Rough Set -Algorithm for Clustering Categorical Data Using Mean Attribute (MMA) Dependency Based Measure

[Muftah Mohamed Baroud, Siti Zaiton Mohd Hashim, Siti Mariyam Shamsuddin, Anazida Zainul]

***Abstract*—Different cluster techniques based on the Rough Set Theory (RST) have been used for attribute selection and grouping objects displaying similar characteristics. On the other hand, a majority of these clustering techniques cannot tackle uncertainty. Furthermore, these processes are computationally complicated and less accurate. In this study, the researchers have explored the limitations of the two rough set theory based techniques, i.e., the Maximum Dependency Attribute (MDA) and the Maximum Indiscernible Attribute (MIA). They also proposed a novel approach for selecting the clustering attributes, i.e., the Maximum Mean Attribute (MMA). They compared the performances of the MMA, MDA and the MIA techniques, using the UCI dataset. Their results validated the performance of the MMA with regards to its accuracy and computational complexity.**

***Keywords*— clustering, rough set theory, categorical data, dependency of attribute, performance.**

I. Introduction

Cluster analysis refers to a novel data mining tool which can be applied for grouping the data having similar characteristics (1, 2, and 6). Many techniques have been proposed for clustering the categorical data (2, 3, and 10). For the categorical data, all data objects consist of multi-valued attributes, which differ from the numerical data. Therefore, the RST-based clustering processes that select the attributes from the categorical data are very popular. These processes are able to select the best clustering attribute amongst the various attributes. The Maximum Dependency Attributes (MDA) approach was proposed by Herawan et al. (3), for selecting the clustering attributes. This process was based on the RST and considered the attribute dependency on a database Also, the Maximum Indiscernible Attribute (MIA) process was proposed by Uddin et al. (1) for selecting the clustering attributes this process was based on the indiscernibility relation concept, which combined several clusters and selected the clustering attribute having the maximal indiscernibility (12). The MIA process performed better than the MDA technique. The MDA and the MIA values were determined using the traditional rough sets that were based on an equivalence class structure that was considered a computationally complex and expensive process (4). This makes it difficult to select the most appropriate clustering attribute that shows a lower performance using a few of the datasets. Thus, according to some researchers (1, 3, 9), the best process which uses categorical data for estimating the clustering attributes, with a lower computational complexity and a higher accuracy, has still not been proposed. Therefore, a novel process for clustering the categorical data needs to be determined. In this study, the researchers proposed a novel Mean Maximum Attribute (MMA), technique for selecting the clustering attributes by considering the mean dependency values for the attribute. They also defined the concept of a maximum mean attribute for assessing the performance of the clustering attribute selection process. Their experimental results indicated that the proposed technique showed a higher accuracy for clustering attribute selection, in comparison to the MDA and MIA processes. The remaining paper has been organized in the following manner: Section II described the RST used in the information systems; Section III analysed the two rough set theory based MDA and MIA techniques. Section IV describes the limitations of using the RST based techniques, while Section V describes the MMA methodology and presents an illustrative case study. Section VI compares the MMA, MDA and the MIA results. Finally, Section VII presents the conclusions of the study.

II. (Rough Set Theory (RST)).

Pawlak introduced the RST (5) in the 1980s. This technique handles uncertainty and identifies the cause-effect relationship in the databases for database learning and data mining. This approach improves the data clustering and ensures uncertainty management in the relational databases (12). The information system is a convenient technique, which represents the objects based on their attribute values. The information system has been described earlier (4). It consists of a 4-tuple (quadruple), , wherein U refers to a nonempty finite set consisting of various objects; A denotes a non-empty finite set of attributes, a; while, while *Va* represents the domain (or value set) of the attributes; denotes the total function so that, for each , known as the information functions. Based on this information system, some definitions commonly used in the RST are as follows:

Definition 2: Let be the information system; B is any subset of A; while *X* is a subset of U. Also, the B-lower approximation for *X*, which is represented by *B(X)* and the B-upper approximation of *X*,

Muftah Mohamed Baroud, Siti Zaiton Mohd Hashim, Siti Mariyam Shamsuddin, Anazida Zainul

School of Computing N28, UTM Faculty of engineering.
Skudai, Johor Bahru, Malaysia.

which is described by *B(X)*, respectively, can be defined as:

$$\underline{B}(X)=\{x\in U \mid [X]_B\subseteq X\} \& \overline{B}(X)=\{x\in U [x]_B\cap X\neq\phi\} \quad (1).$$

Furthermore, the approximation accuracy for any, subset $X\subseteq U$ with respect to $B\subseteq A$, which is represented as

