Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 4, December 2020, pp.2575-2587

Fast attribute selection based on the rough set boundary region

[1]Muftah Mohamed Baroud, [2]Siti Zaiton Mohd Hashim, [3]Jamal Uddin Ahsan, [4]Anazida Zainal, [5]Hussein Khalaff
[1,4]School of Computing, N28, UTM Skudai Johor Bahru, Universidad Teknologi, Malaysia
[2]Institute for Artificial Intelligence and Big Data, Universiti Malaysia Kelantan (UMK), Kelantan, Malaysia
[3]Computer Science Department, IQRA National University, Peshawar KP Pakistan
[5]Windsor Regional Hospital, Ouellette Campus, 1030 Ouellette Ave, Windsor N9A 1E1. Canada

ABSTRACT

The problem of clustering exists in numerous fields such as bioinformatics, data mining, and the recognition of patterns. The function of techniques is to suitably select the best attribute from numerous contending attribute(s). RST-based approaches for definite data has gained significant attention, but cannot select clustering attributes for optimum performance. In this paper, the focus is on the processes that exhibit a similar degree of results to an identical attribute value. First, the MIA algorithm was identified as the supplement to the MSA algorithm, which experiences set approximation. Second, the proposition that MIA accomplishes lesser computational complexity through the indiscernibility relation measurement was highlighted. This observation is ascribed to the relationship between various attributes, which is markedly similar to those induced by others. Based on the fact that the size of the attribute domain is relatively small, the selection of such an attribute under such circumstances is problematic. Failure to choose the most suitable clustering attribute is challenging and the set is defined rather than computing the relative mean where it can only be implemented with a distinctive category of the information system, as illustrated with an example. Lastly, a substitute method for selecting a clustering attribute-based RST using Mean Dependency degree attribute(s) (MD) was proposed. This involved selecting the upper value of an mean attribute(s) as a clustering attribute through a considerable targeting procedure for the rapid selection of an attribute to settle the instability in selecting clustering attributes. Thus, the comparative performance of the selected clustering attributes-based RST techniques MSA and MIA was conducted.

Keywords: Clustering, Data, Algorithms, Attribute Selection, Dependency, Rough set Theory

Corresponding Author:

Muftah Mohamed Baroud
School of Computing
Universiti Teknologi Malaysia
N28, UTM Skudai Johor Bahru
E-mail: muft08@gmail.com

1. Introduction

Over two decades now, data mining has greatly developed into a multidisciplinary field that encompasses statistics, machine learning, databases, and other associated subject areas that data is a source of information [1]. Clustering is typically considered as an integral part of the data mining process. The procedure is described by the following. Assume $D = \{x_1, x_2, x_3, \dots, x_n\}$ is a collection of n objects, whereas the term x_n is an N dimensional vector for a specific typical space. The clustering process aims to organise different objects in a manner that within a similar cluster, other objects display a great level of similarities, whereas objects in different clusters exhibit a great degree of variation [1, 2]. Currently, the most extensively adopted algorithm for clustering is the classic k-means, which is applied in practically all fields. The k-means clustering is defined as the non-overlapping or separated ('crisp') clusters with double affiliations whereby an object either falls into a cluster or not. Conversely, numerous actual uses are differentiated by circumstances whereby a more appropriate depiction will be overlying clusters. Thus, the contribution is limited because the techniques are unable to handle uncertainty [3]. Accordingly, the development of robust algorithms that can withstand ambiguity during the clustering of definite data is required. However, clustering techniques primarily used for

categorical data cannot solve data uncertainty. This is considered to be a major issue in several process operations, as no sharp boundary among the different groups was found [4-6].

Recently, the categorical data for clustering has gained significant interest from the research community for data mining [7-11]. An example of a categorical data clustering technique is applied through the introduction of various attributes for clustering. Hence, the selected attribute is used to individually distribute the objects until the clustering of the entire objects is accomplished. Due to the practical problem caused by numerous candidate attributes, the user must independently select the best attribute for clustering objects according to some predefined condition. Categorical data cannot be ordered naturally, unlike numerical data. Therefore, the methods use in clustering numerical data cannot be used for clustering categorical data. Additionally, limited research has been conducted on categorical data clustering. Therefore, its contribution is limited because the techniques cannot handle uncertainty [3].

