A Neighbourhood Rough Set Based Clustering Algorithm and its Applications

B. K. Tripathy [1], Akarsh Goyal [2]

[1] School of information technology and Engineering, VIT University, Vellore, India

[2] University of Southern California, USA

[1]tripathybk@vit.ac.in, [2]akarsh.goyal15@gmail.com

Abstract:

The process of Data clustering puts similar objects into the same group. The attributes under consideration may be numerical or categorical. A type of partition attribute based clustering algorithm dealing with categorical attributes only through rough sets was started in 2007 with the Min-Min Roughness (MMR) algorithm due to Parmar et al. It was generalised to the Min Mean roughness (MMeR) algorithm dealing with heterogeneous attributes by Kumar et al in 2009. Here, the numeric attributes are transformed into categorical ones. It was further improved through the Min Standard Deviation Roughness (SDR) algorithm and the Standard deviation Standard Deviation Roughness (SSDR) algorithm by Tripathy et al in 2011. A natural approach to deal with both types of attributes together, which extends all these algorithms is the Min Mean Neighbourhood Roughness (MMeNR) algorithm, which can be applied to uncertainty based heterogeneous attributes. This algorithm is shown to be superior to the above algorithms through the computation of F-measure and bench marked data sets like Teacher Assistant Evaluation Data Set and Acute Inflammations Data Set from UCI repository. Spatial datasets like the Forest Fire and the Abandoned Mine Land Inventory Data are used to show the superiority of MMeNR among all.

Keywords: Rough Set, Clustering, Neighbourhood, Geospatial Data, Epidemiology

Introduction

Huge amounts of data are generated every day in real world and study of these datasets is very important to understand their properties and infer their behaviour. Through the analysis carried out over these datasets in the form of finding rules of association, their similarity or dissimilarity is very much likely to provide a basis to study their behaviour in a group or otherwise. One technique in data mining which is helpful in analysis of data is the unsupervised learning phenomena of clustering data. This is a process of putting objects which are similar to each other basing upon some characteristics into a group and the objects which are dissimilar or less similar into different groups. One of the applications of clustering is to study any object from a cluster and similar characteristics are expected from the other elements of the cluster. So our hard earned knowledge from dealing with one element is utilised for gaining similar characteristics for other elements also in the same cluster, which otherwise would have been needed to be repeated. Clustering (Huang, 1998) helps in making groups of elements having similar characteristics and paves the way for studying their properties in granules instead of doing it individually. Clustering as an accessory step is utilized in several studies and application areas. (Chen et al, 2011) have used it for image segmentation. In (Chuang et al, 2006) it is used it for spatial data and image segmentation. (Chaira et al, 2011) have used clustering for tumour detection. (Endo et al. 2012) have developed an objective based clustering using rough set theory. (Gong et al, 2013) have used kernel metric and local information to develop a clustering algorithm, which does not use fuzzy set of rough set. For developing rules for diagnosis of liver disease using clustering was developed in (Karthik et al, 2011). Hybrid models of fuzzy set and rough set together for the study of fuzzy preorder and fuzzy topologies was carried out in (Tiwari et al, 2013). For intelligent medical diagnosis a framework was proposed by taking rough set and formal concept analysis in (Tripathy et al, 2011). (Lal et al, 2010) have studied clustering using rough set theory from granular computing point of view. One drawback of the above methods is that only numerical data sets or categorical data sets can be handled by using them but not both. An opinion comes out from the studies carried out by (Gibson, Kleinberg, & Raghavan, 2000; Guha, Rastogi, & Shim, 2000). The major reason behind this is that the multi-valued property of the categorical attributes. In this direction, two of the notable contributions come from (Guha et al., 2000) and (Gibson et al. 2000). But, most of the modern day datasets are having uncertainty inherent in them, which cannot be dealt with these algorithms. So, as a whole the above algorithms are not suitable for handling uncertainty based categorical data sets.

Several clustering algorithms based upon the uncertainty based models of fuzzy set (Tripathy et al 2015), rough set (Tripathy et al 2015), intuitionistic fuzzy set (Tripathy et al 2015) and their hybrid models (Dubois and Prade 1990; Tripathy et al 2015). These methods require as an input from the user, the required number of clusters, which is many times undesirable. These are depending upon the optimization of an objective function which in some sense captures the distances between objects and is needed to be minimised. Hence, the automatic detection of number of clusters by the algorithms became a necessity. A method which determines the splitting attribute for clustering was proposed by (Parmar et al., 2007) in the form of proposed the Min-Min-Roughness (MMR) algorithm. As in some of the algorithms mentioned above, this algorithm also cannot handle numeric data sets. It was observed by Kumar et al in 2009 that the selection of the splitting attribute in MMR is not logically correct. So, in order to rectify this drawback of MMR, to make it more efficient and applicable to heterogeneous data sets, a more powerful and efficient algorithm was proposed by them in the form of Min Mean Roughness (MMeR) in (Kumar et al, 2009). It was established by them that the replacement of min in MMR by mean in MMeR improves its accuracy in the form of purity ratio. Replacing the mean used in MMeR by standard deviation, which is known to be a better measure of central tendency than mean, an algorithm called as SDR (Tripathy and Ghosh 2011a) was developed and proved to be more efficient than MMeR. It was observed that the algorithm SSDR (Tripathy and Ghosh 2011b) proposed by replacing the first min in SDR by standard deviation does not improve the accuracy any further. For a detailed discussion on data clustering algorithms using rough sets a chapter by (Tripathy et al, 2013) can be referred.

