JCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Vol-6, Special Issue-4, May 2018 E-ISSN: 2347-2693

Comparative Analysis of Roughness with Maximum Dependency Attribute

M.Jancy Rani[1*], A. Pethalakshmi[2]

[1*,2] Department of Computer Science, M.V.Muthiah Govt.Arts College for Women, Dindigul, Tamil Nadu, India.

[*] *Corresponding Author:* jancymca1623@gmail.com, *Tel.: +00-12345-54321*

Available online at: www.ijcseonline.org

***Abstract*—** Rough set theory is a powerful mathematical tool that has been applied widely to extract knowledge from many databases. It deals with inexact and incomplete data. Cluster analysis means finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters that are meaningful, useful, or both. Data clustering techniques are valuable tools for researchers working with large databases of multivariate data. Cluster analysis is used in various applications viz., Pattern Recognition, Data Analysis, Image Processing and so on. This paper analyses Roughness and Maximum Dependency Attribute clustering algorithms that minimizes the need for subjective human intervention and compare the purity analysis between these two methods. Purity analysis percentage is calculated from the result of final clusters. Six datasets are used in this research work for comparing the roughness and maximum dependency attribute algorithm to describe the cluster solution by using the purity analysis (PA).

***Keywords*—** Rough Clustering, Equivalence Classes, Roughness, Maximum Dependency Attribute, Purity Analysis.

I. Introduction

A. Data Mining

Data mining refers to extracting knowledge from large amounts of data. Data mining is often treated as synonym for another popularly used term, Knowledge Discovery in Databases (KDD) [4]. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

B. Rough Set Theory (RST)

In 1982, Pawlak introduced the theory of Rough sets. A Rough Set is a mathematical approach to deal with uncertainty and vagueness of an information system [8]. An information system can be presented as a table with rows analogous to objects and columns analogous to attributes. Each row of the table contains values of particular attributes representing information about an object. It can be used for clustering, feature selection, feature extraction, data reduction, decision rule generation, and pattern extraction (templates, association rules).

Over the years, RST has become a topic of great interest to researchers and has been applied to many domains, in particular to knowledge databases. This success is due in part to the following aspects of the theory:

- Only the facts hidden in data are analyzed,
- No additional information about the data is required,
- Minimal knowledge representation is obtained.

The paper is organized as follows: Section 2 deals with the related works. Section 3 describes about Rough Clustering algorithms. Section 4 discusses the results and discussion. Section 5 analysis the Comparison between Roughness and Maximum Dependency Attributes (MDA) Algorithm and Section 6 concludes the paper.

II. Related Works

Data clustering is a thrust area of research for statistics, biology, machine learning as well as data mining researchers which resulted in the development of a huge variety of successful clustering algorithms. There are several researches and implementations have been done on Rough Clustering model. In this paper, we reviewed some of the research papers on clustering algorithm and mainly a Rough Clustering model.

Pethalakshmi et al, proposed A Novel Approach in Clustering via Rough Set. They enhanced a vision of UCAM algorithm is represented by hybridizing with RSAR, which reduces attributes in the data set without affecting its originality. The hybridization of UCAM with RSAR helps to reduced computational complexity, processing time and to increase the cluster uniqueness. This approach also reduced the overheads of fixing the cluster size and initial seeds as in K-Means. It fixes threshold value to obtain a unique clustering. The proposed method improves the scalability and reduces the clustering error. This approach ensures that the total mechanism of clustering is in time without loss in correctness of clusters [10].

Pethalakshmi et al, described Hybridization of Rough set with K-Means and Fuzzy C-Means through affinity measure for unique clustering. Fuzzy C-Means provided membership degree and rough set provided attribute reduction [9].

PawanLingras et al, presented their results from conventional and rough Set based K-Means algorithm. Web visitors for three courses were used in the experiments. It was expected that, the visitors would be classified as studious, crammers, or workers. Since some of the visitors might not precisely belong to one of the classes, the clusters were represented using interval sets. The experiments produced meaningful clustering of web visitors using both the conventional and rough set based approach [7].