RST is a practical system that could resolve the uncertainty issue observed in the clustering of categorical data. It is a symbolic data analysis technique that was initially developed for cluster analysis [12]. It is also commonly used for clustering categorical data [13]. Originally, RST was developed as an emblematic tool for data analysis and performing cluster analysis [14]. Thus, the RST based techniques are used for assessing the relationships between the attributes based on an indiscernibility cardinality of lower approximation and dependency relationships [15]. Recently, some additions to RST techniques for clustering were suggested to solve the attribute selection problems. Likewise, [16, 17] introduced the notion of "Rough Set Theory" (RST) and introduced some rough set studies, their implementations and directions. For the most part, the clustering of RST-based attribute selection methods used in categorical data have gained great attention. The difficulty of such methods is to choose just one attribute that is most suitable to distinguish the objects from the various attribute candidates. The dependencies observed amongst the selected attributes cannot be captured by the Minimum-Minimum Roughness (MMR) and Minimum Mean Neighbourhood Roughness (MMeNR) of the latent techniques. The three techniques are based on bi-valued attributes including B-Clustering (BC). The application of MMR [18] and "Total Roughness" (TR) techniques [19] are determined by the greatest roughness value for the attribute choice to the RST algorithms. The MMR, however is considered complementary to the TR and provides the TR technique with the same accuracy and computational intricacy due to the same degree of attribute value [20]. The Maximum Dependency Attributes (MDA) approach suggested by [9] employs the indiscernibility relation, which considers the attributes' maximum dependencies for similarity measure with regards to accuracy and computational complexity. The attribute values of the identical clustering are linked to the techniques that cluster the objects with regards to MMR, TR, and MDA techniques [21, 22]. The selecting clustering attributes-based RST tactics such as "The Maximum Significance of Attributes" (MSA) proposed by [21] seeks to measure or select the maximal significance value of the best attribute. This method strengthens the clustering attribute selection process with respect to the precision of the selection of attributes. The MSA is regarded as a better technique compared to the MDA with regards to evaluation measures such as rough accuracy [23]. However, MSA does not consider similarity attributes as in the case of the MIA. This is because the approach is mainly based on the same selecting clustering attribute procedure and calculation of closely related attributes. Consequently, the maximum degree of cardinality for the selecting clustering attribute is found in the similarity attribute values. In another case, compared to the current values for equal techniques, an attribute domain is comparatively small in size. On the other hand, choosing the attribute that has identical value to best clustering attribute is unsuitable, since the selection of the objects cannot be carried out [13, 24].

"The Maximum Indiscernibility Attribute" (MIA) approach has been developed by [23]. The technique typically employs the indiscernibility relation measure to select the clustering attributes or an attribute that reveals the acquired clusters. Therefore, the number of clusters generated is observed by the cardinal values of the indiscernible relation. This theory of selecting attributes with the highest indiscernible cardinality relation stems from the suggestion that the number of clusters is large. The MIA shows a deterioration that implies that when the number of clusters is greater, finding the differences between the similarity and dissimilarity objects in a cluster is challenging. The major problem is related to selecting the clustering attribute-based RST, which is an inefficient selection of the best attribute. Some of the RST methods are unable to directly handle the selection process of clustering attributes.

The selection process requires several computation steps before convergence is attained, which leads to loss of information [25, 26]. The selection of multiple clusters could complicate the results, analysis, and increase computation costs [27-29]. The selection step is part of clustering, which employs process attributes to collect connected objects before splitting the entire objects into the clusters [30-32]. The techniques are useful in recursively acquiring extra clusters. At uninterrupted repetitions, the leaf node consumes additional objects

selected for the supplementary process of splitting. Therefore, one of the foremost challenges of the RST is the algorithm of reasoning, analysis, and decision-making processes for similar data [33]. A problem might occur when the attribute with fewer numbers is selected. The problem occurs when an equal value attribute is chosen because the clustering attribute prevents the clustering operation [34]. The inefficiencies could lead to failure in making the correct decision for data clustering. It is because the efficiency of the technique is very crucial for certain problems of large size in this age of data [35]. In the meantime, the two algorithms i.e. MSA and MIA have equal drawbacks since all the attributes considered are selected. The first drawback is involved clustering of objects caused by partitioning attribute using lower approximation wherein a single attribute is greatly similar to other prompted attributes. The second drawback is the choice of the attribute o the clustering depending on the highest value of the attribute degree measurement. Therefore, the selection of inefficient attributes for the cluster leads to high computational complexity. The highly complex computation indicates that the algorithm will be operationally complex but inefficient in choosing the clustering attributes. Based on the concepts presented by RST, the set approximation and measures of significant, roughness and indiscernibility are simple steps for measurement based on the connexion between an attribute. The rough set method for selecting clustering attributes calculates the maximal and minimum values for the best attribute selection, that is a costly computational to solve the issue. Furthermore, it is merely applicable for simple group of data or can only be implemented on specific kinds of information systems. Considering all the aforementioned drawbacks, the existing techniques for clustering require further improvements. Thus, there is an urgent need to develop a novel method for selecting clustering attributes that require a lower computational cost to evaluate the attributes in the boundary region. The proposed method will also examine their relevance to the positive region, which could enhance the clustering of the RST and reduce its uncertainty in the boundary region. This research aims to minimise computational complexity by enabling the rapid selection of attributes for data clustering by employing the mean dependency measure by selecting the maximal value of the best attribute. Thus, the method of creating an attribute selection for addressing an in-built instability can be successfully generated at the clustering centre. At consequent repetitions, the leaf node with extra objects has been designated regarding additional cutting to accomplish enhanced precision, which is employed recursively to acquire additional clusters. Furthermore, an improved method for selecting the clustering attribute-based technique using MSA and MIA is required. The relationship between attributes based on the mean dependencies is one way to selecting clustering attribute. Therefore, a procedure for selecting clustering attributes based on the RST technique is proposed in this paper. In an information system, the Mean Dependence measure of attribute(s) (MD) is determined through the RST. for selecting the upper value for the cluster attributes. Besides, the target of the noticeable procedure is for rapidly selecting an attribute to settle the instability in selected clustering attributes. The calculation and comparison of the results of an advanced method for selecting clustering attributes based on the RST method with existing algorithms such as the MSA and MIA are called a test case. The rest of the current paper is arranged as: the RST employed in the information systems and the limitations of using the RST based techniques are presented in part two. In third part, the three RST-based techniques, MSA and MIA are analysed along with descriptions of the method selecting clustering attributes based on the RST methodology along with an illustrative case study. Section 4 and 5 compares the selected clustering attributes methods based on the RST algorithms; MSA and MIA. Lastly, section 6 provides the conclusions of the study.