$\alpha_B(x)$ can be estimated by

$$\alpha_B(X)=\frac{\underline{B}(X)}{\overline{B}(X)} \quad (2).$$

In this study, |*X*| represents the cardinality of *X*. Also, the approximation accuracy is interpreted by using the popular MarcZeweski-Steinhaus (MZ) metric [10]. The researchers applied the MZ metric for the upper and lower approximations for the subset $X\subseteq U$ in an information system *S*, and obtained:

$$D(\underline{B}(X),\overline{B}(X))=1-\frac{\left|\underline{B}(X)\cap\overline{B}(X)\right|}{\left|\overline{B}(X)\cup\underline{B}(X)\right|}=\frac{\left|\underline{B}(X)\right|}{\left|\overline{B}(X)\right|}=1-\alpha_B(X) \quad (3).$$

Definition 3: Let *S = (U, A, V, f)* be the information system; *G* and *H* represented the subsets of *A*. Also, *G* is seen to be dependent on *H* in the degree *k*, which is described as $H\Rightarrow_k G$. The degree, k, is defined as:

$$K=\frac{\sum_{X\in U/D}\underline{H}(x)}{|U|} \quad (4).$$

Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities.

For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. Analysis of the RST-Based Techniques.

This section includes the MDA technique and the MIA processes.

A. MDA Technique (Herawan et al.)

Consider any attribute *ai, aj, ka (ai)* where in *ai* is seen to depend on the *aj* in the degree k, and can be obtained using the following equation:

$$Kaj(ai)=\frac{\sum_{X\in U/q}\left|aj(x)\right.}{|U|} \quad (5).$$

Thereafter, the Max-Dependency (MD) of the attribute a_i ($a_i\in A$) is determined as the $MD\ (a_i) = Max\ (ka_1\ (a_i)),.k_{ai}\ (a_i\ ),.., k\ a_m\ (a_i))$ after the values of *m* from $MD\ (a_i)$, $i = 1, 2,..., m$ are obtained. The MDA process can select the clustering attribute having the maximal MD values as follows:

$$MDA = Max\ (MD\ (a_1),..., MD\ (a_i),..., MD\ (a_m)).$$

B. MIA Technique (Uddin et al.).

Consider T as a subset of *A*, wherein the 2 elements $x, y\in U$ are T indiscernible (by a set of attributes, $T\subseteq A$ in s) if $\delta(x,t)=\delta(y,t)$ for each $t\in T$. Hence, each subset *T* of *A* can induce an equivalence indiscernibility relationship and a specific clustering process as indicated by the *IND(T)*. The *U* clustering is induced by the *IND (T)* in *S* and is presented as *U/*T. The equivalence class in the cluster *U/T* contains $x\in U$, which is indicated by *T[x]*. Furthermore, the cardinality of the indiscernibility relationship of the the attribute(s) indicates the number of the clusters that can be obtained using that attribute and are determined as:

$$card(IND(T)=\left|IND(T)\right| \quad (6).$$

IV. Limitations of the Rough Set-Based Process.

In an earlier report (8), the researchers used a test study to compare the MDA and MIA processes. Their results showed that these techniques could be determined using a traditional rough set attribute. A similar clustering process, based on the partition classes, is used for clustering the objects. This helps in selecting few attributes, which indicates the lower performance of this process. This is described using an example below:

Example 1 (1): Table 1 presents Pawlak's car performance dataset. This case study consists of 6 cars (m = 6) with 3 (n = 3) conditional attributes, which are: a=Terrain familiarity, b=Gasoline level, c=distance.

TABLE I. Pawlak’s Car Performance Dataset (1)

U	a	b	c	d
1	Poor	Low	Short	<30
2	Poor	Low	Short	<30
3	Good	Low	Medium	<30
4	Good	Medium	Short	30 … 50
5	Poor	Low	Short	<30
6	Poor	High	Long	>50

TABLE II. The degree of dependency of all the attributes described in Table 1 using the MDA process.

Attribute (depends on)	**Degree of dependency**		**MDA**	**Second MDA**
a	b	c	0.3	
	0.3	0.3		
b	a	c		
	0.3	0.3	0.3	
c	a	b		
	0	0.3	0.3	0.3

TABLE III. The degree of dependency of all the attributes described in Table 1 using the MDA process.

Attributes	**Indiscernibility Relation Cardinality**	**MIA**	**Second MIA**
a	2		
b	3	3	
c	3	3	
B + c	4	3+3	4

Owing to a higher resemblance between the MDA and the MIA processes, it is difficult to determine the best clustering attributes. It is also difficult to select the similar maximal degree values, simultaneously. Here, the researchers analysed the efficiency of the MMA technique to select the best dataset. They have also explored the issues faced by the different processes while selecting the best clustering attributes.

