ∈A
and $a_i \neq a_j$.

This is the neighborhood roughness [25] of a_i in reference to a_j.

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

$$MeNR(a_i=\alpha)=(\sum_{j=1,j\neq i}^{n} NR_{a_j}(X/a_i=\alpha))/(n-1).$$

Definition 8 (Min mean neighborhood roughness): Here we define the minimum of mean neighborhood roughness defined in definition 7. Let β,γ,δ,χ . . . 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

MMeNR(a_i) = Min(MeNR(a_i = α), MeNR(a_i = β), MeNR(a_i= γ), MeNR(a_i= δ), MeNR(a_i = χ), . . .).

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

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

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) = $\frac{|\mathrm{eq}_{B_i}-\mathrm{eq}_{C_i}|}{no_i}$, 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 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) =
$$\frac{\text{The number of data occuring in both the i}^{th}\text{ cluster and its corresponding class}}{\textit{the number of data in the data set}}$$

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

III. EXAMPLE

Let us understand the concept of neighbourhood with the help of an example. Below a sample table is given –

TABLE I. Sample Set

Object	A	B
X_1	1	0.1
X_2	2	0.20
X_3	2	0.45
X_4	3	0.5
X_5	3	0.4

In this table, 5 objects from X_1 to X_5 are given. Two attributes are there to describe it. The first one is a categorical attribute A and the second one is a numerical attribute B. We can find out the neighborhoods of the objects with respect to both the attributes.

The neighborhood when we take attribute A is given as –

$\delta(X_1)$ = {X_1}, $\delta(X_2)$ ={X_2, X_3}, $\delta(X_3)$ ={X_2, X_3}, $\delta(X_4)$={X_4, X_5}, $\delta(X_5)$={X_4, X_5}

Let the threshold δ = 0.1. The neighborhood when we take attribute B is given as –

$\delta(X_1)$= {X_1, X_2}, $\delta(X_2)$ ={X_1, X_2}, $\delta(X_3)$={ X_3, X_4, X_5}, $\delta(X_4)$={ X_3, X_4, X_5}, $\delta(X_5)$={ X_3, X_4, X_5}

Let us calculate the lower and upper approximations of the given attribute values with respect to the other attribute values. Here, each object is a value and the neighborhood generated by it is an equivalence class.

So when we calculate the approximations for attribute A with respect to B we get –

$\underline{NX}_1=\Phi$, $\overline{NX}_1$= {X_1, X_2}

$\underline{NX}_2=\Phi$, $\overline{NX}_2$= {X_1, X_2, X_3, X_4, X_5}

$\underline{NX}_3=\Phi$, $\overline{NX}_3$= {X_1, X_2, X_3, X_4, X_5}

$\underline{NX}_4=\Phi$, $\overline{NX}_4$= {X_3, X_4, X_5}

$\underline{NX}_5=\Phi$, $\overline{NX}_5$= {X_3, X_4, X_5}

The lower and upper approximation for attribute B with respect to A is –

$\underline{NX}_1$= {X_1}, $\overline{NX}_1$ = {X_1, X_2, X_3}

$\underline{NX}_2$= {X_1}, $\overline{NX}_2$= {X_1, X_2, X_3}

$\underline{NX}_3$= {X_4, X_5}, $\overline{NX}_3$= {X_2, X_3, X_4, X_5}

$\underline{NX}_4$= {X_4, X_5}, $\overline{NX}_4$= {X_2, X_3, X_4, X_5}

$\underline{NX}_5$= {X_4, X_5}, $\overline{NX}_5$= {X_2, X_3, X_4, X_5}

Hence we have correctly explained the neighborhood set and now we move on to the entire algorithm in the next section.