2. Rough Set Theory (RST)

This section discussed the fundamental definitions, conceptions, and operations of RST. The indiscernibility relation measures set approximation concepts, and dependency is presented. Furthermore, the minimization of data is explored along with discussions on the search techniques. The RST was developed to address the uncertainty and vagueness of data sets, concepts and properties in RST.

2.1. Fundamental Definitions, Concepts, and Operations of RST

2.1.1. Definition 1

Assume the information system is defined as $I = (U, K, V, \varepsilon)$, for $U = \{u_1, u_2, u_3, \dots, u_{|U|}\}$: A complete, fixed collection of objects, $K = \{k_1, k_2, k_3, \dots, k_{|U|}\}$, : Fixed complete set of attributes, $V = \bigcup_{k \in K} V_k, V_a$. The domain (value set) of attribute a, $\varepsilon = U \times K \to V$: A statistical function defined as $\varepsilon(u, k) \in V_k$.

2.1.2. Definition 2

Let $S, T \in K$ in a decision system, indiscernibility determines the relationship of equivalence between objects in I. For every $s \in S$ in I, there will be an indiscernibility relation $IND_I(S) = \{(a_1 = a_2) \in U^2 | \forall\, s \in S, S(a_{1)} = S(a_2)\}$. $IND_I(S)$ also represented by $[x]_S$ is termed an S -indiscernibility relation. If two objects $(a_1, a_2) \in IND_I(S)$, therefore, the aforementioned objects can be indistinguishable or indiscernible w.r.t. *S*.

2.1.3. Definition. 3

The S-lower estimate of X, represented by $\underline{S}\,(X)$ and S-upper estimates of X, represented by $\bar{S}(X)$, are described as $\underline{S}\,(X) = \{x \in U | [x]_S \subseteq X$ and $\bar{S}(X) = \{x \in U | [x]_S \cap X \neq \emptyset\}$, respectively. The precision of a set belonging to some subset $X \subseteq U$ concerning $S \subseteq K$, represented by $\propto_B (X)$ is determined as; $\propto_B (X) = \frac{|\underline{S}(X)|}{|\bar{S}(X)|}$.

2.2. Rough accuracy

The total roughness (TR) of any set $(a_i = \beta_k)$, $k = 1,2, \ldots \ldots n$ linked to the attribute a_j based on the attribute a_i where $i \neq j$, is represented by $X|a_i = \beta_k$ $(X|a_i = \beta_k)$, is described as follows:

$$Ra_i(X|a_i = \beta_k) = \left|\frac{Xa_j\,(a_i=\beta_k)}{Xa_j\,(a_i=\beta_k)}\right| k = 1,2, \ldots \ldots n \quad (1)$$

The TR is typically adopted to determine the rough precision of choosing the clustering attribute. The greater the TR, the greater the precision of the chosen clustering attributes $a_i \in A$ as relates to the attribute $a_j \in A$ where $\neq j$, is represented by *Rough* a_j (a_i) and assessed as follows:

$$\text{Total Roughness } (a_i) = \frac{\sum_{J=1}^{|A|} Rough_{a_j}(a_i)}{|A|-1} \quad (2)$$

The precision of the set in Eq (1) could likewise be understood based on the renowned metric proposed by Marczeweski-Steinhaus (MZ) [8, 36]. Based on the application of the MZ metric to the set calculation of a subset $X \subseteq U$ in the information system S, the relation has been gained:

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

2.3. Dependency on Attributes

2.3.1. Definition 5

Suppose the dependency of attribute $I = (U, K, V, \varepsilon)$ is the information system and let a_i and a_j signify the subset of K. The dependency attribute a_i and a_j of the point k $(0 < k < 1)$ can be represented as $a_i \Longrightarrow_{a_j}$. The point k is given by the relation [17]:

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

3. MSA and MIA Techniques Review and Contrast

The significance of a single attribute (MSA) and the indiscernibility of attributes (MIA) techniques are analysed and compared in this section.

3.1. Significance of Single Attribute (MSA)

The study by [21] presents another RST based technique that is called the Maximum Significance of Attributes (MSA).

3.1.1. Definition 5

Let $Q = (U,\ F,\ V,\ \beta)$ and $D, C \subseteq F$, where D ≠ C in information systems. The MSA utilizes the degree of attribute significance with a selection of the maximal or significant value as the best attribute of clustering. The aim is to deal with the assumed significance of a single attribute in the selecting clustering attributes.

$f_i \in$ A related to $f_j \in F$.

$\sigma f_j\,(f_i) = \gamma_{\hat{f}}\,(f_j) - \gamma_{\hat{\hat{f}}}\,(f_j)$ Proposed [16] (5)

Where $\grave{F} = F - \{f_j\},\ \ddot{F} = F' - \{f_i\}$ (6)

According to the method of choosing a clustering attribute-based (MSA) technique, the most appropriate clustering attribute was selected relying on the highest significance value. In case of similar maximum significance, the degrees appear as two or more attributes and then the next highest degree needs to be considered and the process will continue unless the tie is broken.