Some of the algorithms mentioned in the above paragraph can handle hybrid data sets. But this is achieved by transforming the numerical attributes into categorical forms. Also, these algorithms are unlikely to be used for huge data sets as it is. The constraint of equivalence relation being used to define the rough set models used puts the constraint of dividing the universe into elementary granules and so is applicable to only a limited type of attributes. In (Hu et al., 2008) the basic rough set model was extended to propose a generalised rough set model called the neighbourhood rough set, which can efficiently handle both types of attributes without the requirement of one being transformed into the other type like it was done for MMeR, SDR or SSDR algorithms. The neighbourhood relations used are capable of handling decision classes through approximation with the help of neighbourhood granules. In this paper, we have tried to use this advantage of neighbourhood rough sets in proposing a generalised algorithm, which provides a natural way of handling both numerical and categorical attributes simultaneously and is better than MMeR, SDR and SSDR. The new algorithm is termed as the Min-Mean Neighbourhood Roughness (MMeNR) algorithm. To facilitate the comparison among these algorithms we use the two well-known and widely used measures of F-measure and purity ratio as it was done in the previous algorithms. Also, we have taken two of the Bench marked data sets from UCI repository in the form of the Teacher Assistant Evaluation data set and Acute Inflammations data set on which all these algorithms are applied for clustering and compared through the two measures mentioned above.

Several efforts have been made to use spatial properties of data in clustering. These have resulted in solutions to real life problems occurring from time to time. One of the earliest instances is in (Meng et al, 2005), where the transmission and distribution of Severe Acute Respiratory Syndrome (SARS) in Beijing was studied, which was extended in (Wang et al., 2006). A study of epidemic which occurred in Istanbul in the form of measles is found in (Ulugtekin et al., 2007). In (Chi & Zhu, 2008), statistical modelling and regression were used to study population dynamics. (Bai et al., 2010) studied the cause of defects in birth and how it is affected due to environmental factors using discernibility. An extensive deliberation on application of clustering algorithms in biomedical research is found in (Xu et al, 2010). (Kumar et al, 2017) have used neighbourhood based rough set theory for the classification and diagnosis of cardiac diseases.

A study on Neighbourhood rough set based Approximation Modelling for Spatial Epidemiology was carried out in (Tripathy et al., 2016. Incidence-Prevalence Patterns in Spatial Epidemiology via Neighborhood Rough Sets are explored in (Tripathy, B. K., & Sharmila Banu, K. (2017). A compilation of Neighbourhood rough set systems in Big Data Analysis is done In (Tripathy, 2017). Neighbourhood rough sets were used for knowledge acquisition through MapReduce was proposed in (Hiremath et al, 2015). Another application of neighbourhood based rough sets for knowledge acquisition using MapReduce from Big Data over Cloud Computing is studied in (Vishwakarma et al, 2014).

Neighbourhood rough set model has several advantages over the basic rough set model, like dealing with inconsistent data and better capability to reduce high dimensional data into lesser dimensions which make it suitable to use in analysis of spatial data. As our proposed algorithm is based upon neibourhood based rough sets, we have selected some such data sets for geospatial data analysis and epidemiology in order to compare and analyse its superiority over other rough set based competing algorithms applications. Besides proving its superiority, it also shows as how effective it can be when used for real-world applications. It may be noted that neighbourhood rough sets from the multigranular computing point of view was studied for their algebraic as well as topological properties in (Tripathy et al, 2014).

The presentation sequence in this paper is as follows. Next we present definitions and notations used in the rest of the paper followed by the proposed algorithm. We perform experiment with two bench marked data sets followed by introduction of spatial data, spatial epidemiology and perform experimentation on two spatial datasets to show that the proposed algorithm is superior to its contemporary algorithms in this direction also even on such datasets and show its applicability for such cases also.

Definitions and Notations

Neighbourhood rough set model was proposed by (Hu et al. 2008), which is a generalised model taking care of categorical and numeric attributes in a dataset and capable of all characteristics of basic rough sets; like dimensionality reduction and rule generation. The definition still follows the approximation of a set through two approximations; lower and upper approximations. The neighbourhoods are described through a threshold value. Precisely, the Minkowsky distance (Hu et al., 2008) characterizes a neighbourhood and its shape. Rule generation and reduction of number of rules techniques have been developed, which also leads to a reasoning process also.

The neighbourhood granules are outcomes of the neighbourhood relations defined on the universe. This also leads to the formulation of a uniform framework.

Let us denote the information system IS by the pair (U, A) where U is a universe of discourse and A is a set of attributes $\{A_1, A_2, \ldots A_n\}$. If IS happens to be a decision system then A can be decomposed into two sets C and D such that $A = C \bigcup D$ and $C \bigcap D = \phi$. The elements of 'C' are called the condition attributes and the elements of 'D' are called the decision attributes. The following definitions are taken from (Hu et al, 2008).

Definition 1 (Neighbourhood): Let B $\subseteq$ C. A distance function (metric) Δ over U is defined as $\Delta : U \times U \rightarrow R$ such that $\forall x_i, x_j, x_k \in U$

- $\Delta(x_i, x_j) \geq 0; \Delta(x_i, x_j) = 0 \text{ iff } x_i = x_j;$
- $\Delta(x_i, x_j) = \Delta(x_j, x_i)$ and
- $\Delta(x_i, x_k) \leq \Delta(x_i, x_j) + \Delta(x_j, x_k)$

Let Δ^B be a metric over (U, B). Then for any real number δ we define the $\delta-\text{neighbourhood}$ of $x_i \in U$ with respect to B as

$$\delta_B(x_i) = \{x_j \mid x_j \in U, \Delta^B(x_i, x_j) \leq \delta\}.$$

Definition 2 (Neighbourhood Granules): B₁ is a set of numerical attribute such that $B_1 \subseteq A$ and B₂ is a set of categorical attribute such that $B_2 \subseteq A$.The neighbourhood granule of sample x induced by B_1, B_2 and $B_1 \bigcup B_2$ are defined as

For numerical attributes -

$$\delta_{B_1}(x) = \{x_i \mid \Delta_{B_1}(x, x_i) \leq \delta, x_i \in U\}; \tag{1}$$

For categorical attributes -

$$\delta_{B_2}(x)=\{x_i \mid \Delta_{B_2}(x,x_i)=0, x_i \in U\}; \tag{2}$$

For hybrid attributes -

$$\delta_{B_1 \cup B_2}(x)=\{x_i \mid \Delta_{B_1}(x,x_i)\le \delta \, and \, \Delta_{B_2}(x,x_i)=0, x_i \in U\}, \tag{3}$$

Therefore Definition 2 is applicable to numerical, categorical data and their mixture.