IV. Proposed Algorithm

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

```
1.  Process MMeNR(U, k)
2.  Start
3.  CNC = 1 (CNC is the current number of clusters).
4.  U is the ParentNode.
5.  The first loop is Loop1:
6.  Check whether CNC $\neq$1 and CNC < k
7.  If so then :
8.  ParentNode = Proc ParentNode (CNC)
9.  End if
// the ParentNode clustering begins
10. For every $a_i$ calculate the equivalence classes.
11. Determine neighborhood using Definition 3.
12. Find Neighborhood Roughness using Definition 4
    and 6.
13. Next
14. Calculate the MMeNR given by definition 8.
15. Take the minimum of all MMeNR given by
    different attributes from $a_i$, $a_j$, …..
16. Next the splitting attribute is determined. Let it be
    $a_i$.
17. On $a_i$ binary split is performed.
18. This could be done by taking the equivalence
    class whose roughness value is nearer to the
    roughness of the splitting attribute $a_i$.
19. The number of leaf nodes is equal to CNC.
20. Go back to Loop 1.
21. Stop
22. ParentNode (CNC) // Procedure
23. Start
24. Let i = 1
25. Till i < CNC the process is done
26. If the Avg-distance of cluster i is already
    calculated
27. Goto label
28. Else if not done
29. n = Count (Cluster i elements).
30. Avg Dist (i) = 2* (DR)/(n*(n-1)).
31. label :
32. i++
33. Loop
34. Find Max (Avg-distance (i))
35. Send back (Entities in cluster i) which have Max
    (Avg-distance (i))
36. Stop
```

V. Empirical Evaluation

PYTHON has been used to write the codes and implement the whole technique. Purity ratio is used to determine that which method is better. 'Teacher assistant evaluation' and 'Acute inflammation' data sets have been used by us to make an informed comparison and come to a proper conclusion.

A. Experiment 1 (Teacher Assistant Evaluation Data Set)

151 objects are there in this set. Also 4 categorical features and 1 numerical feature is there to describe the objects. The classification is in terms of 3 classes. The clusters formed by applying MMeNR and MMeR [17] [22] are shown in TABLE I and TABLE II respectively. The decision classes are C1, C2 and C3.

TABLE II. Purity Ratio of Teacher Assistant Evaluation Data Set using MMeNR

Cluster Number	C1	C2	C3	Purity
1	2	6	15	0.652
2	0	0	2	1
3	47	44	35	0.373
Overall Purity				0.675

TABLE III. Purity Ratio of Teacher Assistant Evaluation Data Set using MMeR

Cluster Number	C1	C2	C3	Purity
1	2	6	15	0.652
2	3	5	12	0.6
3	44	39	25	0.41
Overall Purity				0.554

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

0.554 and 0.675 are the overall purities given by MMeR and MMeNR respectively.

We can see that MMeNR has higher accuracy as shown by Table 1 and Table 2.

B. Experiment 2 (Acute Inflammations Data Set)

The number of entities in this set are 120. 5 categorical characteristics and 1 numerical feature is there to describe each entity. The decision attribute has two values either 'yes' or 'no'. Only two clusters will be formed in this algorithm. C1 and C2 are the decision classes in the Tables III and IV.

TABLE IV. Purity Ratio of Acute Inflammation Data Set using MMeNR

Cluster Number	**C1**	**C2**	**Purity**
1	70	21	0.77
2	0	29	1
Overall Purity			0.885

TABLE V. Purity Ratio of Acute Inflammation Data Set using MMeR

Cluster Number	**C1**	**C2**	**Purity**
1	49	10	0.831
2	10	51	0.836
Overall Purity			0.8335

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

0.8335 and 0.885 are the overall purities given by MMeR and MMeNR respectively when used on acute inflammations data set. The purity ratio of MMeNR is higher than MMeR in this case as well.

In the domain of heterogeneous clustering we can thus conclude that MMeNR has better clustering efficacy and accuracy as compared to the other algorithms. This can be seen from the two examples given above. This is due to its intrinsic ability of dividing the universe of objects into elemental concepts. This helps to process both the numerical as well as categorical attributes simultaneously in a data set.

VI. CONCLUSION

In the real world databases heterogeneous 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 MMeNR, which is more efficient than most of the earlier algorithms which have been made in this direction. The uncertainty and heterogeneity in data is handled using neighborhood 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. The accuracy is higher than that of MMeR. Also we have made a logical and coherent analysis of how taking the neighborhood granules in the numerical spaces can help with clustering of hybrid data. We can also extend it in the future by using techniques which incorporate fuzziness of the data as well and make the algorithm more stable.

REFERENCES

[1] D. Parmar, T. Wu, and J. Blackhurst, “MMR: An algorithm for clustering categorical data using Rough Set Theory,” Data & Knowledge Engineering, vol. 63, pp. 879 – 893, 2007.

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

[3] Z. Huang, “Extensions to the k-means algorithm for clustering large data sets with categorical values,” Data Mining and Knowledge Discovery, vol. 2, no. 3, pp. 283–304, 1998.