3.2. Indiscernibility of Attributes (MIA)

The study by [23] on the Indiscernibility of Attributes (MIA) is proven to be more effective than the earlier techniques such as MSA in some situations.

3.2.1. Definition 7

Let $Q = (U, F, V, \beta)$ be the information system. Hence, there is a need to assign a VS domain or VS value set to each $S - F$ attribute to which $S: U - VS$. The second step involves the decision for each cardinality set attributes [37]. Hence, Eq (7) is adapted to establish cardinality of Indiscernibility of attributes that are given as:

$$Card\left(Ind\ (T)\right) = |Ind(T)| \quad (7)$$

Let T be the subsection of A where the binary elements $x, y \in U$ are perceived as T-indiscernible. The Indiscernibility of the group of attributes, $T \subseteq A$ in S, if $\delta(x,t) = \delta(y,t)$ for each $t \in T$ cardinality of the indiscernibility correlation. For any available attribute present in the sum of clusters, it depicts the sum of the clusters, which can be determined based on the attribute expressed in Eq (7). In case of similar maximum significance, the degrees appear by two or more attributes then the next highest degree needs to be considered and the process will continue unless the tie is broken. According to the method of choosing clustering attributes based on the MIA and MSA techniques, the paramount attribute for clustering is selected from the maximum degree of Indiscernibility Relation (IR) and Significant measurement (SM).

Example 1. Based on the MIA [23] and MSA [21] techniques, the procedures to find the indiscernibility and significant measures, degree and values are presented. The calculations of the MIA technique are based on Eq (12) and [17], whereas the MSA technique is based on Eqs (9) and (10). According to the promotion dataset of credit cards presented in [38], the attribute maximal cardinality value of the IR and SM is selected as the attribute of the clustering. Anyways, if the values of the MIA and MSA degrees are similar to others, then a couple of attributes that are tied are considered pending when the tie is ruined. However, in reviewing MSA and MIA algorithms, it still is timewasting in computing the IR and all attributes SM degree, the outlined techniques can only be used for specific data groups. Based on the analysis, the problems and concerns related to choosing the best clustering attributes using the two techniques are explored. To illustrate the problem, the following example, based on two test cases are considered to comparatively appraise the correctness and or difficulty of the MIA and MSA approaches.

Case 1: Table 1 presents the promotion datasets for credit cards reported in the study by [38]. As observed, the five categorical attributes (n = 5) include; credit card insurance, life insurance promotion, magazine promotion, sex and watch promotion. The outlined attributes consider all binary discrete values (i.e. l = 2) with yes or no along with ten objects (m = 10). Hence, the best attribute is based on the criteria in Table 1 (from Eq. 12).

Table 1. Subset of the promotion dataset for the Acme Credit Card [38]

Person(s)	Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Sex
1	Yes	No	No	No	M
2	Yes	Yes	Yes	No	F
3	No	No	No	No	M
4	Yes	Yes	Yes	Yes	M
5	Yes	No	Yes	No	F
6	No	No	No	No	F
7	Yes	No	Yes	Yes	M
8	No	Yes	No	No	M
9	Yes	No	No	No	M
10	Yes	Yes	Yes	No	F

M: Male, F: Female

Conversely, the planned attribute selection process can lead to a problem i.e. cardinality of attribute $a_i \in A$ related to the group of attributes after computation of indiscernibility relations. The quality of the MIA will typically not be preserved by the initial choice. The adapted methodology is also unsuitable for the entire forms of the data set. The indiscernibility is similar to that of the other data set in situations as 2 or more attributes of a certain group of data are is similar to the highest cardinal value for the relationship. The MIA method delivers a similar maximum degree of indiscernibility cardinality for these data sets.

Table 2. Indiscernibility Relations Cardinality from Database with categorized attribute values

Attribute(s)	**Indiscernibility Relation(s) Cardinality Degree**	**MIA**
Magazine Promotion	2	-
Watch Promotion	2	-
Life Insurance Promotion	2	-
Credit Card Insurance	2	-
Sex	2	-

Table 3. Level of significance for all attributes in Table 1 using the MSA method

Attribute(s)	**Significance Degree**				**MSA**
Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Sex	0.2
	0.2	0.2	0	0.2	
Watch Promotion	Magazine Promotion	Life Insurance Promotion	Credit Card Insurance	Sex	0.1
	0.1	0	0	0.1	
Life Insurance Promotion	Magazine Promotion	Watch Promotion	Credit Card Insurance	Sex	0.3 0.3
	0.2	0	0.3	0.3	0.2
Credit Card Insurance	Magazine Promotion	Watch Promotion	Life Insurance Promotion	Sex	0.3 0.3 0
	0	0	0.3	0.3	
Sex	Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	0.3 0.2 0.1
	0.1	0.1	0.3	0.2	

