Clustering mixed data using neighbourhood rough sets

Sharmila Banu Kather* and B.K. Tripathy

School of Computer Science and Engineering,
VIT University,
Vellore, Tamilnadu, India
Email: sharmilabanu.k@vit.ac.in
Email: tripathybk@vit.ac.in
*Corresponding author

Abstract: Data in varied nature and huge quantities are being generated every day. They range from tabulated, structured and semi-structured as well as numerical or categorical in terms of attributes. Data pre-processing presents data in a favourable format to apply analytics algorithm and derive knowledge therein. Data analytics has revolutionised millennial mankind unwinding the knowledge and patterns mined from data. Clustering is an unsupervised learning pattern which has popular algorithms based on distance, density, dimensions and other functions. These algorithms are operational on numerical attributes and special algorithms for data involving categorical features are also reported. In this paper we propose a straight forward way of clustering data involving both numerical and categorical features based on neighbourhood rough sets. It does not include calculation of any extra parameters like entropy, saliency, dependency or call for discretisation of data. Hence its complexity is lesser than algorithms proposed for categorical or mixed data and offers better efficiency.

Keywords: clustering; mixed; categorical and numerical data; continuous data; rough sets; neighbourhood rough sets; granulation.

Reference to this paper should be made as follows: Kather, S.B. and Tripathy, B.K. (2020) 'Clustering mixed data using neighbourhood rough sets', *Int. J. Advanced Intelligence Paradigms*, Vol. 15, No. 1, pp.1–16.

Biographical notes: Sharmila Banu Kather received her BE in Computer Science and Engineering in 2003 from the University of Madras, Tamil Nadu, India and MTech in Computer Science and Engineering in 2005 from VIT University, Vellore, Tamil Nadu, India. Currently, she is a Doctoral candidate at VIT University, Vellore, Tamil Nadu, India. Her research work focuses on the applications of neighbourhood rough sets in clustering and spatial epidemiology. She has been working as an Assistant Professor (Senior) in the School of Computer Science and Engineering at VIT University, Vellore, India. She has two years of industry experience and seven years of teaching experience.

B.K. Tripathy is a Senior Professor in School of Computer Science and Engineering, VIT University, Vellore, India. He has received research/academic fellowships from UGC, DST, SERC and DOE of Govt. of India. He has published more than 380 technical papers and has produced 26 PhDs, 13 MPhils and 3 MS (by research) under his supervision. He has published two books on soft computing and computer graphics. He is a life/senior member of IEEE, ACM, IRSS, CSI and IMS. He is an editorial

board member/reviewer of more than 70 journals. 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, multi-sets and list theory.

1 Introduction

Data clustering is the process of forming groups of similar objects from a given set of data basing upon some similarity criterion such that the elements in any group are more similar to each other than elements in different groups. Data clustering has several applications in computer science like image processing, pattern recognition, database anonymisation, medical diagnosis, weather forecasting, etc. The clusters are regrouped based on a distance function again and again until the desired number of clusters is reached. This approach has found useful applications where the data has been numerical. In efforts to address the categorical nature of data, Ganti et al. (1999), Gibson et al. (2000) and Guha et al. (2000) have proposed suitable approaches. Some algorithms advocate converting nominal features to numerical type and apply numerical algorithm (Gibson et al., 2000). Expectation maximisation algorithm involves probabilities and using hypergraphs needs disjoint partitions (Dempster et al., 1977). Dissimilarity among the objects with respect to the attributes is used as the bottom line for clustering. In these algorithms, multiple passes are run to obtain the required number of clusters. The stability of clusters thus attained changes with the values assigned in the beginning. Besides, most of these algorithms place an object categorically in one grouping only. A representation of measurements of real-world cannot be crisp or an object can belong in more than one group.

The earlier instance of data clustering algorithm is the hard C-means algorithm (HCM) (MacQueen, 1967). However, the modern day databases have inherent uncertainty in themselves. So, HCM loses its efficiency when applied to most of them. HCM generates clusters which are disjoint from each other in the sense there is no chance of a data element belonging to more than one clusters. This is a big restriction and hence necessitated the development of clustering algorithms based on uncertainty-based models. Uncertainty models are applied in image processing, classification and clustering of vague and imprecise data (Ripon et al., 2016; Inbarani and Kumar, 2015; Roy et al., 2014; Ripon et al., 2016). This list of models is pretty long. We have ANN (McCulloch and Pitts, 1943; Hebb, 1949), fuzzy sets (Zadeh, 1965), rough sets (Pawlak, 1982), intuitionistic fuzzy sets (Atanassov, 1986) and also their hybrid models like the rough fuzzy (Dubois and Prade, 1990) and rough intuitionistic fuzzy sets (Tripathy et al., 2013). Also, several variants of these basic uncertainty-based algorithms also exist; like their Kernelised versions, possibilistic versions and Kernelised possibilistic versions. The fuzzy C-means (FCM) algorithm was initially proposed by Ruspini (1969) and the objective function approach was introduced by Bezdek et al. (1984). Several improved versions of FCM are found in literature (Tripathy et al., 2013; Ruspini, 1969; Bezdek et al., 1984; Xu and Wu, 2010; Maji and Pal, 2007). The intuitionistic fuzzy set model is a generalisation of the fuzzy sets and so the intuitionistic fuzzy C-means (IFCM) (Xu and Wu, 2010) has been found to be superior to FCM. There are two different approaches to