For an information system (U, A) and a metric Δ over it, for any pre-defined real number δ, the family of neighbourhood granules $\{\delta(x_i) \mid x_i \in U\}$ forms a cover over U. This neighbourhood granule system induces a neighbourhood relation N over U. This neighbourhood relation is reflexive and symmetric but not transitive. The objects in a neighbourhood granule are similar to each other.

Definition 3: The lower and upper approximations of X where X ⊆ U can be obtained for a neighbourhood relation N over U defined as (U, N) as

$$\underline{N}X=\{x_i \mid \delta(x_i)\subseteq X, x_i \in U\}, \tag{4}$$

$$\overline{N}X=\{x_i \mid \delta(x_i)\cap X \ne \phi, x_i \in U\}. \tag{5}$$

$\underline{N}X$ and $\overline{N}X$ are called the lower and upper neighbourhood approximations of X respectively. We have, $\underline{N}X \subseteq X \subseteq \overline{N}X$. X is rough with respect to N iff $\underline{N}X \ne \overline{N}X$. The uncertainty region of X is denoted by BN(X) and is given by $\overline{N}X-\underline{N}X$.

Definition 4 (Neighbourhood Roughness): The neighbourhood roughness of X with respect to B is denoted by $NR_B(X)$ and is given by (6).

$$NR_B(X)=1-\frac{\underline{N}X}{\overline{N}X} \tag{6}$$

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 5 (Relative neighbourhood roughness): Let the lower and upper approximations of X with respect to {aj} be given by $\underline{NX_{a_j}(a_i=\alpha)}$ and $\overline{NX_{a_j}(a_i=\alpha)}$, then

$$NR_{a_j}(X/a_i=\alpha)=1-\frac{\left|\underline{NX_{a_j}(a_i=\alpha)}\right|}{\left|\overline{NX_{a_j}(a_i=\alpha)}\right|}, \text{ where } a_i,a_j \in A \text{ and } a_i \ne a_j. \tag{7}$$

This is the neighbourhood roughness of a_i in reference to a_j.

Definition 6 (Mean neighbourhood roughness): The mean neighbourhood roughness for the equivalence class $a_i=\alpha$ is denoted by $MeNR(a_i=\alpha)$ and is defined as

$$MeNR(a_i=\alpha)=(\sum_{j=1, j\ne i}^{n} NR_{a_j}(X/a_i=\alpha))/(n-1). \tag{8}$$

Definition 7 (Min mean neighbourhood roughness): Let β,γ,δ,χ . . . and so on be the values other than α of the attribute aᵢ. 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

$$\mathrm{MMeNR(a_i)} = \mathrm{Min(MeNR(a_i} = \alpha), \mathrm{MeNR(a_i} = \beta), \mathrm{MeNR(a_i} = \gamma), \mathrm{MeNR(a_i} = \delta), \ldots). \tag{9}$$

Definition 8 (Distance of relevance): DR for relevance of things is:

$$DR(B,C) = \sum_{i=1}^{n}(b_i, c_i) \tag{10}$$

Here B and C are objects and b_i and c_i are their values respectively, under the ith 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 9 (Precision and Recall): It is the measure of the number of relevant items out of those picked. Let TP = True positive, FP = False positive and FN = False negative. Then

$$\mathrm{P} = \mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{11}$$

Recall is the measure of number of relevant items picked out of the total number of relevant items.

$$\mathrm{R} = \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{12}$$

Definition 10 (F-measure): It measures the accuracy of the process. Its range is [0, 1]. In fact, it is the harmonic mean of P and R. Precisely,

$$F = \frac{2 * P * R}{P + R} \tag{13}$$

Example

Through this example we explain the concept of neighbourhood.

Table 1. 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 the table above, we have 5 objects from x_1 to x_5 and. two attributes A and B, which describe them. The attributes A and B are respectively categorical and numerical by nature. Let the neighbourhood of an object 'x' be denoted by $\delta(x)$. Then from Table 1, applying (2), we get the neighbourhoods of each of the objects with respect to the attribute A as:

$$\delta(x_1)=\{x_1\}, \delta(x_2)=\{x_2,x_3\},\ \delta(x_3)=\{x_2,x_3\},\ \delta(x_4)=\{x_4,x_5\},\ \delta(x_5)=\{x_4,x_5\}$$

Taking the value $\delta = 0.1$, from Table 1 and applying (1), we get the neighbourhoods of objects with respect to attribute B as follows:

$$\delta(x_1)=\{x_1,x_2\},\ \delta(x_2)=\{x_1,x_2\},\ \delta(x_3)=\{x_3,x_4,x_5\},\ \delta(x_4)=\{x_3,x_4,x_5\},\ \delta(x_5)=\{x_3,x_4,x_5\}.$$

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 neighbourhood generated by it is an equivalence class. With respect to the attribute A we have three classes; class of '1'= $X_1=\{x_1\}$, class of '2' = $X_2=\{x_2,x_3\}$ and class of '3' = $X_3=\{x_4,x_5\}$. Similarly, with respect to the attribute we have two classes; $Y_1=\{x_1,x_2\}$ and $Y_2=\{x_3,x_4,x_5\}$.