Hence, it has 1st (2), 2nd (2), 3rd (2), 4th (2) and 5th (2) similar maximum. The method is unable to choose an attribute as its most suitable attribute for clustering in this case. It is justified that the maximum of similar values cannot be chosen. The MIA shows a deterioration that implies that when the number of clusters is similar, the search for the variances between the similarity/dissimilarity between objects in the cluster is very difficult. The "Life Insurance Promotion" attribute has the highest value of all attributes in Table 3, i.e. 0.3 before the tie is broken or the next category of attributes is considered. In this case, the second level conforming with the Life Insurance Promotion attribute, i.e. 0.3, is identical to the second level of Magazine Promotion, hence the third level corresponding to the "Life Insurance Promotion" attribute, i.e. 0.2. In this instance, the third level applicable to the Life Insurance Promotion attribute, i.e. 0.2, is greater than the third level linked to the Credit Card Insurance attribute, i.e. 0, subject to the MSA algorithm. The Life Insurance Promotion attribute is subsequently designated as the clustering attribute. For the aforementioned condition, the techniques of MIA and MSA face have selection problems or unsuccessfully chose the most suitable clustering attribute since it is impossible to choose the maximal value between similar values. However, it demonstrates indirectly that the planned technique can handle the matter effectively.

4. Attribute selection based on Mean Dependency Attribute (MD)

In this study, an alternate technique for picking the clustering attribute was proposed. Hence, it is necessary to choose the uppermost significant mean degree of dependency as the clustering attribute using an appreciable process for the fast selection of attributes. The process aims to also settle the instability in choosing the clustering attributes. After successive repetitions, the leaf node with the most objects is designated by additional splitting. The aim is to realise additional clusters of the highest quality by recursive clustering. The explanation for the uppermost value of the MD degree of attributes suggests it is more precise for choosing the attribute for clustering as presented in Eqs (9) and (10).
Figure 1 indicates the pseudo-code for the adopted approach. As noted, the methodology is based on the MD of attributes located in RST-based information systems. Usually, the technique consists of a variety of major steps. The initial phase is addressed the deduction of the level of MD between attributes. The MD level of attributes is determined on the basis of the Eq. The next step concerns the determination of the degree of MD. In the final stage, the attribute of the clustering is chosen by the attribute's highest value for the level of MD. The higher mean attribute level is more accurate when choosing the attribute of the clustering. As the greater the value of the degree attributes, the greater the precision of the clustering attribute as defined in the second proposal and the verification based on the MZ metric in Eq (3).

4.1. Theorem 1:

4.1.1. Proposition 1.

This can also be generalized in $\mathrm{I} = (\mathrm{U}, \mathrm{K}, \mathrm{V}, \varepsilon)$, let $\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_m$ and T are any subsets of K. If $\mathrm{r}_m \leq \mathrm{r}_{m-1} \leq \ldots \leq \mathrm{r}_2 \leq \mathrm{r}_1$ are the corresponding values of MD, hence for every $X \subseteq U$

$$\sigma_{\mathrm{T}}(\mathrm{X}) \leq \sigma_{\mathrm{S}_m}(\mathrm{X}) \leq \sigma_{\mathrm{S}_{m-1}}(\mathrm{X}) \leq \ldots \leq \sigma_{\mathrm{S}_2}(\mathrm{X}) \leq \sigma_{\mathrm{S}_1}(\mathrm{X}).$$

The term $|X|$ signifies the cardinality of the value X. Hence, the greater the precision of any given subset of $X \subseteq U$ is more precise (or less imprecise) than the set itself. It results in higher performance when compared to previous methods. Based on the descriptions above, it is a measurement of the space among the attributes. Hence, the mean is the highest precision of the clustering from which the best attribute clustering is achieved. According to earlier highlighted description, the selected clustering attribute is presented.

4.1.2. Proposition 2:

Assuming $\mathrm{I} = (\mathrm{U}, \mathrm{K}, \mathrm{V}, \varepsilon)$, is an information system and the terms S and T are subsections of K. If the term T hinges on S, hence the attribute S has superior roughness precision when compared to the T given as:

$$\sigma_{\mathrm{T}}(\mathrm{X}) \leq \sigma_{\mathrm{S}}(\mathrm{X})$$

4.1.3. Proposition 3:

Let $I = (U, K, V, \varepsilon)$, let $S_1, S_2, \ldots, S_m$ be any subsets of K. In finding the most suitable attribute for clustering, the value of the MD has a less uncertainty measure than the maximum value used to calculate the MD using Eq (8).
In the mean of all the dependency attribute values $(K-1)$ with respect to targeted attribute can be calculated as:

$$M_r = \frac{\sum |r_{|K-1|}|}{|K-1|} \tag{8}$$

Accordingly, the highest value of the MD degree among the set of all attributes (K) can be calculated as a second step as shown below:

$$MM_r = max\ |M_r|_K \tag{9}$$

Proof:
Let $\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_m \subseteq \mathrm{K} \in \mathrm{I} = (\mathrm{U}, \mathrm{K}, \mathrm{V}, \varepsilon)$ obtained using Eqs (9) and (10). Similarly, the relations $\mathrm{r}_{12} = \mathrm{r}_{21} = \mathrm{r}_{31} \ldots = \mathrm{r}_{m1}$, $\mathrm{r}_{13} = \mathrm{r}_{23} = \mathrm{r}_{32} \ldots = \mathrm{r}_{m2}$, $\mathrm{r}_{1m} > \mathrm{r}_{2m} > \mathrm{r}_{3m} \ldots > \mathrm{r}_{m(m-1)}$ and $i = 1,2,3, \ldots, m-1$ are considered.