rough set-based clustering algorithms. The first one is due to Mazlack et al. (2000) and the second one is the rough C-means (RCM) due to Lingras and West (2004). While the first one uses the partitioning attribute approach the second one is in the same line as the FCM and IFCM. The hybrid rough fuzzy C-means (RFCM) algorithm was first introduced by Mitra et al. (2006) and immediately generalised by Maji and Pal (2007). The rough intuitionistic fuzzy C-means (RIFCM) algorithm was introduced in Bhargava et al. (2013) and as expected has been found to be the best among its family of algorithms. All these algorithms use the Euclidean distance between the tuples for similarity measure. This has some problems like the linear separability of data. In order to handle these problems kernel functions were used instead of the Euclidean distance to measure similarity and the Kernelised versions of all the above algorithms; KFCM (Graves and Pedrycz, 2010), KIFCM (Lin, 2014). KRCM (Tripathy et al., 2012), KRFCM (Bhargava and Tripathy, 2013) and KRIFCM (Tripathy et al., 2014) have been developed. Also, possibilistic versions of all these two families of C-means algorithms have been established. It has been observed that the kernels do not provide any definite trend as none of the Kernelised algorithms is found to be the most efficient for all types of data sets (Mittal and Tripathy, 2015). However, our focus in this paper is in the second direction; that is that of Mazlack et al. (2000) and Darshit et al. (2007). One problem with rough sets is that they are suitable for categorical attributes and are not directly applicable to numeric data sets unless some measures are taken for their applicability. However, an improved version of MMR called the MMeR was developed by Kumar and Tripathy (2009), which handle numeric datasets with some discretisation procedure being added. Clustering algorithms like squeezer (He et al., 2002), link-based clustering (He et al., 2004), extension of k-means (Huang, 1998) and based on hypergraphs (Han et al., 1997) have been proposed for categorical data. But they assume the data is precise and are not uncertain. The proposed algorithm MMNR handles uncertainty and does not include discretisation or any other conversion. Neighbourhood systems were introduced by Lin (1988). However, the neighbourhood-based rough sets were introduced by Hu et al. in 2008 and these types of rough sets are capable of handling hybrid data sets very efficiently. So, in order to extend the capability of MMR to handle the hybrid datasets, we propose an improved algorithm using neighbourhood rough sets instead of basic rough sets which we call as the min-min-neighbourhood rough set (MMNR) algorithm. We establish through experimentation that MMNR is much more efficient than MMR in clustering datasets.

Mazlack et al. (2000) proposed a rough set-based technique in order to select partitioning attributes for clustering. They use a measure called total roughness to determine the crispness of the partitions. Different ways of partitioning can lead to different total roughness. By comparing total roughness for different partitioning the most suitable one can be selected. However, for partitioning, the method starts with binary valued attributes and uses the total roughness criterion only for multi-valued attributes. So, partitioning is done on a binary attribute even though the total roughness for a multi-valued attribute is lower. This problem is handled very efficiently in the MMR algorithm proposed by Darshit et al. (2007), which performs clustering of the objects basing upon all the attributes. In addition, MMR proposes a new way to measure data similarities based on the roughness concept. MMR utilises a measure termed mean roughness comparable to that proposed by Mazlack et al. (2000) based on rough set theory. In this work, we have proposed an algorithm which works for multi-valued categorical and numerical attributes based on granulation and

approximation – neighbourhood rough sets. Neighbourhood rough sets intuitively prevent loss of information as it can deal with continuous data without having to discretise it as needed by rough sets to establish indiscernibility. Loss of information is reported with discretisation (Jensen and Shen, 2004). Neighbourhood rough sets are an improved alternative with prospect of the natural data grouping and are very useful in domains with continuous data like weather data, time-series data, remote sensing data etc., and group them without discretising.

2 Approximate reasoning and rough set theory

The notion of fuzzy set was introduced by Zadeh in the year 1965 in order to extend the modelling capability of the crisp sets and it has been found to be a fruitful model in many real life applications. But fuzzy set has the drawback of depending upon the definition of membership function, which cannot be specified in a unique manner. Also, like many other models of uncertainty before it and also in the field of artificial intelligence, the notions of uncertainty and vagueness are not differentiated. The notion of rough set was introduced by Pawlak in the year 1982 which is the only model which differentiates these two notions. It does not depend upon any other notion like the dependence of fuzzy sets on membership function and statistical models upon probability. More importantly it follows the characteristic of defining uncertainty using the boundary region approach proposed by Frege (1948), the father on modern logic. The notion is motivated by definition of the notion of knowledge proposed by Pawlak. He says that human knowledge is dependent upon their capability to classify objects. From the mathematical point of view Pawlak took the concept of equivalence relations, which are known to induce classifications on the universes on which they are defined. A subset of a universe is rough or not with respect to one or more equivalence relations defined on the universe. To achieve this he introduced two crisp sets associated with any subset of the universe and one or more equivalence relations, called the lower and upper approximations of the subset. If it happens that the lower and upper approximations are identical then the subset is definable with respect to the equivalence relation or the indiscernibility relation generated by the group of equivalence relations, which is the intersection of these relations. We define this notion mathematically below.