Jun-Hao Zhang et al, proposed the implementation of Rough-Fuzzy K-Means clustering algorithm used in MATLAB. The assistance of the lower and upper approximation of rough sets, the Rough Fuzzy K-Means clustering algorithm might improve the objective function and further the distribution of membership function for the traditional Fuzzy K-Means clustering. However, the algorithm only had theoretical ideas rather than concrete realizations. To make it better applied to practice, using Matlab, a mathematical programming tool to implement rough Fuzzy K-Means clustering algorithm was discussed. Moreover, steps of implementation were given in detail. The foresaid contributions might provide clustering learners and non-computer professional researchers with a simple, convenient, efficient and feasible implementation method [5].

Manish Joshi et al, described three different non-crisp clustering algorithms namely Fuzzy C-means (FCM), Rough K-means (RKM) and Interval set K-means (IKM) algorithms. The principles of interval set clustering were discussed and the processes of blending the conventional K-means algorithm to adapt the interval set concepts were also described. These algorithms were on a real life data set (Web usage data). They used an Intra Cluster Variation (ICV), an unambiguity measure, and execution time as measures for comparing the three algorithms [6].

Tien-Chin Wang et al, described the Fuzzy Set Theory (FST) and Rough Set Theory (RST) used to extract Decision Rules from a Decision table [11].

Chandranath Adak, proposed an unsupervised technique to detect the changed region of multi-temporal images on a same reference plane with the help of rough clustering. The proposed technique is a soft-computing approach, based on the concept of rough set with rough clustering and Pawlak's accuracy. It is less noisy and avoids pre-deterministic knowledge about the distribution of the changed and unchanged regions. To show the effectiveness, the proposed technique was compared with some other approaches [3].

III. Methodology

Overview of Rough Clustering Algorithms:
In Rough Clustering, two algorithms are performing to find out the Clustering values. They are Roughness algorithm and Maximum Dependency Attribute algorithm.

A. Roughness Algorithm

In Roughness Algorithm first we find out equivalence classes [1]. It is also known as indiscernibility Relation. The Roughness algorithm is shown in Figure 1.

Calculate Roughness

```
1. Load the Dataset
2. Calculate Roughness.
    2.1 Find Lower and Upper Approximations.
    2.2 Roughness = 1 – (Lower Approximation / Upper Approximation)
3. Draw Conditional Attribute Table.
4. Final Clustering Table.
```

Algorithm: Roughness

```
Input: Dataset without clustering attribute
Output: Clustering attribute
Begin
Step 1. Compute the equivalence classes using the indiscernibility relation on each attribute.
Step 2. Calculate Roughness using the formula:
Roughness = 1- (Lower Approximation / Upper Approximation)
Step 3. Select a clustering attribute based on the maximum roughness of attributes.
Step 4. Final Clustering Table.
End
```

Figure 1: The Roughness Algorithm

B. Maximum Dependency Attributes (MDA)

MDA is the second algorithm of Rough Clustering. The MDA algorithm is shown in Figure 2.

Calculate Maximum Dependency Attribute

```
1. Load the Dataset.
2. Calculate Maximum Dependency Attribute.
    Calculate $D = \Sigma U|CA|B(X)/|U|$
3. Draw Conditional Attribute Table.
4. Final Cluster Table.
```

Here,
U – Universal Set
CA – Conditional Attribute

Algorithm: MDA

```
Input: Dataset without clustering attribute
```

```
Output: Clustering attribute
Begin
Step 1. Compute the equivalence classes using the
        indiscernibility relation on each attribute.
Step 2. Determine the dependency degree of attribute ai with
        respect to all aj, where i ≠ j.
Step 3. Calculate Maximum Dependency Attribute using the
        formula: D = ΣU|CA|B(X)/|U|
Step 4. Select the maximum of dependency degree of each
        attribute.
Step 5. Select a clustering attribute based on the maximum
        degree of dependency attributes.
End
```

Figure 2: The MDA Algorithm

III. Purity Analysis (PA) for Roughness and MDA Algorithm

Purity analysis percentage is calculated from the result of final clusters. The below formula is used to find out the purity analysis.

$$\text{Purity Analysis} = \frac{\text{Number of final clusters}}{\text{Number of objects}}$$

IV. Results and Discussion

The Rough Clustering Algorithms are developed for both numerical and categorical attributes. Here in this section, the sample patient data set is taken and applied for both Rough Clustering Algorithms. The sample patient data set is tabulated in table 1.

