AN ENHANCED FUZZY ROUGH SET-BASED CLUSTERING ALGORITHM FOR CATEGORICAL DATA

Raval Dhwani Jayant[1], Vashi Viral Mineshkumar[2]

[1]M.Tech (Software Tech.), VIT University, India
[2]M.Tech (Software Tech.), VIT University, India

Abstract

In today's world everything is done digitally and so we have lots of data raw data. This data are useful to predict future events if we proper use it. Clustering is such a technique where we put closely related data together. Furthermore we have types of data sequential, interval, categorical etc. In this paper we have shown what is the problem with clustering categorical data with rough set and who we can overcome with improvement.

----------***----------

1. INTRODUCTION

In database Categorical variable is variable which have to take one of the values from set of finite values attached to it. Value of this categorical variable relies to values of other related variables this as a whole is known as categorical data [1]. Now clustering based on what category data falls is issues here. Because unlike categorical data of blood group when we know that data will definitely fall in only one group i.e. A, B, AB e.t.c. this are crisp categorical data but we also have some categorical data which are not crisp and have may-be scenario e.g. for giving loan there will be some situation when it is tough to decide if its "risky" or "safe" and thus sometimes same set of data will fall in "risky" category where as sometimes in "safe". Another example is disease treatment when sometime we go for treatment A sometimes for B for same medical data. Clustering this kind of data to get correct cluster set we cannot apply existing rough set clustering technique as it only work with data which have at least one crisp relation using which clustering is done. With little modification in lower approximation formulation and use of fuzzy logic we can improve conventional rough set technique to better cluster the data so that all closely related data fall in same cluster.

In following section we will go through basics of rough set and fuzzy logic and then with example we will show how rough set clustering works problem in rough set clustering and Improved fuzzy rough set clustering all with proper example.

2. LITERATURE SURVEY

Rough Set Theory: - It was first proposed by Pawlak[6] for processing incomplete information in information systems.
It is defined by Pawlak[7] as Formal approximation of a set which have strong one to one relation or as called are crisp set based on its low and high set of approximation is roughs set.

As defined by Kumar and its colleague [2] in Rough cluster lower approximation contains objects that belong to unique cluster and it has no membership to other cluster and upper approximation contains objects which are in this cluster as well as in other clusters.

Rough Set uses following defined Nomenclature and formula to do clustering as defined by Darshit Parmar [9]

U universe or the set of all objects (x1, x2..)

X subset of the set of all objects, ($X \subseteq U$)

xi object belonging to the subset of the set of all objects, $xi \in X$

A the set of all attributes (features or variables) ai attribute belonging to the set of all attributes, $ai \in A$

V(ai) set of values of attribute ai (or called domain of ai)

B non-empty subset of A($B \subseteq A$)

$\underline{XB}$ lower approximation of X with respect to B

$\overline{XB}$ upper approximation of X with respect to B

Rai(X) roughness with respect to {ai}

Ind(B) indiscernibility relation

[xi]Ind(B) equivalence class of xi in relation Ind(B),

Indiscernibility relation Ind(B): - Objects, xi, $xj \in U$ are indiscernible by the attributes set B in A,that is xi, $xj \in$ Ind(B) if and only if $\forall a \in B$ where $B \subseteq A$, a(xi) = a(xj).

Equivalence class [xi]Ind(B): - Set of objects xi with respect to indiscrmibility realtion on B i.e. Ind(B) Having similar values for the set of attributes in B consists of an equivalence classes, [xi]Ind(B).

Lower approximation: - For attributes set of B in A the objects X in U, the lower approximation of X is defined as

$\underline{XB}$=U {xi|[xi]Ind(B) ⊆ X}

Upper approximation: - For attributes set of B in A the objects X in U, the upper approximation of X is defined as

$\overline{XB}$=U {xi|[xi]Ind(B) ∩ X ≠ ∅}

Roughness: - The ratio of the lower approximation and the upper approximation is called roughness it is measured as

RB(X) = 1 - |$\underline{XB}$|/|XB|

If RB(X) = 0, X is crisp or precise with B. If RB(X) < 1, X is rough or vague with B.

It is also defined as the accuracy of estimation.

Fuzzy Logic: - Fuzzy logic is multi value logic where we don't have exact value instead we have approximate value. All Fuzzy logic variables are assigned degree to which they are associated with these degree ranges 0 and 1. It deals with partial truth.[4] We then use linguistic variables to describe this fuzzy logic.