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

$$\underline{N}X_1=\phi,\ \overline{N}X_1=\{x_1,x_2\}$$

$$\underline{N}X_2=\phi,\ \overline{N}X_2=\{x_1,x_2,x_3,x_4,x_5\}$$

$$\underline{N}X_3=\phi,\ \overline{N}X_3=\{x_3,x_4,x_5\}$$

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

$$\underline{N}Y_1=\{x_1\},\ \overline{N}Y_1=\{x_1,x_2,x_3\}$$

$$\underline{N}Y_2=\{x_4,x_5\},\ \overline{N}Y_2=\{x_2,x_3,x_4,x_5\}$$

We now present our proposed algorithm Min Mean Neighbourhood Roughness (MMeNR) below.

Proposed Algorithm

The procedure of MMeNR is given below.

```
Procedure MMeNR(U, k)
Start
N = 1 // Number of clusters at this point
PN = U //Parent node
Loop1:
If(N ≠1 and N< k)
    PN = Proc PN (N)
End if
  // PN clustering begins
$\forall$ $a_i$ determine equivalence classes( $a_i$ )
Determine neighbourhood using Definition 2.
Find Neighbourhood Roughness using Definitions 3 and 5.
Calculate the MMeNR given by definition 7.
Take the minimum of all MMeNR given by different attributes from $a_i$, $a_j$...
Next the splitting attribute is determined. Let it be $a_i$.
```

15. On a_i binary split is performed.

16. This could be done by taking the equivalence class whose roughness value is nearer to the roughness of the splitting attribute a_i.

17. No_of_leafnodes = N.

18. Go to Loop 1

19. Stop

20. Procedure: PN(N)

21. j = 1

22. While (j < N)

23. If the Avg-distance of cluster j is already computed

24. Then goto L

25. Else

26. m = Count (Cluster j elements).

27. Avg Dist (i) = 2* (DR)/ (m*(m-1)).

28. Label :

29. i++

30. Loop

31. Find Max (Avg-distance (j))

32. Send back (Entities in cluster j) which have Max (Avg-distance (j))

33. Stop

Comparison of MMeNR with MMeR AND SDR

PYTHON has been used to write the codes and implement the whole technique. F-measure is used to determine 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.

Experiment 1 (Teacher Assistant Evaluation Data Set)

It is found in the UCI repository. The table has 5 attributes out of which 1 is numeric and the rest are categorical. It has 151 rows. It has 3 clusters. Tables 2, 3 and 4 present the structures of the clusters formed by using the algorithms.

Table 2: F-measure of Teacher Assistant Evaluation Data Set using MMeNR

Cluster Number	**C1**	**C2**	**C3**	**Precision**	**Recall**	**F-measure**
1	2	10	20	0.72	0.92	0.807
2	0	0	2	0.6	1	0.75
3	47	40	30	0.92	0.61	0.733
Overall F-measure						0.763

Table 3: F-measure of Teacher Assistant Evaluation Data Set using MMeR

Cluster Number	**C1**	**C2**	**C3**	**Precision**	**Recall**	**F-measure**
1	2	6	15	0.288	0.652	0.4
2	3	5	12	0.231	0.6	0.334
3	44	39	25	0.9	0.41	0.56
Overall F-measure						0.43

Table 4: F-measure of Teacher Assistant Evaluation Data Set using SDR

Cluster Number	**C1**	**C2**	**C3**	**Precision**	**Recall**	**F-measure**
1	2	8	15	0.308	0.652	0.417
2	0	2	7	0.341	0.8	0.478
3	47	40	30	0.92	0.393	0.54
Overall F-measure						0.478

The F-measure value for MMeNR is the highest among all the F-measures for the three algorithms which establishes its superiority.

Experiment 2 (Acute Inflammations Data Set)

This dataset is also available in the UCI repository and has 6 attributes, out of which 5 are of categorical type and the other is a numeric one. There are 120 rows in this table. There is a decision attribute with "yes" or "no" values for it. Table 5, Table 6 and Table 7 describe the number of elements in the clusters, precision values, recall values and the F-measures for each cluster and the overall F-measure values for the clusters with respect to the respective algorithms.

Table 5: F-measure of Acute Inflammation Data Set using MMeNR

Cluster Number	**C1**	**C2**	**Precision**	**Recall**	**F-measure**
1	44	15	0.881	0.913	0.897
2	15	46	0.836	0.876	0.873
Overall F-measure					0.885

Table 6: F-measure of Acute Inflammation Data Set using MMeR

Cluster Number	**C1**	**C2**	**Precision**	**Recall**	**F-measure**
1	49	10	0.831	0.831	0.831
2	10	51	0.836	0.836	0.836
Overall F-measure					0.8335

Table 7: F-measure of Acute Inflammation Data Set using SDR

Cluster Number	**C1**	**C2**	**Precision**	**Recall**	**F-measure**
1	70	21	1	0.77	0.87
2	0	29	0.681	0.84	0.84

Overall F-measure	0.855

The highest F-measure value is obtained for the MMeNR algorithm, which is 0.885 followed by 0.855 for MMeR and 8.335 for MMeR. So, the proposed algorithm is the most efficient among the three.

Experiments performed on the two UCI repository datasets supports the claim that MMeNR is the best among the rough set based clustering algorithms for hybrid data. In fact it was established in (Tripathy et al, 2011b) that SSDR does not have better efficiency than SDR. Also, MMR is inferior to MMeR. So, among all these partition based rough set clustering algorithms MMeNR is the best.

Comparison of MMeNR with MMeR and SDR on Geospatial Data

The advantage of neighbourhood based rough sets is more distinctly evident when we consider geospatial data. Let us have a closer look at geospatial data analysis to have better appreciation.