Let U be a universe of discourse and P be a set of equivalence relations defined over U. Let Q be a subset of P. We denote by $IND(Q)$, the intersection of the equivalence relations in Q. It is well known that $R = IND(Q)$ is an equivalence relation on U. Let us denote the equivalence class of any element x in U by $[x]_R$. Then the lower and upper approximations of a subset X of U with respect to R are denoted by $\underline{R}X$ and $\overline{R}X$ defined as

$$\underline{R}X = \{x \in U \mid [x]_R \subseteq X\} \tag{1}$$

$$\overline{R}X = \{x \in U \mid [x]_R \cap X \neq \phi\} \tag{2}$$

X is rough with respect to R if and only if $\underline{R}X \neq \overline{R}X$ and R-definable otherwise. When X is rough with respect to R, it has non-empty uncertainty region. We denote the uncertainty region by $BN_R(X)$ and define it as

$$BN_R(X) = \overline{R}X \setminus \underline{R}X \tag{3}$$

Data clustering which started with the HCM (MacQueen, 1967) was extended to the FCM algorithm by Ruspini (1969). This was further extended and the notion of objective function was introduced by Bezdek (1984). This is an important breakthrough in uncertainty-based data clustering and has been followed consistently thereafter to develop many algorithms. In fact the first data clustering algorithm was introduced by Lingras and West in 2004. It has been observed in literature that the hybrid algorithms provide better results. In fact it was observed to be true when two RFCM algorithms were introduced by Mitra et al in 2006 and Maji and Pal in 2007. This was further extended to develop rough intuitionistic fuzzy algorithms by Tripathy et al. (2013) and Bhargava et al. (2013). However, in between an algorithm called the MMR algorithm was introduced by Darshit et al. (2007) from a different angle. This algorithm is based upon the basic rough sets. Hence, it can handle only categorical data. However, hybrid datasets are common in today's world. Clustering of these datasets cannot be done by using MMR algorithm. Neighbourhood-based rough sets are introduced by Hu et al. 2008) in order to handle hybrid data sets. So, it is the aim of this paper to used neighbourhood-based rough sets instead of basic rough sets in developing an algorithm which will extend MMR in the sense that it will be convenient to cluster hybrid datasets using this algorithm.

3 Neighbourhood rough sets

It is well known that rough set theory the two key issues are granulation and approximation. Granulation is to divide the whole universe into several subsets with respect to certain criterion. Any subset of the universe is approximated with respect to these granules through crisp subsets of the universe called the lower and upper approximation of the subset. The basic rough sets introduced by Pawlak use equivalence relations to provide this granulation through the equivalence classes. This is quite adequate for categorical attributes in a database. But attributes having numerical attributes are difficult to handle with this approach without additional techniques. However, neighbourhood relations can be used to handle numerical attributes. The concept of neighbourhood rough sets was introduced by Lin (1988). We formally define these models as follows:

Let us consider an information system $K = (U, A)$, where U is a nonempty finite set of tuples $\{u_1, u_2, \ldots u_n\}$ and A is a set of attributes $\{A_1, A_2, \ldots A_m\}$ defined over it. A decision system is a special type of information system where $A = C \cup D$ such that C is called the set of condition attributes and D is called the set of decision attributes. Next we introduce the concept of a metric over a universe.

3.1 Neighbourhood granulation

Definition 1: A metric (sometimes called as the distance function) d is a mapping from $U \times U \rightarrow R$ such that for $\forall x, y, z \in U$, the following three properties are satisfied:

$$d(x, y) \geq 0 \text{ and } d(x, y) = 0 \text{ if and only if } x = y \tag{4}$$

$$d(x, y) = d(y, x) \tag{5}$$

$$d(x, z) \leq d(x, y) + d(y, z) \tag{6}$$

Definition 2: The δ-neighbourhood of an element $x \in U$ with respect to a subset of attributes B of C is denoted by $\delta_B(x)$ and is defined as

$$\delta_B(x) = \{y \in U \mid d_B(x, y) \leq \delta\} \tag{7}$$

Here, d_B is the metric associated with the set of attributes B.

There are three standard metrics available in literature and are very popular also. These are the Manhattan's distance, the Euclidean distance and Chebyshev's distance. All these are special cases of the Minkowsky distance, which we define below.

Definition 3: Let x and y be two elements in an N-dimensional space determined by the attribute set $A = \{A_1, A_2, \ldots A_N\}$. Let us denote by $F(x, A_i)$, the value of the element x under the attribute A_i, $I = 1, 2, \ldots N$. Then Minkowsky's distance M_p, $p \geq 1$is defined as

$$M_p(x, y) = \left(\sum_{i=1}^{N} | F(x, A_i) - F(y, A_i) |^p \right) \tag{8}$$

The Minkowsky's function reduces to the Manhattan's distance when $p = 1$, to the Euclidean distance when $p = 2$ and to the Chebychev's distance when $p = \infty$.