V. *Maximum Mean Attribute (MMA) Process.*

For overcoming all limitations of the RST-based processes, the researchers have proposed a novel RST-based MMA technique: Definition 5.1: The RST strategy was proposed for handling the uncertainty and the vagueness in the data. In this theory, the information system has a format of $I = (U, K, V, \varepsilon)$ Where in;

$U = \{u1, u2, u3, ..., u|u|\}$: A non-empty finite set of some objects

$K = \{k1, k2, k3, ..., k, |k|\}$: A non-empty finite set of attributes

$V = V_{k \in k} V_k, V_a$: The domain or the value set of the attribute a, $\varepsilon = U \times K \rightarrow V$: An information function which was $\varepsilon(u,k) \in V_k$ and was used for estimating the dependency.

$$r = \delta(S,T) = \frac{\left|Pos_S(T)\right|}{|U|} = \frac{\sum_{i=0}^{m}\left|S_i\right|}{|U|} \qquad (7).$$

The mean dependency value for all attributes ($K - 1$) with regards to the target attribute was determined. If $r = 1$, "*T*" is seen to be completely dependent on the S; for 0 < r < 1, "*T*" partially depends on "S"; and for r = 0, "T" is independent of "*S*". It can be seen that for r = 1, "*T*" completely depends on "S", hence, *IND(S)* ⊆ *IND(T)*. In other words, *U/S* was finer than the *U/T* and was estimated as follows:

$$Mr = \frac{\sum\left|r\left|k-1\right|\right|}{\left|k-1\right|} \qquad (8).$$

Thus, the maximal mean dependency value in the set of attributes (K) is calculated as follows:

$$MMr = \max\left|mr\right|k \qquad (9).$$

As seen from all the above-mentioned equations, the MMA measures the mean dependency between the attributes *S* and *T* and the dependency between *U/S* and *U/T*. The higher the mean dependency value, the greater is the crispness clustering value. Based on all these definitions, the researchers have proposed the Maximum Mean Attribute (MMA) algorithm, described in Figure 1. This algorithm consists of 5 steps: Step 1 includes the calculation of the Target attribute (*T*). Also, the equivalence class structures for every attribute is based on the Specific attribute (*S*) and is indicated by *S*1, *S*2 …, *S*m. Step 2 handles the positive region class specifics, which are estimated using Eq. “7”. Step 3 determines the average dependency value of all the attributes (K − 1) with regards to the target attribute, using Eq. “8”. Step 4 determined the maximal mean dependency value for all the attributes using Eq. “9”. Finally, Step 5 determined all the attributes, and a clustering algorithm was selected based on the maximal maximum mean dependency.

Algorithm MMA

Input: Dataset without clustering attributes

Output: Clustering attribute

1. Compute the equivalence classes using the indiscernibility relation on each attribute
2. Determine the degree of dependency of attribute T on attributes S.
3. Find the mean of (K-1) attributes with respect to target attribute (T).
4. Calculate the maximum mean of (K-1) attributes with respect to target attribute (T)
5. Select the clustering attribute based on maximum dependency.

Fig. 1. Algorithm code for the MMA process

Example 2: In this case study, 4 categorical attributes (n = 4): were considered, i.e., A, B, C and D. These attributes had 2 specific values (1=3), i.e., low, high, small, bad loss, good, large and large and 5 objects (m = 5).

TABLE IV. Modified information system (2).

U	A	B	C	D
1	Low	Bad	Loss	Small
2	Low	Good	Loss	Large
3	High	Good	Loss	Medium
4	High	Good	Loss	Medium
5	Low	Good	Profit	Large

Table 5 presents the results of the MMA process. Thus, as per this technique, the Attribute D showed the best maximum mean value of 0.866. It can be concluded that the MMA is an inexpensive and less computationally complex technique that can successfully select the best clustering attribute in comparison to the MDA and the MUA processes.

TABLE V. The mean maximum degree of all the attributes described in Table 4, using the MMA process.

Attributes	Mean Maximum Dependency			MMA
A	B	C	D	
	0.4	0.4	0.4	0.4
B	A	C	D	
	0.2	0.2	0.2	0.2
C	A	B	D	
	0.2	0.2	0.2	0.2
D	A	B	C	
	1	1	0.6	0.866

VI. Comparison of the Different Processes.

The MDA and the MIA processes used different techniques for selecting the clustering attributes. For evaluating the dataset accuracy, Pawlak described two numerical measures for determining the uncertainties in a dataset, i.e., accuracy and the roughness of the dataset (11). The results of a zoo dataset have been described in Figure 2. The experimental results obtained using the Zoo dataset highlighted the performance of the MMA technique in comparison to the existent MDA and MIA techniques, which used the numerical measurement accuracy for determining the uncertainty within the rough set in the information systems (11).