Geospatial Data Analysis

We have to deal with spatial data irrespective of our field of interest and we do it consciously or subconsciously. This type of data, which is better known as geospatial data deals with any type of data related to particular places on the surface of earth. This data can be in several formats and inherently equipped with data beyond the location specific one on the surface of earth. Additional information to that of location on earth are provided in the form of attributes and there may be several such attributes associated with spatial data. These attributes may be the height, width and capacity of a dam at some location.

The spaces of interest in various studies are spatial in character; i.e. these are related to space. This space can be geographical space or any artificial spaces created by human beings for their studies of different phenomena.

Attempts have been made to store and manipulate such datasets since the proposal of relational databases. This has necessitated the requirement of capability to manage with geometric objects in large volume. The spatial database system deals with objects in space and not images or pictures in space. The notion of spatial database systems are different form the related image database systems as the former deals with objects having locations, relationships and well-defined extents whereas the later deals with raster images (Frank, 1991). An elaborative account of extensible database systems, spatial data structures an index structures, spatial reasoning, geographic information systems, quad trees, and thematic map modelling is found in the proceedings of the First International Symposium on Large Spatial Databases (Buchmann et al, 1989).

We can define a spatial database system as a database system that has a data model which provides spatial data types for implementation, a query language, spatial indexing and algorithms to facilitate spatial joins.

To reduce the losses in natural phenomena like earthquake, cyclone, farming, global warming and stock market, their early prediction is of much importance. Extrapolation cane used to generate information about futuristic events from the existing events. The general approach is to form similarity blocks in the area of study. The specific values for the blocks can be obtained. From these values predictive analysis can be performed to generate futuristic information. So, any of the phenomena occurring block wise can be extrapolated to predict and that will reduce the loss due to early notification. Outliers play an important role in spatial data analysis as they are different from the consistent data and their analysis can generate previously unknown knowledge. For such analysis scatter plots are used.

Groups of spatial data objects can be formed through clustering of such data and thus generating blocks of spatial data, so that the hot spots can be determined which we can be used in analysis of epidemics, analysis of criminals etc. Most importantly, location of events are used as an attribute while forming clusters of spatial data and this helps in identifying hotspots.

Finding technology for GIS is the main force for spatial database system researches. Out of the two views followed in modelling the objects in space view model cities, forests and rivers, while the second view of space partitions of the space.

Spatial Epidemiology

Spatial epidemiology deals with the study of health data mapped geographically (Wang et al., 2006). Several risk factors are involved in the analysis of such data. These may be related to case, ecology, infection, demography or behaviour. Knowledge discovery through spatial data mining is used to study epidemiology in (Ranjan et al, 2012).

Development of decision making systems which can infer from epidemiological data is conducive to the progresses made in the field of soft computing or more generally in data science. Rough set theory is appropriate to such decision making procedures. The decision systems with spatial attributes are no exceptions. In (Tripathy et al, 2016) a study of spatial epidemiology using neighbourhood approximation and rough fuzzy set theory was carried out. Studies like those use spatial auto-correlation leads to the creation of hotspots (Tripathy et al, 2017) and are typically applied to spatial datasets. Tobler's first law states that the objects which are nearer to each other have a stronger relation among themselves than those which are relatively far. This strongly favours the use of neighbourhood rough sets instead of normal rough sets as spatial auto correlation has this core notion.

Dealing with health of a community instead of any individual through the development of Geographic Information Systems (GIS) has shown improved results. Although infrastructure for GIS seems to be costly but in comparison to its advantage in dealing with spatial attributes, it looks advantageous. But, the utility for several situations also adds to this advantage.

Studies of epidemiological data are useful in tracing the prevalence of these epidemics and also to develop plans for improving the occurrence of such epidemics in future. Geo-referencing is perhaps the best characteristic of the study of spatial databases. The knowledge obtained from the occurrence or pervasive information in one region can be mapped efficiently to other regions spatially through comparisons. These studies can provide interventions and will help in finding corrective measures. We have proposed to use Neighbourhood Rough Sets in this work so that the grid based study can be extended from one disease to many other diseases, thus helping in the study of incidence prevalence. The basic concept used is the similarity of regions. If a certain corrective measure works in a specific region then the same measure can be carried out in all other regions which are similar to it.

It has been shown in (Tripathy et al, 2017) that the study of epidemiological datasets using their spatial dimension generates crucial information in the form of patterns which are related to incidence and prevalence of some diseases and help in preventing epidemics.

Two characteristically different geospatial datasets are used in this study which helps in analysing epidemiology.

Forest Fire Dataset

This dataset is available in the UCI repository, which is a multivariate one with 517 instances with 13 attributes each. The outcome of this dataset is to come up with burnt areas due to forest fires. The occurrence of forest fires may be due to human negligence or lightning and causes serious damages in the form of property and also becomes a serious threat to mankind. In spite of serious steps being taken to control by spending enough money it remains a threat.

On the contrary early speculation of its occurrence can be helpful in its control. Some of the solutions which occur are in the form of using Satellite related phenomena, infrared sensors and Meteorological sensors. It involves acquiring satellites which is costly and their maintenance is also requires enough expenditure.

However, this problem has been demonstrated to be solved by analysis techniques developed for geospatial data analysis techniques (Miller, 2004). Data mining techniques like grouping of the conditions have been found to be helpful in correlating areas of Forest Fire and as a consequence their happening can be prevented from the beginning or at an early stage.

Abandoned Mine Land Inventory Data Set

Areas after abandonment of coal mines become sources of health hazards for community due to the leftover garbage, dusts and atmospheric contamination. Mining may be very short lived or may be live for a few decades. The process may not be generating good revenue or maybe there is termination of resources, which lead to abandonment. The tenure may be computed even before the mine opens. The life of a mine depends upon the quality and quantity of minerals coming out of it. Sometimes even if the quantity is not zero, but the revenue generated may be low such that running the mine becomes infeasible.