Definition 4: For two subsets B_1 and B_2 of A where B_1 is a set of numerical attributes and B_2 is a set of categorical attributes, the neighbourhood granules of an element x in U with respect to B_1, B_2 and $B_1 \cup B_2$ are denoted by $\delta_{B_1}(x)$, $\delta_{B_2}(x)$ and $\delta_{B_1 \cup B_2}(x)$ respectively and are defined as

$$\delta_{B_1}(x) = \{y \in U \mid d_{B_1}(x, y) \leq \delta\} \tag{9}$$

$$\delta_{B_2}(x) = \{y \in U \mid d_{B_2}(x, y) = 0\} \tag{10}$$

$$\delta_{B_1 \cup B_2}(x) = \{y \in U \mid d_{B_1}(x, y) \leq \delta \wedge d_{B_2}(x, y) = 0\} \tag{11}$$

3.2 Neighbourhood approximation

Definition 5: Given a metric space (U, d), the family of δ-neighbourhood granules $\{\delta(x) | x \in U\}$ forms a granule system, which is a cover of U instead of being a partition. It is to note that none of the granules $\delta(x)$ is empty and their union is the whole universe U.

Definition: let (U, A) be an information system. Then we define a neighbourhood relation N on U given by a matrix $M(N) = (r_{ij})$, where

$$r_{ij} = \begin{cases} 1 & \text{if } d(x_i, x_j) \leq \delta; \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Definition 6: Given a set of objects U and a neighbourhood relation N over U, we say that (U, N) is a neighbourhood approximation space. With respect to this relation, we define the lower and upper approximation of a subset X of U as

$$\underline{N}X = \{x \in U \mid \delta(x) \subseteq X\}$$
$$\overline{N}X = \{x \in U \mid \delta(x) \bigcap X \neq \phi\} \tag{13}$$

We say that X is N-rough if $\underline{N}X \neq \overline{N}X,$, otherwise, X is said to be N-definable. We denote the boundary region of X with respect to N by $BN_N(X)$ and define it as:

$$BN_N(X) = \overline{N}X \setminus \underline{N}X \tag{14}$$

Note: the neighbourhood granulation system depends upon the threshold value δ strictly. Varying the value of δ we can get different neighbourhood relations. We have used the δ for the threshold value as well as the neighbourhood granule x. the difference is clear from the context.

3.3 Neighbourhood-based roughness

Definition 7: Neighbourhood-based roughness is defined as 1 – ratio of the number of objects occurring in *neighbourhood-based* lower approximation to the number of object occurring in the upper approximation as in equation (13).

$$NR_{opr}(X) = 1 - \frac{\left|\underline{R}_{opr}(X)\right|}{\left|\overline{R}_{opr}(X)\right|} \tag{14}$$

If this measure is less than 1, it is rough with respect to the attribute sub set *opr* and if it is equal to 1, there is no vagueness and it is crisp.

Definition 8: Mean neighbourhood roughness is the mean of neighbourhood roughness of each unique value set (of attributes). Neighbourhood roughness of each attribute with respect to every other attribute is tabulated. If v_1, v_2, v_3, … ,v_n are the values in the domain of an attribute f, then we calculate a neighbourhood-based lower and upper approximation of that attribute for each of the values – v_1, v_2, v_3, … ,v_n with respect to every other attribute and take a mean of those values. Equation (15) calculates neighbourhood roughness of an attribute *fa* with respect to another attribute *fb* for the attribute value $v1$where *fa* and *fb* are any two attributes between which Neighbourhood roughness is calculated.

$$NR_{fb}\left(X \mid fa = v1\right) = 1 - \frac{\left|\underline{X_{fb}(fa = v1)}\right|}{\left|X_{fb}\ (fa = v1)\right|} \tag{15}$$

Similarly, neighbourhood roughness between *fa* and *fb* are calculated for each of the domain values in the value set V (*fa*) of *fa* and their mean is calculated. This neighbourhood roughness has to be calculated between every pair of attributes. Equation (13) gives the neighbourhood-based upper and lower approximation of a subset based on an attribute. Further, the ensuing novel algorithm applies this to calculate the neighbourhood-based roughness between two attributes, categorical or numerical. Hence equation (13) is upheld in the following form.

$$\begin{aligned} \underline{X_{fb}\ (fa = v_i)} &= \left\{ x_i \in U \middle| \delta(x_i\ ) \subseteq X,\ x_i \in U\ \right\} \\ \overline{X_{fb}\left(fa = v_i\right)} &= \left\{ x_i \in U\ | \delta\left(x_i\right) \cap X,\ x_i \in U \right\} \end{aligned} \tag{16}$$

Neighbourhood of each x_i in U is defined as the union of the neighbourhoods x_i of *fa* and *fb*.

$$\delta\left(x_i\right) = \delta\left(x_i\left(fa\right)\right) \cup \delta\left(x_i\left(fb\right)\right) \tag{17}$$

Mean neighbourhood roughness

$$fb(fa) = \frac{NR_{fb}\left(X \mid fa = v1\right) + NR_{fb}\left(X \middle| fa = v2\right) + \ldots + NR_{fb}\left(X \middle| fa = vn\right)}{\left|V(fa)\right|} \tag{18}$$

If there are *k* attributes, *mean neighbourhood roughness* of an attribute is calculated with respect to all the attributes and the attribute with the *minimum* of that *mean neighbourhood roughness* is chosen for partitioning.