Table 1: Patient (Dengue) Data Set

Patient	Muscle Pain	Headache	Temperature
1	Yes	Yes	High
2	Yes	No	High
3	No	Yes	Very High
4	No	Yes	Normal
5	No	Yes	Normal
6	No	Yes	High
7	No	No	Normal

Patient Data set Description

The Patient Data set contains three Attributes. All the attributes are categorical. All the attributes are considered as input attributes. The Patient Data set provides the details about the patient (Muscle Pain, Headache, and Temperature) and which could be useful to cluster the patients according to their present details.

A. Roughness Algorithm

Iteration 1

i) Equivalence Classes

X (Muscle Pain) = {{1,2}, {3,4,5,6,7}}

X (Headache) = {{1,3,4,5,6}, {2,7}}

X (Temperature) = {{1,2,6}, {3}, {4,5,7}}

ii) Obtain Lower and Upper Approximation

Output: Muscle Pain

a) $X(\text{Headache=Muscle Pain})_{*}=\{\phi\}$

$X(\text{Headache=Muscle Pain})^{*}=\{1,2,3,4,5,6,7\}$

b) $X(\text{Temperature=Muscle Pain})_{*}=\{3,4,5,7\}$

$X(\text{Temperature=Muscle Pain})^{*}=\{1,2,6\}$

iii) Roughness

Roughness=1 – X (B_{*}) / X (B^{*})

a) X(Headache=Muscle Pain)=1 - 0 / 7 = 1

b) X(Temperature=Muscle Pain)=1 - 4 / 3 = 0.33

iv) Resultant Table

Decision Attribute: Muscle Pain

Attributes	Crisp	Rough	Class
Headache	1	-	C1
Temperature	-	0.33	C2

Iteration 2

Output: Headache

Resultant Table

Decision Attribute: Headache

Attributes	Crisp	Rough	Class
Muscle Pain	1	-	C1
Temperature	-	0.83	C2

Iteration 3

Output: Temperature

Resultant Table

Decision Attribute: Temperature

Attributes	Crisp	Rough	Class
Muscle Pain	-	0.6	C2
Headache	1	**-**	C1

Final Result of Roughness (Clustering Table)

Conditional Attributes	**Muscle Pain**	**Head Ache**	**Tempe rature**
Muscle Pain	-	C1	C2
Headache	C1	-	C2
Temperature	C2	C1	-

From the above final result table, the dominating attribute is Headache. Therefore Headache of (C1) = 2

The final clusters of the attribute headache is: {1,3,4,5,6}, {2,7}

B. *Maximum Dependency Attributes (MDA)*

Iteration 1
Output: Muscle Pain

a) $$K = \frac{|\left(\frac{\text{Headache}}{\text{Muscle Pain}}\right)*|}{|u|} = \frac{0}{7} = 0$$

b) $$K = \frac{|\left(\frac{\text{Temperature}}{\text{Muscle Pain}}\right)*|}{|u|} = \frac{4}{7} = 0.57$$

Iteration 2
Output: Headache

a) $$K = \frac{|\left(\frac{\text{Muscle Pain}}{\text{Headache}}\right)*|}{|u|} = \frac{0}{7} = 0$$

b) $$K = \frac{|\left(\frac{\text{Temperature}}{\text{Headache}}\right)*|}{|u|} = \frac{1}{7} = 0.14$$

Iteration 3
Output: Temperature

a) $$K = \frac{|\left(\frac{\text{Muscle Pain}}{\text{Temperature}}\right)*|}{|u|} = \frac{2}{7} = 0.28$$

b) $$K = \frac{|\left(\frac{\text{Headache}}{\text{Temperature}}\right)*|}{|u|} = \frac{0}{7} = 0$$