These abandoned mines become sources of hazard to safety as the work is left as it is after the mine is closed. Acid drainage, soil erosion and heavy metal contamination pollute the ponds, rivers and streams. The attack of such health hazards is similar to community hazards and is not having lesser effect than the epidemiology. Also, old mines pose physical health hazards. Study of such problems is important as people are exposed to such environments in large number. Incidence prevalence of these abandoned land mine areas can be carried out through application of appropriate clustering algorithms (Wang et al, 2006). These studies will lead to taking corrective measures in these effected areas or can predict the prevalence of health hazards.

Experimental Setup

The above mentioned two data sets are used to establish the applicability of the MMeNR algorithm in geospatial data analysis and epidemiology. The common characteristic of the concerned datasets is that they contain regression based decision classes. The pre-processing steps required for preparing the datasets include visualization of clusters through scatterplots, removal of outliers using the plot, getting a rough estimate for the number of clusters, binning the datasets which helps to make them ready for clustering to take place. F-measure is used as an estimation of the efficiency of the algorithm as such reflecting the results expected when it is applied to real world datasets.

Experiment 1 (Forest Fire Data Set)

As mentioned above, this is a dataset from UCI repository, wherein there are 6 attributes each of numeric and categorical kind with a total of 517 entities. As the decision is based upon the area attribute having continuous data, it was needed to be converted to interval bins before clustering takes place. It was observed that the number of clusters is 5. This leads to binning of the attribute values to assign them to these 5 intervals. Table 8 reflects the cluster counts. Precision, recall and F-measures obtained thereof.

Table 8: Overall F-measure for MMeNR on Forest Fire Dataset

Cluster Number	**C1**	**C2**	**C3**	**C4**	**C5**	**Precision**	**Recall**	**F-measure**
Cluster 1	3	1	0	0	0	0.57	0.67	0.616
Cluster 2	2	1	0	0	0	0.61	0.8	0.69
Cluster 3	4	1	0	0	0	0.62	0.75	0.68
Cluster 4	0	16	0	0	0	0.84	1	0.913
Cluster 5	258	0	5	5	4	0.97	0.95	0.96
Overall F-measure								0.7718

Table 9: Overall F-measure for MMeR on Forest Fire Dataset

Cluster Number	**C1**	**C2**	**C3**	**C4**	**C5**	**Precision**	**Recall**	**F-measure**
Cluster 1	2	1	0	0	0	0.071	0.67	0.127
Cluster 2	4	1	0	0	0	0.15	0.8	0.104

Cluster 3	3	1	0	0	0	0.12	0.75	0.207
Cluster 4	0	16	0	0	0	0.84	1	0.913
Cluster 5	238	20	5	5	4	0.93	0.91	0.919
Overall F-measure								0.454

Table 10: Overall F-measure for SDR on Forest Fire Dataset

Cluster Number	**C1**	**C2**	**C3**	**C4**	**C5**	**Precision**	**Recall**	**F-measure**
Cluster 1	4	2	0	0	0	0.463	0.51	0.486
Cluster 2	3	1	0	0	0	0.41	0.7	0.502
Cluster 3	5	3	0	0	0	0.52	0.55	0.534
Cluster 4	2	14	0	0	0	0.72	0.93	0.408
Cluster 5	252	6	5	5	4	0.91	0.89	0.899
Overall F-measure								0.566

Hence by using our proposed algorithm on this geospatial based dataset we get an overall F-measure of 0.7718 which is much higher than that for MMeR and SDR, which are 0,454 and 0.566 as shown in Tables 9 and 10 respectively.

Experiment 2 (Abandoned Mine Land Inventory Sites Data Set)

We have already discussed about the characteristics of this dataset. It is supposed to create health hazards and needs to be analysed at an early stage to take care of public health and the safety of the people around them. There are 2 numerical and 5 categorical attributes in this dataset and 313 tuples. As the decision attribute is continuous one in the form of area, to discretize it, the dataset is first converted into interval bins for the purpose of clustering. Using scatterplot technique we find that the optimal number of clusters to be formed is 3. As a consequence the values are binned to put them into 3 numbers of intervals. In Table 11, we reflect the 3 different cluster values, precisions, recalls and F-measures of the clusters of this dataset after the clustering process is carried out using the proposed MMeNR algorithm.

Table 11: Overall F-measure for MMeNR on Abandoned Mine Land Inventory Dataset

Cluster Number	**C1**	**C2**	**C3**	**Precision**	**Recall**	**F-measure**
1	77	5	2	0.79	0.916	0.86
2	163	23	7	0.61	0.845	0.714
3	29	4	2	0.78	0.828	0.803
Overall F-measure						0.792

Table 12: Overall F-measure for MMeR on Abandoned Mine Land Inventory Dataset

Cluster Number	**C1**	**C2**	**C3**	**Precision**	**Recall**	**F-measure**
1	70	9	5	0.71	0.83	0.765
2	158	28	7	0.58	0.74	0.65
3	24	5	6	0.62	0.764	0.683

Overall F-measure	0.699

Table 13: Overall F-measure for SDR on Abandoned Mine Land Inventory Dataset

Cluster Number	C1	C2	C3	Precision	Recall	F-measure
1	73	8	3	0.75	0.85	0.79
2	160	25	8	0.61	0.76	0.676
3	25	4	3	0.7	0.68	0.68
Overall F-measure						0.715

The accuracy of MMeNR is 0.792 as is obtained in Table 11 and in comparison the same for the other two algorithms under comparison are lower at the values 0.699 and 0.715 as shown in Tables 12 and 13 respectively.