Table 4. Calculation of MD value in information system

Attribute(s)	Dependency Degree				Mean Dependency Value(s)	Higher of attribute(s) value(s)
S_1	S_2	S_3		S_m		$r_{12}, r_{13}, \dots, r_{1m}$
	r_{12}	r_{13}	$\dots$	r_{1m}	$\sum r_{1i}$	
S_2	S_1	S_3		S_m		$r_{21}, r_{23}, \dots, r_{2m}$
	r_{21}	r_{23}	$\dots$	r_{2m}	$\sum r_{2i}$	
S_3	S_1	S_2		S_m		$r_{31}, r_{32}, \dots, r_{3m}$
	r_{31}	r_{32}	$\dots$	r_{3m}	$\sum r_{3i}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_m	S_1	S_2		S_{m-1}	$r_{m1}, r_{m2}, \dots, r_{m(m-1)}$	
	r_{m1}	r_{m2}	$\dots$	$r_{m(m-1)}$		$r_{m1}, r_{m2}, \dots, r_{m(m-1)}$

In the case of maximum dependency value, it requires at least $m-1$ steps whereas, for MD value, it requires just one step to find the best clustering attribute. Hence, the MD value has less uncertainty measure computationally than maximum IR and MS value.

MD Method

Input: Data set (information system) lacking the picking clustering attribute **Output:** Selecting the best attribute clustering

Begin

Step 1. For each attribute calculate the induced by MD measure attributes.

Step 2. For each, every pair attribute applies MD degree compute of all attribute with respect to another attribute.

Step 3. Select the attribute with the highest value of the mean degree as clustering attribute.

End

Figure 1. Method based RST of selecting clustering attribute

5. Comparison Test

The drawbacks of the MSA and MIA techniques are explained by example 1. Table 1 displayed the promotion data set of credit cards as described in the literature [38] .The comparison test aims to demonstrate that the planned technique can overcome the drawbacks of previous techniques MSA and MIA. Table 5 displays the results of the planned approach with the best clustering attributes based on the first test case for one small dataset.

Table 5. MD attributes degree in Table 1 using the proposed method

Attribute(s)	Mean dependency Degree				Higher of Attribute Value
Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Sex	0.222
	0	0.1	0.1	0	
Watch Promotion	Magazine Promotion	Life Insurance Promotion	Credit Card Insurance	Sex	0
	0	0	0	0	

Attribute(s)	Mean dependency Degree				Higher of Attribute Value
Life Insurance Promotion	Magazine Promotion 0.5	Watch Promotion 0	Credit Card Insurance 0.5	Sex 0	0.111
Credit Card Insurance	Magazine Promotion 0.2	Watch Promotion 0	Life Insurance Promotion 0.2	Sex 0.2	0.066
Sex	Magazine Promotion 0	Watch Promotion 0	Life Insurance Promotion 0	Credit Card Insurance 0.4	0.044

As observed, the attribute "Credit Card Insurance" has the uppermost value of MD degree of attributes, which is 0.066. Therefore, Credit Card Insurance is designated as the clustering attribute. The approach to determine the degree of mean attribute dependency in any information system is carried out using Eqs (9) and (10) as illustrated in Example 1. The findings of the suggested technique of choosing the clustering attribute were compared with several baseline approaches such as the MIA. The highest values of the SM and IR degree measures for choosing clustering attribute were determined. Typically, the greater the number of repetitions, the higher the complexity of computation, whereas low degrees of accuracy for selecting clustering attributes presents difficulties. Hence, the best clustering attribute that results in fewer numbers of attributes cannot be selected. Therefore, the suggested technique has the best clustering attributes in the first try of the three datasets obtained by further clusters. Lastly, the findings demonstrate that the suggested method is an oversimplification that has low computational complications compared to the MSA, ITDR and MIA techniques.