3.4 Minimum of minimum neighbourhood-based roughness algorithm

function MMNR (*U*, *c*)
Begin
Initialise current_cluster_count cn to 1
Initialise BaseSet cluster_sets to *U*
Loop for *c* times
  BaseSet = fnBaseSet (cn, cluster_sets)
  For each $A_l \in A$ (l = 1 to M) where M is the number of attributes; considering all objects in BaseSet
    Determine the neighbourhood of each object in A_l
    For each $A_p \in A$ (p = 1 to M and $p \neq$ l)
Determine neighbourhood roughness of A_l with respect to A_p, $NR_{A_p}\left(A_l\right)$
    End for
    Determine minimum of neighbourhood roughness, $MNR_{A_l} = Minimum\left(NR_{A_p}\left(A_l\right)\right)$
  End for
  Determine minimum of $MNR_{A_l} = Minimum\left(NR_{A_p}\left(A_l\right)\right)$ where l = 1, 2, … , M (for all attributes)
  Using minimum of *MNR*, determine the attribute about which BaseSet is partitioned.
  Partition the BaseSet
  Assign current_cluster_count = number of partitions arrived at
End loop
End
Set *fn* BaseSet (current_cluster_count, cluster_sets)

Begin
Initialise v to 1
While (v < = current_cluster_count) repeat
  Count_in_v = number of objects in cluster_sets (v)
  $v = v + 1$
End while
Determine the cluster with the largest value for count_in_v, max_count_in_v
Return the cluster with max_count_in_v
End

The MMNR algorithm partitions the objects to obtain better clusters from mixed data. It accepts the number of clusters c as input and stops after generating c clusters from the above sequence. It is essential to highlight that the neighbourhood roughness of one attribute with respect to every other attribute in the base set is calculated using equation (14). The subset along which the neighbourhood-based lower and upper approximation are determined, is the union of neighbourhoods of objects of attributes A_1 with A_p using equation (17).

Table 1 consists of a sample dataset which contains both categorical and numerical data. Of all the algorithms that cluster categorical data, approximation based algorithms has been effective. Rough sets-based clustering algorithms proposed by Mazlack et al. (2000) and Darshit et al. (2007) have laid new pathways to explore more possibilities. These algorithms use a *roughness* measure that is calculated across attributes. But they can be applicable only to categorical data as they apply equivalence classes. Kumar and Tripathy (2009) proposed MMeR which could handle mixed data. Although it has been effective in clustering, it calls for conversion logic from numeric to categorical and proceed with the clustering process. Our proposed work finds the neighbourhood of attributes either categorical or numerical. For numerical cases, neighbourhood is calculated based on Minkowsky' distance formula.

Table 1 Sample data set

	A1	*A2*	*A3*	*A4*	*A5*	*A6*	*A7*
1	Good	Dark	Opaque	0.2	region1	Torrid	Grade index2
2	Average	Grey	Translucent	0.85	region3	Tropical	Grade index 3
3	Fair	Light grey	Transparent	0.31	region2	Temperate	Grade index 1
4	Average	Dark	Translucent	0.74	region2	Torrid	Grade index 2
5	Fair	Light grey	Translucent	0.82	region1	Torrid	Grade index 3
6	Good	Very dark	Opaque	0.72	region3	Tropical	Grade index 1
7	Fair	Light grey	Opaque	0.9	region1	Arctic	Grade index 1
8	Fair	Light grey	Transparent	0.64	region1	Torrid	Grade index 1
9	Good	Very dark	Opaque	0.56	region3	Tropical	Grade index 3
10	Average	Very dark	Translucent	0.4	region3	Torrid	Grade index 3

Considering two x_i, x_j objects in a dimension, $f(x, A_i)$ represents the value of object x in attribute A_i, a Minkowsky distance defined by

$$\delta_p\left(x_i, x_j\right) = \left(\sum_{j=1}^{M} \left| f\left(x_i, A_i\right) - f\left(x_j, A_i\right) \right|^p \right)^{1/p} \tag{19}$$

$\delta_p(x_i, x_j)$ is Manhattan distance if $p = 1$, Euclidean distance if $p = 2$ and Chebychev distance if $p = \infty$. $\delta_A(x_i)$ is the neighbourhood of data object x_i and its size is based on the threshold δ. Neighbourhood roughness is calculated based on the neighbourhood and the roughness value can evaluate to maximum or minimum based on this threshold.

Table 2 Neighbourhood roughness calculation of A1 with respect to every other attribute

A1	*X1 = fair*	*X2 = average*	*X3 = good*	*Average*
A2	0	0.8333	1	0.6111
A3	0.5714	0	1	0.5238
A4	0.875	1	0.8333	0.9028
A5	1	1	1	1
A6	0.7142	1	1	0.9048
A7	1	1	1	1

Table 2 shows neighbourhood roughness of attribute A1 for each of its subset corresponding with a value with respect to all other attribute. The average of roughness is determined and considered as neighbourhood roughness between A1 and other attributes respectively which is tabulated in Table 4. Table 3 shows a similar calculation of Neighbourhood roughness between A2 and other attributes like Table 2.