Final Result of MDA (Clustering Table)

Conditional Attributes	**Muscle Pain**	**Headache**	**Temperature**
Muscle Pain	-	C1	C2
Headache	C1	-	C2
Temperature	C2	C1	-

Here 0's are C1 and others are C2. From the above final result table, the dominating attribute is Temperature. Therefore Temperature of (C2) = 2

The Final Clusters of the attribute Temperature is: {{1,2,6}, {3}, {4,5,7}}

C. *Purity Analysis (PA) for Roughness and MDA Algorithm*

For Roughness Algorithm

Purity Analysis = *100 = 28.57

For MDA Algorithm

Purity Analysis = * 100 = 42.85

V. Comparision between Roughness and Maximum Dependancy Attributes (MDA) Algorithm

The different datasets are collected from UCI machine learning repository [2] and applied two types of clustering algorithms and also calculate purity analysis. The resultant table is presented in Table 2. Among the two algorithms, Maximum Dependency Attribute algorithm provides maximum purity analysis than roughness algorithm.

Table 2: Purity Analysis of Roughness with MDA

S. No	Data Set	Roughness Purity Analysis(PA) in Percentage	MDA Purity Analysis(PA) in Percentage
1	Supermarket	29%	43%
2	Mushroom	13%	20%
3	Heart Disease	40%	60%
4	Credit Card Promotion	20%	30%
5	Diabetics	20%	30%
6	Dengue	29%	43%

VI. Conclusion

In this paper, we designed and analyzed two algorithms of Rough Clustering that was generalized to read any categorical as well as numerical data set with any number of attributes. The Rough Clustering algorithms were tested on different data sets from the UCI machine learning repository. This paper was also presented the comparison of roughness and maximum dependency attribute to describe the cluster solution by using the purity analysis (PA). We observed that MDA algorithm outperformed than Roughness algorithm for all the data sets.

References

[1] Arun K Pujari, "Data Mining Techniques", Universities Press (India) Private Limited, 2010.

[2] C. L. Blake and C.J. Merz, "UCI Repository of Machine Learning Databases", Irvin, University of California, http://www.ics.uci.eduction."/~mlearn/, 1998.

[3] Chandranath Adak, "Rough Clustering Based Unsupervised Image Change Detection".

[4] J.Han and M.Kamber, "Datamining: Concepts and Techniques", Morgan Kaufmann Publishers, 1992.

[5] Jun-Hao Zhang, Ming-Hu Ha, Jing Wu, "Implementation of Rough K-Means Clustering Algorithm in MATLAB", 9th International Conference on Machine Learning and Cybernetics, pp.2084-2087, July 2010.

[6] Manish Joshi, Yiyu Yao, Pawan Lingras, Virendrakumar C.Bhavsar, "Rough, Fuzzy, Interval Clustering For Web Usage Mining", 10th International Conference on Intelligent Systems Design and Applications, pp.397- 402, 2010.

[7] Pawan Lingras, Rui Yan, Chad West, "Comparison of Conventional And Rough K-Means Clustering", LNAI 2639, Springer – Verlag Berlin Heidelberg,pp.130-137,2003.

[8] Z. Pawlak, "Rough sets", International Journal of Computer and Information Sciences, vol. 11, pp. 341–356, 1982.

[9] Pethalakshmi.A, A.Banumathi, "Refinement of K-Means And Fuzzy C-Means", International Journal of Computer Applications, Volume 39, Paper Number: 17, Feb 2012.

[10] Pethalakshmi.A, A.Banumathi, "A Novel Approach in Clustering via Rough set", IJSR, vol. 2, Issue 7, July 2013.

[11] Tien-Chin Wang, Lisa Y.Chen, Hsien-Da Lee, "Fuzzy Entropy-Based Rough Set Approach for Extracting Decision Rules", pp.5636-5639, IEEE 2007.