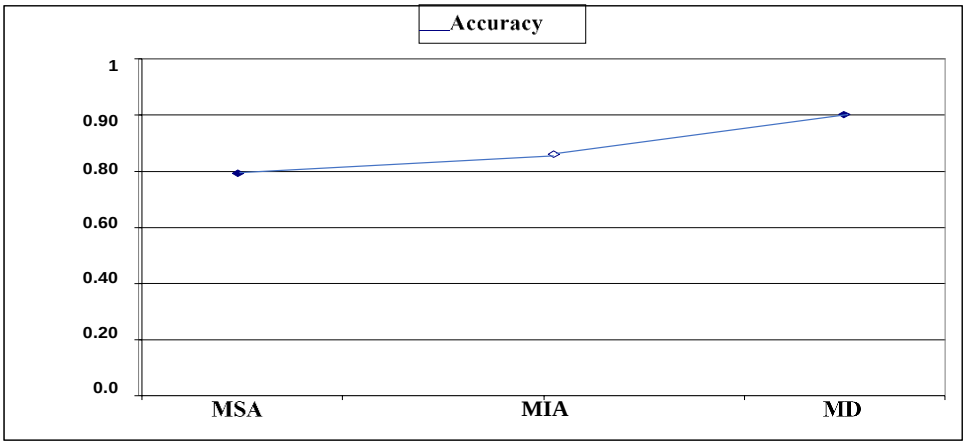

Figure 2. The accuracy of MSA and MIA techniques and MD method

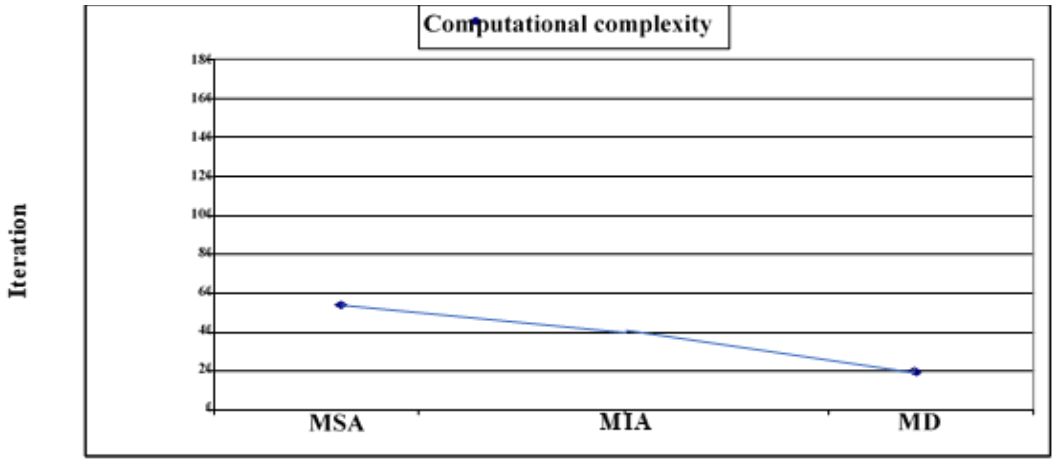

Figure 3. The accuracy of MSA and MIA techniques and MD method

5.1. Pawlak's Car Performance Data Set

Table 6 has been adopted from the literature [16]. Based on the study, a total of six cars ($U = 6$) along with three ($n = 3$) provisional attributes (a = Terrain familiarity; b = Gasoline level; c = Distance) exist in the study.

Table 6. Car performance Data Set [16]

(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

Based on the MSA technique, the importance of degrees of the entire attributes is briefly presented in Table 7. As observed, the attributes *b* and *c* exhibit the values 1 and 0.67 for the first and second maximal values, respectively. Concerning their similar maximal importance degrees, the MSA approach is hampered by the challenge of choosing the best attribute among *b* and *c*. However, MIA effectively picked the most suitable clustering attributes, based on the maximum cardinality of the IR. The cardinality of the IR of the respective attributes of the data set for car performance described by [16] is presented in Table 8. The attributes *b* and *c* according to the tests possess greater but equal IR cardinality, which is 3. Hence, the most promising groupings of *b* and *c* attributes are selected based on the MIA method. The cardinality of the IR is the only likely grouping for *b* and *c* is 5, which is the maximum value. Therefore, this subsequent grouping of attributes, i.e. *b* and *c* is designated as the most suitable clustering selection based on the MIA method.

Table 7. Degree of significance for the entire attributes in the Pawlak's Car Performance Data Set

Attribute(s)	Significant Degree		MSA
a	b	c	
	0.67	0.67	-
b	a	c	1
	0.67	1	0.67
c	a	b	1
	0.67	1	0.67

Table 8. Cardinality IR values the Car Performance Data Set from (Pawlak, 2012)

Attribute(s)	Indiscernibility Relation(s) Cardinality Degree	MIA
a	2	-
b	3	-
c	3	-
b+c	5	5

Table 9. Degree of the importance of the entire attributes from the data set of car performance [16]

Attribute(s)	Mean Dependency Degree		Higher of Attribute Value
a	b	c	
	0.167	0.167	0.668
b	a	c	
	0	0.167	0.167
c	a	b	
	0.167	0	0.167

Based on Table 9, the value of the attribute *a* is 0.668, which is considered the higher value of the mean degree of attributes. Therefore, this is considered the most appropriate attribute for clustering based on the suggested technique.