Table 3 Neighbourhood roughness calculation of A2 with respect to every other attribute

A2	*X1 = dark*	*X2 = grey*	*X3 = light grey*	*X4 = very dark*	*Average*
A1	1	1	0	1	0.75
A3	1	1	0.5714	1	0.8925
A4	0.8	1	0.8889	1	0.9225
A5	1	1	1	1	1
A6	1	1	1	0.7143	0.9285
A7	0	1	1	1	0.75

Table 4 Neighbourhood roughness across attributes

	A1	*A2*	*A3*	*A4*	*A5*	*A6*	*A7*
A1		0.6111	0.5238	0.9028	1	0.9058	1
A2	0.7500		0.8295	0.9223	1	0.9285	0.7500
A3	0.6238	0.9333		0.9583	1	0.9074	1
A4	1	0.9281	1		1	0.9635	1
A5	1	0.6670	1	0.9167		0.7639	1
A6	1	0.8820	1	0.9750	1		0.9500
A7	1	0.6250	1	0.9333	1	0.9333	

From the neighbourhood roughness of each attribute with respect to every other attribute, the attribute having minimum value is chosen for the split of universal set. The split is based on the domain of the chosen attribute. From the synthetic data used for demonstration, attribute A1 has the minimum value when compared with other attributes. The domain of A1 imposes three neighbourhoods which is partitioned the universe into two, one based on *good* and *average* and another based on *fair.* This is due to the minimum of neighbourhood roughness contributed by *fair* being more than the other two put together. After the first iteration two partitions are obtained and the partition with more number of elements is considered for the subsequent iteration. This process repeats till the desired number of clusters is realised.

4 Cluster analysis

Evaluating the purity of clusters obtained is essential and has been reported in Guha et al. (2000), Darshit et al. (2007) and Kumar and Tripathy (2009). It is observed as the ratio of cardinality of data occurring in the cluster as well as in the neighbourhood to the cardinality of the universe of objects considered for clustering.

$$\text{Cluster Purity} = \frac{\text{cardinality of data occuring in the cluster as well as in the neighbourhood to the cardinality}}{\text{cardinality of the universe of objects}} \tag{20}$$

$$\text{Total Purity} = \frac{\text{Sum of purity of all clusters}}{\text{Total number of clusters}} \tag{21}$$

5 Experimental results

Using the cluster purity approach mentioned in Section 4, we applied MMNR on UCI repository's *zoo* and *soybean* datasets (Lichman, 2013). Soybean dataset comprises of four diseases – *diaporthe stem canker, charcoal rot, rhisoctonia root rot ns phytophthora rot.* The number of clusters was chosen as four as there are four categories of diseases. Upon applying MMNR, the total purity of clusters generated was 92% as depicted in Table 5.

Table 5 MMNR applied on the *soybean* data

Cluster	*Disease1*	*Disease2*	*Disease3*	*Disease4*	*Purity*
1	0	10	0	0	1
2	10	0	0	0	1
3	0	0	8	17	0.68
4	0	0	2	0	1
					0.92

The algorithm was applied to cluster the zoo dataset with over 100 data items and around 18 attributes. There are seven categories of animals in the zoo dataset and hence the number of clusters was chosen as seven. The total purity of the clusters generated was 82.5% as tabulated in Table 6.

Table 6 MMNR applied on the *zoo* data

Cluster	*Class1*	*Class2*	*Class3*	*Class4*	*Class5*	*Class6*	*Class7*	*Purity*
1	0	20	0	0	0	0	0	1
2	0	0	0	0	0	0	0	1
3	0	0	0	0	1	0	0	1
4	39	0	2	0	3	0	0	0.886
5	0	0	2	12	0	0	0	0.857
6	2	0	0	1	0	0	1	0.5
7	0	0	0	0	0	8	9	0.529
								0.825

When compared with other hybrid algorithms we see that it clusters mixed data using an *iterative* and *merge* logic which will account for extra computations. This approach does not need extra logic to include numerical data and neither conversion to nominal form. It proposes natural grouping of numerical data in numerical form without any extra computation. With n data items and m attributes, it calculates neighbourhoods in nm time and roughness in n^2 time and hence generates clusters in polynomial time. MMNR offers the same complexity as MMR algorithm but only has the extra potential to account for mixed data. An intuitive heuristic is used for partitioning which is reduction-based and not on the domain. This accounts for progressive splitting which enhances the clustering process with natural grouping. It is data-driven and can be applied across various domains.

6 Results and discussion

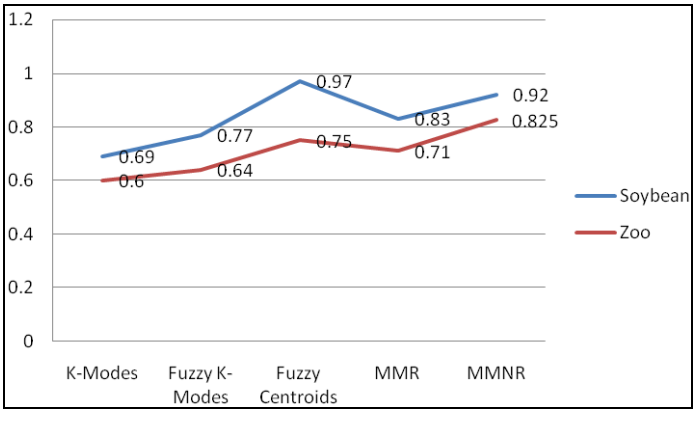
The proposed MMNR algorithm is compared with other algorithms. We can observe from Table 7 that it outperforms all algorithms MMR (Darshit et al., 2007), K-modes (Chaturvedi et al., 2001), fuzzy k-modes (Huang and Ng, 1999) except fuzzy centroids (Kim et al., 2004). There is always a concern on the choice of memberships attributed to data items and this requires heavy iterations. Further, MMNR determines clusters without multiple passes and hence will be scalable for dataset of any size and deliver with competent purity. One potential scope for enhancement would be with choosing clusters for further partitions. The algorithm chooses the cluster with maximum number of data items. This approach can be refined to identify a cluster with farthest neighbourhood for subsequent iteration which will improve the scalability metric and can be applied for very large data sets. Our proposed algorithm accepts *number of clusters* as the only input and generates the clusters based on this input without any exclusive conversion logic and hence is more robust and can work across different domains efficiently.