[4] R. Johnson and W. Wichern, “Applied Multivariate Statistical Analysis,” Prentice Hall, New York, 2002.

[5] S. Wu, A. Liew, H. Yan, and M. Yang, “Cluster analysis of gene expression data based on self-splitting and merging competitive learning,” IEEE Transactions on Information Technology in BioMedicine, vol. 8, no. 1, pp. 5–15, 2004.

[6] D. Jiang, C. Tang, and A. Zhang, “Cluster analysis for gene expression data: a survey”, IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 11, pp. 1370–1386, 2004.

[7] K. Wong, D. Feng, S. Meikle, and M. Fulham, “Segmentation of dynamic pet images using cluster analysis,” IEEE Transactions on Nuclear Science, vol. 49, no. 1, pp. 200–207,2002.

[8] S. Haimov, M. Michalev, A. Savchenko, and O. Yordanov, “Classification of radar signatures by autoregressive model fitting and cluster analysis,” IEEE Transactions on Geo Science and Remote Sensing, vol. 8, no. 1, pp. 606–610,1989.

[9] R. Mathieu and J. Gibson, “A Methodology for large scale R&D planning based on cluster analysis,” IEEE Transactions on Engineering Management, vol. 40, no. 3, pp. 283–292, 2004.

[10] D. Gibson, J. Kleinberg, and P. Raghavan, “Clustering categorical data: an approach based on dynamical systems,” The Very Large Data Bases Journal, vol. 8, no. 3-4, pp. 222–236, 2000.

[11] S. Guha, R. Rastogi, and K. Shim, “ROCK: a robust clustering algorithm for categorical attributes,” Information Systems, vol. 25, no. 5, pp. 345–366, 2000.

[12] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” Journal of the Royal Statistical Society, vol. 39, no. 1, pp. 1–38, 1977.

[13] Z. Pawlak, “Rough Sets,” Int. Jour of Computer and information Sciences, vol.11, pp.341- 356, 1982.

[14] Z. Pawlak, “Rough Sets- Theoretical Aspects of Reasoning About Data” Norwell: Kluwar Academic Publishers, 1992.

[15] Q. Hu, D. Yu, J. Liu, and C. Wu, “ Neighborhood rough set based heterogeneous feature subset selection”, Information Sciences, vol. 178, no. 18, pp. 3577-3594, 2008.

[16] S.U. Kumar and H.H. Inbarani, “A Novel Neighborhood Rough Set Based Classification Approach for Medical Diagnosis,” Procedia Computer Science, vol. 47, pp. 351-359, 2015.

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

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

[19] B.K. Tripathy and A. Ghosh, “SSDR: An Algorithm for Clustering Categorical Data Using Rough Set Theory”, Advances in Applied Science Research, Vol.2, no. 3, pp. 314-326, 2011.

[20] B.K. Tripathy, A. Goyal, and A.S. Patra, “Clustering Categorical Data Using Intuitionistic Fuzzy K-mode”, International Journal of Pharmacy and Technology, Vol. 8, no. 3, pp. 16688-16701, September - 2016.

[21] B.K. Tripathy, A. Goyal, and A.S. Patra, “ A Comparative Analysis of Rough Intuitionistic Fuzzy K-mode for Clustering Categorical Data”, Research Journal of Pharmaceutical, Biological and Chemical Sciences, Vol. 7, no. 5, pp. 2787-2802, 2016.

[22] B.K. Tripathy, A. Goyal, R. Chowdhury, and A.S. Patra, “MMeMeR: An Algorithm for Clustering Heterogeneous Data using Rough Set Theory”, Communicated to International Journal of Intelligent Systems and Applications, 2017.

[23] B.K. Sharmila and B.K. Tripathy, “Exploring incidence-prevalence patterns in spatial epidemiology via neighborhood rough sets”, International Journal of Healthcare Information Systems and Informatics, Vol. 12, no. 1, pp. 30-43, 2017.

[24] B.K. Sharmila and B.K. Tripathy, “Neighborhood Rough Sets Based Spatial Data Analytics”, Encyclopedia of Information Science and Technology, Fourth Edition, IGI Global, 2018.

[25] B.K. Sharmila and B.K. Tripathy, “Clustering Mixed Data using Neighborhood Rough Sets”, International Journal of Advanced Intelligence Paradigms, September - 2016.