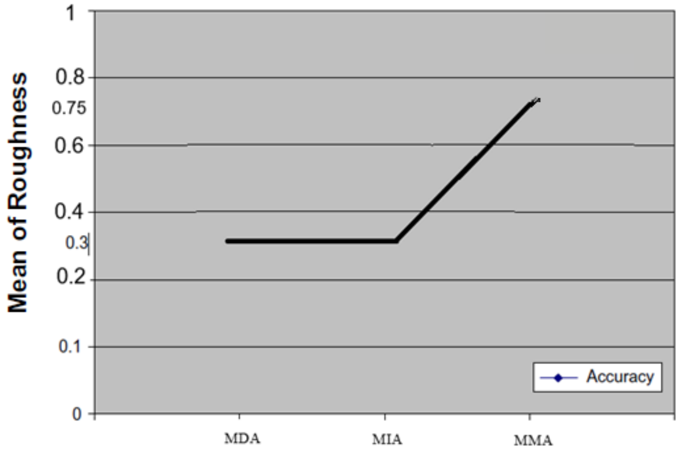

Fig. 2. A comparison of the accuracies of the MDA, MIA and MMA processes using the Zoo dataset (Parmar 3).

VII. Conclusion.

In this study, the researchers proposed a new RST-based process for clustering the categorical data and handling the uncertainty. The currently existing processes like the Maximum Dependency Attribute (MDA) and the Maximum Indiscernibility Attribute (MIA), used for selecting the clustering attributes have a few limitations, like computational complexity and lower accuracy. The researchers analysed the limitations of the 2 techniques and proposed the Maximum Mean Attribute (MMA) technique, which was based on the average dependency of the attributes. All experimental results indicated that this technique could overcome the limitations of the earlier 2 techniques. Furthermore, when a few experiments were carried out using the UCI dataset; the researchers noted that the MMA technique showed a higher accuracy and a lower computational complexity for selecting the best clustering attributes. This proposed technique could cluster the different benchmarked categorical data.

References

[1] J. Uddin, R. Ghazali, and M. M. Deris, “An Empirical Analysis of Rough Set Categorical Clustering Techniques,” PLoS One, vol. 12, no. 1, p. e0164803, 2017.

[2] D. Parmar, T. Wu, and J. Blackhurst, “MMR: An algorithm for clustering categorical data using Rough Set Theory,” Data Knowl. Eng., vol. 63, no. 3, pp. 879–893, 2007.

[3] T. Herawan, M. M. Deris, and J. H. Abawajy, “A rough set approach for selecting clustering attribute,” Knowledge-Based Syst., vol. 23, no. 3, pp. 220–231, 2010.

[4] M. S. Raza and U. Qamar, “An incremental dependency calculation technique for feature selection using rough sets,” Inf. Sci. (Ny)., vol. 343–344, pp. 41–65, 2016.

[5] Z. Pawlak, “Rough set theory and its applications,” J. Telecomm. Inf. Technol., vol. 29, no. 7, pp. 7–10, 1998.

[6] Z. Pawlak and A. Skowron, “Rudiments of rough sets,” Inf. Sci. (Ny)., vol. 177, no. 1, pp. 3–27, 2007.7

[7] Y. Y. Yao, “Information granulation and rough set approximation,” Int. J. Intell. Syst., vol. 16, no. 1, pp. 87–104, 2001.

[8] Abraham, R. Falcón, and R. Bello, Rough Set Theory: A True Landmark in Data Analysis, no. February 2016. 2009.

[9] W. Hassanein and A. Elmelegy, “An Algorithm for Selecting Clustering Attribute using Significance of Attributes,” Eur. Sci. J., vol. 10, no. 3, pp. 381–400, 2014.

[10] “A Rough Set Approach in Choosing Partitioning Attributes Lawrence J. Mazlack (mazlack@uc.edu),AijingHe (ahe@ececs.uc.edu) Yaoyao Zhu (yzhu@ececs.uc.edu).

[11] J. Liang, J. Wang, and Y. Qian, “A new measure of uncertainty based on knowledge granulation for rough sets,” Inf. Sci. (Ny)., vol. 179, no. 4, pp. 458–470, 2009.

[12] J. Uddin, R. Ghazali, and M. M.Deris, “Recent Advances on Soft Computing and Data Mining,” vol. 549, 2017.

[13] Data Availability Statement: The data set utilized in this manuscript titled New Rough Set Approach for Clustering Categorical Data Using Mean Attribute Dependency Based Measure are taken from

UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/zoo