Table 7 Comparative analysis of algorithms

Data	*K-modes*	*Fuzzy k-modes*	*Fuzzy centroids*	*MMR*	*MMNR*
Soybean	0.69	0.77	0.97	0.83	0.92
Zoo	0.60	0.64	0.75	0.71	0.825

Graphical representation also highlights the performance of MMNR in consensus with the above tabulation. It should also be noted that MMNR works efficiently with increasing amount of data.

Figure 1 Graphical representation of efficiency of algorithms (see online version for colours)



7 Conclusions

Algorithms that handle uncertainty in data and in clustering process are the need of the hour. The MMNR algorithm treats mixed data without distorting them and neither calls for any conversion logic. It retains originality and treats the data in entirety thereby offering freedom from information loss. Except on the choice of the number of clusters, any other intervention is not called for. The algorithm offers the promise of approximation strategies and efficiency in polynomial time. The neighbourhood threshold considered is 0.1 to calculate the neighbourhood of numerical attribute. The size of the neighbourhood influences the roughness calculated. On our future endeavour, we will explore an interval of neighbourhoods. We will also explore the clusters generated in each iteration and choose an alternative strategy to pick the cluster for further iteration from the current choice of cluster with maximum cardinality. We will also study the possibilities of hybrid approximation approaches.

References

Atanassov, K.T. (1986) 'Intuitionistic fuzzy sets', *Fuzzy Sets and Systems*, Vol. 20, No. 1, pp.87–96, Saleha (Rough Intuitionistic Fuzzy Sets).

Bezdek, J.C., Ehrlich, R. and Full, W. (1984) 'FCM: the fuzzy c-means clustering algorithm', *Computers & Geosciences*, Vol. 10, Nos. 2–3, pp.191–203.

Bhargava, R. and Tripathy, B. (2013) 'Kernel based rough-fuzzy c-means', *International Conference on Pattern Recognition and Machine Intelligence*, Springer, Berlin Heidelberg, pp.148–155.

Bhargava, R., Tripathy, B.K., Tripathy, A., Dhull, R., Verma, E. and Swarnalatha, P. (2013) 'Rough intuitionistic fuzzy C-means algorithm and a comparative analysis', *Proceedings of the 6th ACM India Computing Convention*, p.23.

Chaturvedi, A., Green, P.E. and Caroll, J.D. (2001) 'K-modes clustering', *Journal of Classification*, Vol. 18, No. 1, pp.35–55.

Darshit, P., Terasa, W. and Jennifer, B. (2007) 'MMR: an algorithm for clustering categorical data using rough set theory', *Data & Knowledge Engineering*, Vol. 63, pp.879–893.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society, Series B (Methodological)*, pp.1–38.

Dubois, D. and Prade, H. (1990) 'Rough fuzzy sets and fuzzy rough sets', *International Journal of General System*, Vol. 17, Nos. 2–3, pp.191–209.

Frege, G. (1948) 'Sense and reference', *The Philosophical Review*, Vol. 57, No. 3, pp.209–230.

Ganti, V., Gehrke, J. and Ramakrishnan, R. (1999) 'CACTUS – clustering categorical data using summaries', *Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.73–83.

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

Graves, D. and Pedrycz, W. (2010) 'Kernel-based fuzzy clustering and fuzzy clustering: a comparative experimental study', *Fuzzy Sets and Systems*, Vol. 161, No. 4, pp.522–543.

Guha, S., Rastogi, R. and Shim, K. (2000) 'ROCK: a robust clustering algorithm for categorical attributes', *Information Systems*, Vol. 25, No. 5, pp.345–366.

Han, E., Karypis, G., Kumar, V. and Mobasher, B. (1997) 'Clustering based on association rule hypergraphs', *Workshop on Research Issues on Data Mining and Knowledge Discovery*, pp.9–13.

He, Z., Xu, X. and Deng, S. (2002) 'Squeezer: an efficient algorithm for clustering categorical data', *Journal of Computer Science & Technology*, Vol. 17, No. 5, pp.611–624.

He, Z., Xu, X. and Deng, S. (2004) 'A link clustering based approach for clustering categorical data', *Proceedings of the WAIM Conference* [online] http://xxx.sf.nchc.org.tw/ftp/cs/papers/0412/0412019.pdf (accessed 7 May 2016).

Hebb, D. (1949) *The Organization of Behavior*, Wiley, New York.

Hu, Q., Yu, D., Liu, J. and Wu, C. (2008) 'Neighborhood rough set based heterogeneous feature subset selection', *Information Sciences*, Vol. 178, No. 18, pp.3577–3594.

Huang, Z. (1998) '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.