After the two studies, we find that the proposed algorithm MMeNR works better than its competitors even for geospatial datasets and hence has a better prospect than them in the study of geospatial data analysis and epidemiology. The information gained through it from the datasets is very much beneficial in the study of epidemiology through the geospatial data as it has the capability of handling irregular information, handling both types of categorical and numerical attributes, use of minimal attribute set through the computation of reducts and handling epidemiology in the form of spatial data analysis.

Conclusions

Data clustering algorithms are either applicable to only nominal or only numerical datasets. The first of the rough set based algorithm MMR, was capable of handling categorical data only. Its extensions in the form of MMeR, SDR and SSDR although handle hybrid datasets they require artificial transformation of numerical into categorical form. The MMeNR algorithm proposed in this work is not only naturally applicable to hybrid datasets but also more efficient than MMeR, SDR and SSDR which is established through F-measure. It is shown that this is the case for geospatial data also. Taking the neighbourhood granules in the numerical spaces helps in clustering of spatial hybrid data and epidemic datasets.

References

1. Bai, H., Ge, Y., Wang, J.-F., & Lan Liao, Y. (2010). Using rough set theory to identify villages affected by birth defects: the example of Heshun, Shanxi, China. *International Journal of Geographical Information Science*, 24(4), 559–576. https://doi.org/10.1080/13658810902960079
2. Buchmann, A., Gfinther, O., Smith, T.R., & Wang, Y.E. (1989). Eds. Design and implementation of large spatial Databases, *Proceedings of the First International Symposium on Large Spatial Databases*, Santa Barbara. https://doi.org/10.1007/3-540-52208-5
3. Chaira, T., & Anand, S. (2011). A Novel Intuitionistic Fuzzy Approach for Tumor/Hemorrhage Detection in Medical Images, *Journal of Scientific and Industrial Research*, 70(6), 427-434. http://nopr.niscair.res.in/handle/123456789/11922
4. Chen, L., Chen, C. P., & Lu, M. (2011). A multiple kernel fuzzy c-means algorithm for image segmentation, *IEEE Transactions on Systems, Man and Cybernetics*, Part B (Cybernetics), 41(5), 1263-1274. https://doi.org/10.1109/TSMCB.2011.2124455
5. Chi, G., & Zhu, J. (2008). Spatial regression models for demographic analysis. *Population Research and Policy Review*, 27(1), 17–42. https://doi.org/10.1007/s11113-007-9051-8
6. Chuang, K., Tzeng, H.L., Chen, S., Wu, J., & Chen, T. J. (2006). Fuzzy c-means clustering with spatial information for image segmentation, *Computerized medical imaging and graphics*, 30 (1), 9-15. https://doi.org/10.1016/j.compmedimag.2005.10.001

7. Dubois, D., & Prade, H. (1990). Rough fuzzy sets and fuzzy rough sets, *International Journal of General System*, 17(2-3), 191-209. https://doi.org/10.1080/03081079008935107
8. Endo, Y., & Kinoshita, N. (2012). On objective based Rough C-means clustering, *Proceedings of the IEEE conference on Granular Computing*, 1-6. https://doi.org/10.1109/GrC.2012.6468682
9. Frank, A. (1991). Properties of geographic data: Requirements for spatial access methods, *Proceedings of the Second International Symposium on Large Spatial Databases*, Zürich. https://doi.org/10.1007/3-540-54414-3_40
10. Gibson, D., Kleinberg, J., & Raghavan, P. (2000). Clustering categorical data: An approach based on dynamical systems. *The VLDB Journal*, 8(3–4), 222–236. https://doi.org/10.1007/s007780050005
11. Gong, M., Liang, Y., Shi, J., Ma, W., & Ma, J. (2013). Fuzzy c-means clustering with local information and kernel metric for image segmentation, *IEEE Transactions on image processing*, 22(2), pp.573-584. https://doi.org/10.1109/TIP.2012.2219547
12. Guha, S., Rastogi, R., & Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5), 345–366. https://doi.org/10.1016/S0306-4379(00)00022-3
13. Hiremath, Shruthi, Pallavi Chandra, Anne Mary Joy & Tripathy, B. K. (2015). Neighbourhood rough set model for knowledge acquisition using MapReduce, *Int. J. Communication Networks and Distributed Systems*, 15(2/3), 212-234. https://doi.org/10.1504/IJCNDS.2015.070975
14. Hu, Q., Yu, D., Liu, J., & Wu, C. (2008). Neighborhood rough set based heterogeneous feature subset selection. *Information Sciences*, 178(18), 3577–3594. https://doi.org/10.1016/j.ins.2008.05.024
15. Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3), 283–304. https://doi.org/10.1023/A:1009769707641
16. Karthik, S., Priyadarshini, A., Anuradha, J., & Tripathy, B. (2011). Classification and rule extraction using rough set for diagnosis of liver disease and its types, Adv. Appl. Sci. Res, 2(3), 334-345.
17. Kumar, P., & Tripathy, B. K. (2009). MMeR: an algorithm for clustering heterogeneous data using rough set theory. *International Journal of Rapid Manufacturing*, *1*(2), 189-207. https://doi.org/10.1504/IJRAPIDM.2009.029382
18. Kumar, S. U., & Inbarani, H. H. (2017). Neighborhood rough set based ECG signal classification for diagnosis of cardiac diseases. *Soft Computing*, *21*(16), 4721-4733. https://doi.org/10.1007/s00500-016-2080-7
19. Lal, K. M., & Acharjya, D. P. (2010). Knowledge Granulation, Association Rules and Granular Computing, *Proceedings of Second National Conference on Advanced Technologies in Electrical Engineering*, Virudhunagar, Tamil Nadu, India, 43-46.
20. Parmar, D., Wu, T., & Blackhurst, J. (2007). MMR: An algorithm for clustering categorical data using Rough Set Theory. *Data & Knowledge Engineering*, 63(3), 879–893. https://doi.org/10.1016/j.datak.2007.05.005
21. Ranjan, S. N., Sinha, A. K., & Singh, J. B. (2012). The study of knowledge discovery with spatial Data Mining in Epidemiology Database, *International Journal of Engineering Research and Technology (IJERT)*, 1(6), 1-11. ISSN: 2278-0181
22. Tiwari, S. P., & Srivastava A. K. (2013). Fuzzy rough set, fuzzy preorders and fuzzy topologies, *Fuzzy sets and systems*, 210, 63-68. https://doi.org/10.1016/j.fss.2012.06.001
23. Tripathy, B. K., & Ghosh, A. (2011a). SDR: An algorithm for clustering categorical data using rough set theory. In *2011 IEEE Recent Advances in Intelligent Computational Systems*, 867-872. https://doi.org/10.1109/RAICS.2011.6069433
24. Tripathy, B. K., & Ghosh A. (2011b). SSDR: An Algorithm for Clustering Categorical Data Using Rough Set Theory, *Advances in Applied science Research*, 2(3), 314-326. https://doi.org/10.1109/RAICS.2011.6069433
25. Tripathy, B. K., Acharjya, D.P., & Cynthya, V. (2011). A Framework for Intelligent Medical Diagnosis Using Rough Set With Formal Concept Analysis, *International Journal of Artificial Intelligence and Applications*, 2(2), 45-66. https://doi.org/10.5121/ijaia.2011.2204
26. Tripathy, B. K., & Ghosh, A. (2013). Data clustering algorithms using rough sets. In *Handbook of Research on Computational Intelligence for Engineering, Science, and Business*, 297-327. IGI Global.