Fuzzy Rough Set: - I. Jagielska, C. Matthews, T. Whitfort,[8] stats that Knowledge, Information or data hidden ie which creates pattern in information systems can also be found by describing them data in form of some set of decision rules using application of neural networks, fuzzy logic, genetic algorithms, and rough sets and thus automating the process of information gathering.

Chris Cornelis, Martine De Cock, Anna Maria Radzikowska[5] has given Fuzzy Rough set also known as fuzzy hybridization of Rough Set where we use rough set to find dependency which help to cluster the universe where as fuzzy logic is use to represent the pattern in data through linguistic variables. This combination of fuzzy with rough is also called as soft computing or hybrid intelligent system.

3. WORKING OF ROUGH SET

In table 1 we have 2 attribute A1 and A2 and 10 objects now to show rough set clustering works we take here A1 as base attribute to cluster 10 objects set

	A1	A2
1	Big	Blue
2	Medium	Red
3	Small	Yellow
4	Medium	Blue
5	Small	Yellow
6	Big	Green
7	Small	Yellow
8	Small	Yellow
9	Big	Green
10	Medium	Green

Steps

1. **Find elementary set of A1 and A2**
   X(A1=Small) = {3,5,7,8}
   X(A1=Medium)={2,4,10}
   X(A1=Big)={1,6,9}
   X(A2=Blue)={1,4}
   X(A2=Red)={2}
   X(A2=Yellow)={3,5,7,8}
   X(A2=Green)={6,9,10}

2. **Find Lower and Upper Approximation of A1 sets with Respect to A2 sets**
   Lower Approximation of X(A1=Small)={3,5,7,8}
   Upper Approximation of X(A1=Small) ={3,5,7,8}
   Lower Approximation of X(A1=Medium) ={2}
   Upper Approximation of X(A1=Medium) ={1,2,4,6,9,10}
   Lower Approximation of X(A1=Big) ={No Lower Approximation}
   Upper Approximation of X(A1=Big) ={As No Lower Thus No Upper Approximation}

3. **Find Roughness of each elementary set of A1**
   Roughness X(A1=Small)=1-(4/4)=0
   Roughness X(A1=Medium)=1 (1/6)=0.833
   Roughness X(A1=Big)=N/A

4. **Cluster objects based on elementary set**
   As X(A1=Small) have lower Roughness we will split into two Cluster one with Small and Other with Big, Medium

4. PROBLEM WITH ROUGH SET

With traditional rough set when we have no crisp relation in data we cannot do clustering let us take table 2 as example

	A1	A2
1	Big	Blue
2	Medium	Pink
3	Small	Red
4	Medium	Orange
5	Small	Blue
6	Big	Red
7	Small	Yellow
8	Small	Pink
9	Big	Orange
10	Medium	Yellow

X(A1=Small) = {3,5,7,8}
X(A1=Medium)={2,4,10}
X(A1=Big)={1,6,9}
X(A2=Blue)={1,5}
X(A2=Pink)={8,2}

X(A2=Orange)={9,4}
X(A2=Yellow)={10,7}
X(A2=Red)={3,6}

Lower Approximation of X(A1=Small)={ No Lower Approximation }

Upper Approximation of X(A1=Small) ={ As No Lower Thus No Upper Approximation }

Lower Approximation of X(A1=Medium) ={ No Lower Approximation }

Upper Approximation of X(A1=Medium) ={ As No Lower Thus No Upper Approximation }

Lower Approximation of X(A1=Big) ={No Lower Approximation}

Upper Approximation of X(A1=Big) ={As No Lower Thus No Upper Approximation}

As we don't have lower and upper approximation for any X we cannot find Roughness

5. PROPOSED WORK

In our proposed clustering method we have new lower approximation formulation through which we will find lower approximation with this we will use fuzziness degree of each attribute value to decide which set will be taken as lower approximation if we don't have crisp set and more than one partial set which cannot be distinguished by given lower approximation formula. Once we have clustered data we can write Fuzzy linguistic rule base for future reference. This type of clustering helps to deal with problem in categorical data which we have mentioned in problem statement.

	A1	A2
1	Big	Blue
2	Medium	Pink
3	Small	Red
4	Medium	Orange
5	Small	Blue
6	Big	Red
7	Small	Yellow
8	Small	Pink
9	Big	Orange
10	Medium	Yellow

Steps

1. **Assign Weight to each Attribute value of A1 and A2**
   Weight
   Small=0.1
   Medium=0.2
   Big=0.3
   Blue=0.7
   Pink=0.6
   Red=0.8
   Yellow=0.4
   Orange=0.5

2. **Find Elementary set of A1 and A2**
   X(A1=Small) = {3,5,7,8}
   X(A1=Medium)={2,4,10}
   X(A1=Big)={1,6,9}
   X(A2=Blue)={1,5}
   X(A2=Pink)={8,2}
   X(A2=Orange)={9,4}
   X(A2=Yellow)={10,7}
   X(A2=Red)={3,6}

3. **Find Lower and Upper Approximation for Set of A1 with Respect to A2 using near match fomulation**
   To find lower approximation we find set nearest to base set X with formulation (No. of Match with base set/total no. of element in set) for all set which are having subset of base set we find this. Out of this we will take one with minimum value.