Huang, Z. and Ng, M.K. (1999) 'A fuzzy k-modes algorithm for clustering categorical data', *IEEE Transactions on Fuzzy Systems*, Vol. 7, No. 4, pp.446–452.

Inbarani, H. and Kumar, S. (2015) 'Hybrid TRS-FA clustering approach for web2.0 social tagging system', *International Journal of Rough Sets and Data Analysis (IJRSDA)*, Vol. 2, No. 1, pp.70–87.

Jensen, R. and Shen, Q. (2004) 'Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 12, pp.1457–1471.

Kim, D., Lee, K., Lee, D. (2004) 'Fuzzy clustering of categorical data using fuzzy centroids', *Pattern Recognition Letters*, Vol. 25, No. 11, pp.1263–1271.

Kumar, P. and Tripathy, B.K. (2009) 'MMeR: an algorithm for clustering heterogeneous data using rough set theory', *International Journal of Rapid Manufacturing*, Vol. 1, No. 2, pp.189–207.

Lichman, M. (2013) *UCI Machine Learning Repository*, University of California, School of Information and Computer Science, Irvine, CA [online] http://archive.ics.uci.edu/ml (accessed 4 February 2016).

Lin, K.P. (2014) 'A novel evolutionary kernel intuitionistic fuzzy-means clustering algorithm', *IEEE Transactions on Fuzzy Systems*, Vol. 22, No. 5, pp.1074–1087.

Lin, T.Y. (1988) 'Neighborhood systems and approximation in relational databases and knowledge bases', *Proceedings of 4th International Symposium on Methodologies of Intelligent Systems.*

Lingras, P. and West, C. (2004) 'Interval set clustering of web users with rough k-means', *Journal of Intelligent Information Systems*, Vol. 23, No. 1, pp.5–16.

MacQueen, J. (1967) 'Some methods for classification and analysis of multivariate observations', *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, No. 14, pp.281–297.

Maji, P. and Pal, S.K. (2007) 'RFCM: a hybrid clustering algorithm using rough and fuzzy sets', *Fundamenta Informaticae*, Vol. 80, No. 4, pp.475–496.

Maji, P. and Pal, S.K. (2007) 'Rough set based generalized fuzzy-means algorithm and quantitative indices', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, Vol. 37, No. 6, pp.1529–1540.

Mazlack, L., He, A., Zhu, Y. and Coppock, S. (2000) 'A rough set approach in choosing partitioning attributes', *Proceedings of the ISCA 13th International Conference*, 2000, pp.1–6.

McCulloch, W. and Pitts, W. (1943) 'A logical calculus of ideas immanent in nervous activity', *Bulletin of Mathematical Biophysics*, Vol. 5, No. 4, pp.115–133.

Mitra, S., Banka, H. and Pedrycz, W. (2006) 'Rough and fuzzy collaborative clustering', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, Vol. 36, No. 4, pp.795–805.

Mittal, D. and Tripathy, B.K. (2015) 'Efficiency analysis of kernel functions in uncertainty based c-means algorithms', *2015 International Conference in Advances in Computing, Communications and Informatics (ICACCI)*, pp.807–813.

Pawlak, Z. (1982) 'Rough sets', *International Journal of Computer & Information Sciences*, Vol. 11, No. 5, pp.341–356.

Ripon, S.H., Kamal, S., Hossain, S. and Dey, N. (2016) 'Theoretical analysis of different classifiers under reduction rough data set: a brief proposal', *International Journal of Rough Sets and Data Analysis (IJRSDA)*, Vol. 3, No. 3, pp.1–20.

Ripon, S.H., Kamal, S., Hossain, S. and Dey, N. (2016) 'Theoretical analysis of different classifiers under reduction rough data set: a brief proposal', *International Journal of Rough Sets and Data Analysis (IJRSDA)*, Vol. 3, No. 3, pp.1–20.

Roy, P., Goswami, S., Chakraborty, S., Azar, A.T. and Dey, N. (2014) 'Image segmentation using rough set theory: a review', *International Journal of Rough Sets and Data Analysis (IJRSDA)*, Vol. 1, No. 2, pp.62–74.

Ruspini, E.H. (1969) 'A new approach to clustering', *Information and Control*, Vol. 15, No. 1, pp.22–32.

Tripathy, B.K., Bhargava, R., Tripathy, A., Dhull, R., Verma, E. and Swarnalatha, P. (2013) 'Rough intuitionistic fuzzy c-means algorithm and a comparative analysis', *Proceedings of ACM Compute*, pp.21–22.

Tripathy, B.K., Ghosh, A. and Panda, G.K. (2012) 'Kernel based K-means clustering using rough set', *2012 International Conference in Computer Communication and Informatics (ICCCI)*, pp.1–5.

Tripathy, B.K., Tripathy, A., Govindarajulu, K. and Bhargav, R. (2014) 'On kernel based rough intuitionistic fuzzy C-means algorithm and a comparative analysis', *Advanced Computing, Networking and Informatics*, Vol. 1, pp.349–359, Springer International Publishing, Switzerland.

Xu, Z. and Wu, J. (2010) 'Intuitionistic fuzzy C-means clustering algorithms', *Journal of Systems Engineering and Electronics*, Vol. 21, No. 4, pp.580–590.

Zadeh, L.A. (1965) 'Fuzzy sets', *Information and Control*, Vol. 8, No. 3, pp.338–353.