https://doi.org/10.4018/978-1-4666-2518-1.ch012

27. Tripathy, B. K., & Mitra, A. (2014). On Algebraic and Topological Properties of Neighbourhood Based Multigranular Rough Sets. In *2014 International Conference on Computer Communication and Informatics*, 1-6. IEEE. https://doi.org/10.1109/ICCCI.2014.6921770
28. Tripathy, B.K., & Anuradha, J. (2015). *Soft Computing- Advances and Applications*, Cengage Learning publishers, New Delhi. ISBN: 9788131526194
29. Tripathy, B. K., & Sharmila Banu (2016). Rough Fuzzy Set Theory and Neighbourhood Approximation Based Modelling for Spatial Epidemiology, *Handbook of Research on Computational Intelligence Applications in Bioinformatics*, (Eds: Sujata Das and Bidyadhar Subudhi), IGI publications, Chapter-6, pp.108-118. https://doi.org/10.4018/978-1-5225-0427-6.ch006
30. Tripathy, B. K., & Sharmila Banu, K. (2017). Exploring incidence-prevalence patterns in spatial epidemiology via neighborhood rough sets. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, *12*(1), 30-43. https://doi.org/10.4018/IJHISI.2017010103
31. Tripathy, B. K. (2017). Rough Set and Neighborhood Systems in Big Data Analysis. In: *Computational Intelligence Applications in Business Intelligence and Big Data Analytics (Eds: Vijayan Sugumaran, Arun Kumar Sangaiah, Arunkumar Thangavelu), CRC press,* Auerbach Publications. pp. 261-282. ISBN 9781498761017
32. Ulugtekin, N., Alkoy, S., & Seker, D. Z. (2007). Use of a geographic information system in an epidemiological study of measles in Istanbul. *Journal of International Medical Research*, *35*(1), 150–154. https://doi.org/10.1177/147323000703500117
33. Vishwakarma, H.R., Tripathy, B. K., & D.P. Kothari (2014). Neighbourhood Based Knowledge Acquisition Using MapReduce from Big Data over Cloud Computing, *Proceedings CSIBIG14*, pp.183-188. https://doi.org/10.1109/CSIBIG.2014.7056958
34. Wang, J., McMichael, A. J., Meng, B., Becker, N. G., Han, W., Glass, K., & Zheng, X. (2006). Spatial dynamics of an epidemic of severe acute respiratory syndrome in an urban area. *Bulletin of the World Health Organization*, *84*, 965-968. https://doi.org/10.2471/BLT.06.030247
35. Xu, R., & Wunsch, D. (2010). Clustering algorithms in biomedical research: A review, *IEEE Rev. Biomed Eng*, 3, 120-154. https://doi.org/10.1109/RBME.2010.2083647

Conflict of Interest: The authors declare that they have no conflict of interest.

Author Biographies:

B.K. Tripathy is working at present as a professor and the Dean of SITE School, VIT, Vellore, India. He has published more than 600 technical papers in international journals, conference proceedings and edited research volumes. He has supervised 52 candidates for research degrees. Dr. Tripathy has published two books, 6 research volumes, monographs and has guest edited some research journals. Prof. Tripathy is in the editorial board or reviewer of more than 100 international journals of repute. Also, he has delivered keynote speeches in several international conferences, invited talks in FDPs, virtual conferences and seminars. Dr. Tripathy is a senior member of IEEE, ACM, IRSS and CSI and life member of many other professional bodies. His current research interest includes Fuzzy Sets and Systems, Rough Set theory, Data Clustering, Social Network Analysis, Neighbourhood Systems, Soft Sets, SIOT, Big Data Analytics, Multiset theory, Decision Support Systems, DNN and Pattern Recognition.

Akarsh Goyal is a Data Engineer at Hulu. He is based in Los Angeles. He completed his MS in Computer Science (specialization Data Science) from University of Southern California. He has been extensively involved in working in the field of data as part of my internships, research works and currently his full-time. He has done extensive research work in soft computing, data mining, machine learning & big data and has a good research publication record of past 4 years in various top journals. He is also a reviewer for Elsevier and has received certificates by Elsevier for outstanding contribution in reviewing in the realm of machine learning and fuzzy systems.