   Lower Approximation of X(A1=Small)={3,6}=1/2=0.50

   {1,5}=1/2=0.50
   {10,7}=1/2=0.50
   {8,2}=1/2=0.50

   Now here as all set have same value we will look in fuzzy weight allocated and select one with higher rate.

   Here we have Red having higher weight than other attribute value and thus we select {3,6} as lower approximation

   Upper Approximation of X(A1=Small) ={3,5,7,8,1,6,10,2}

   Lower Approximation of X(A1=Medium) ={8,2}

   Upper Approximation of X(A1=Medium) ={2,4,10,8,9,7}

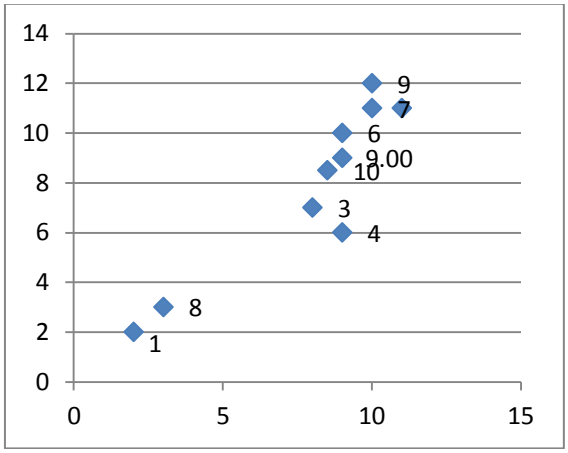
   Lower Approximation of X(A1=Big) ={3,6}

   Upper Approximation of X(A1=Big) ={1,6,9,4,3,5}

4. **Find Roughness**
   Roughness X(A1=Small)=1-(2/8)=0.75
   Roughness X(A1=Medium)=1-(2/6)=0.66
   Roughness X(A1=Big)=1-(2/6)=0.66

5. **Cluster Data based on roughness**
   As X(A1=Medium)and X(A1=Big) have similar lower Roughness we will split into two Cluster

based on weight as X(A1=Big) have higher weight we will split into one with Big and Other with Small,Medium



6. Fuzzy Linguistic Rule Base

Where Color is A2=Red A1=Big

6. CONCLUSION

Rough set can only be applied where we have at least one crisp relation of the data. Thus in this paper we proposed clustering technique with mixture of rough set with improved lower approximation formulation and fuzzy logic which works on any kind of categorical data whether data is having crisp relationship between its attribute or not we can do clustering. As shown with example proposed technique works better than traditional technique.

REFERENCES

[1]. http://en.wikipedia.org/wiki/Categorical_variable

[2]. Pradeep Kumar, P. Radha Krishna, Raju. S. Bapi, Supriya Kumar D "Rough clustering of sequential data"

[3]. Duo Chen, Du-Wu Cui, Chao-Xue Wang, Zhu-Rong Wang "A Rough Set-Based Hierarchical Clustering Algorithm for Categorical Data" School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, 710048, China

[4]. http://en.wikipedia.org/wiki/Fuzzy_logic

[5]. Chris Cornelis, Martine De Cock, Anna Maria Radzikowska, "Fuzzy Rough Sets: from Theory into Practice" Computational Web Intelligence Dept. of Applied Mathematics and Computer Science Ghent University, Krijgslaan 281 (S9), 9000 Gent, Belgium

[6]. Z. Pawlak, "Rough sets", International Journal of Computer and Information Science (1982).

[7]. Z. Pawlak, "Rough Sets—Theoretical Aspects of Reasoning about Data", Kluwer Aca. Pub. (1991)

[8]. I. Jagielska, C. Matthews, T. Whitfort, "An investigation into the application of neural networks, fuzzy logic, genetic algorithms, and rough sets to automated knowledge acquisition for classification problems", Neurocomputing (1999)

[9]. Darshit Parmar, Teresa Wu *, Jennifer Blackhurst, "MMR: An algorithm for clustering categorical data using Rough Set Theory"