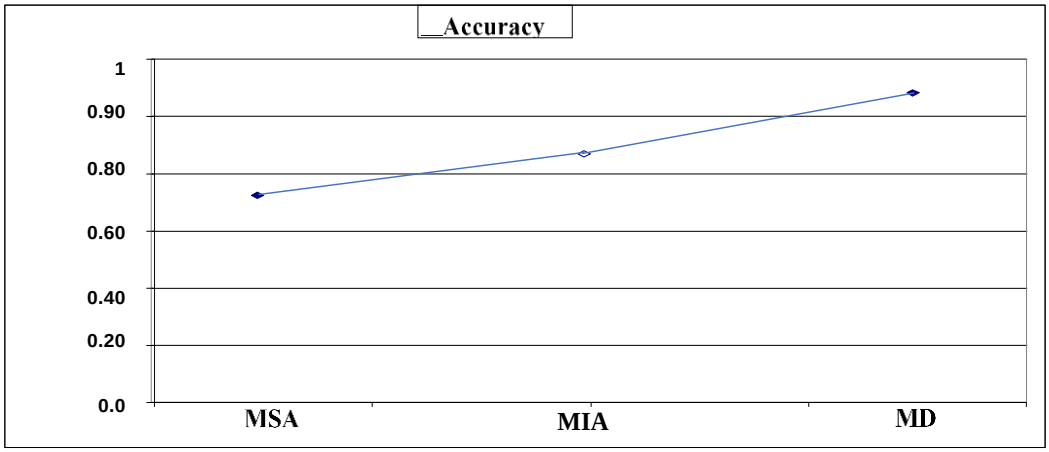

Figure 4. The accuracy of MSA and MIA techniques and MD method

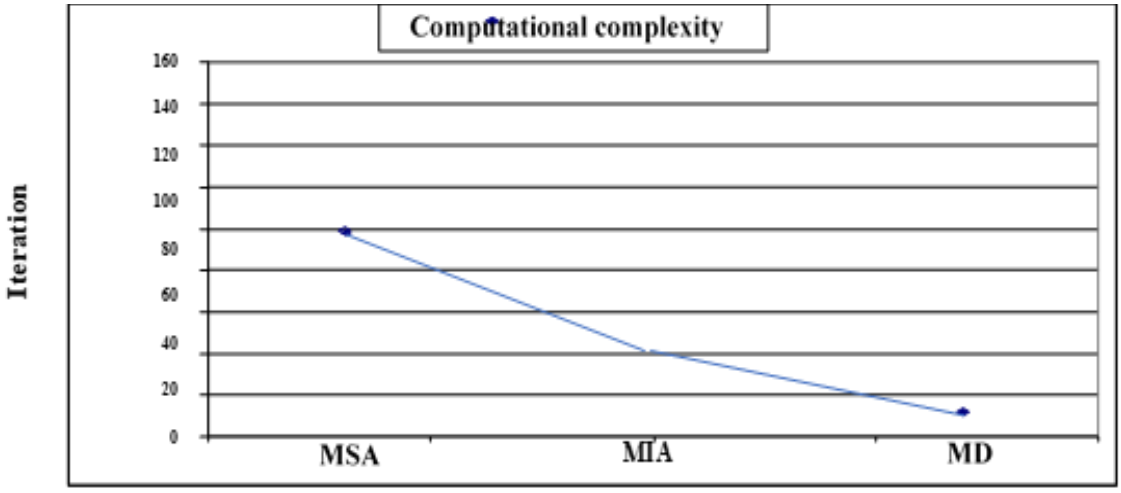

Figure 5. The accuracy of MSA and MIA techniques and MD method

6. Conclusions

The selected strategies for the clustering of categorical data with clustering attributes was recommended based on most popular algorithms such as the MSA and MIA. Although the performance of the outlined techniques seems dissimilar and the computational complexities and accuracy are lacking, there are characteristic similarities between them. This study also presented an overview of two RST-based techniques, namely; the maximum significant attribute (MSA) and the maximum indiscernibility attribute (MIA). Besides, the study recommended an alternative method for selecting clustering attributes. The proposed method is determined by the extent the Mean Dependence of the attribute(s) (MD) attribute for selecting the upper value for the cluster attributes. Besides, the target of the noticeable procedure is for rapidly selecting an attribute to settle the instability in selected clustering attributes. After successive repetitions, the leaf node with many objects is selected by recursive splitting to attain additional clusters for the highest quality of attribute selection. Thus, the increased accuracy but decreased computational complexity was demonstrated by the entire investigational findings that the recommended technique can manage, based on the limitations presented by the earlier algorithms. The suggested alternate method was evaluated and compared with two RST-based techniques, along with the recognised Marczeweski- Steinhaus (MZ) and Total Roughness (TR) metrics to determine the accuracy and rough accuracy and quality of the selected attributes. Empirical results show that the most acceptable or consistency of the attribute selected by the proposed MD method is greater than that of the attributes found by

MSA and MIA. The UCI and the benchmark small of dataset(s) implemented revealed that the performance of the suggested method has lower computational complexity and high accuracy.

References

[1] Han J and Kamber M, "Data Mining: Concepts and Techniques. A. Stephan," ed: San Francisco,, Morgan Kaufmann Publishers is an imprint of Elsevier, 2006.

[2] Ahmed Saadaldeen Rashid Ahmed, Al Barazanchi Israa, Jaaz Zahraa A, and Abdulshaheed Haider Rasheed, "Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set," *Periodicals of Engineering and Natural Sciences*, vol. 7, no. 2, pp. 448-457, 2019.

[3] Tripathy BK and Ghosh Adhir, "SDR: An algorithm for clustering categorical data using rough set theory," in *2011 IEEE Recent Advances in Intelligent Computational Systems*, 2011, pp. 867-872: IEEE.

[4] Kim Dae-Won, Lee Kwang H, and Lee Doheon, "Fuzzy clustering of categorical data using fuzzy centroids," *Pattern recognition letters*, vol. 25, no. 11, pp. 1263-1271, 2004.

[5] Huang Zhexue, "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, 1998.

[6] Belaidan Seetha Letchumy M, Yee Lim Yen, Abd Rahman Nor Azlina, and Harun Khalida Shajaratuddur, "Implementing k-means clustering algorithm in collaborative trip advisory and planning system," *Periodicals of Engineering and Natural Sciences*, vol. 7, no. 2, pp. 723-740, 2019.

[7] Barbará Daniel, Li Yi, and Couto Julia, "COOLCAT: an entropy-based algorithm for categorical clustering," in *Proceedings of the eleventh international conference on Information and knowledge management*, 2002, pp. 582-589.

[8] Guha Sudipto, Rastogi Rajeev, and Shim Kyuseok, "ROCK: A robust clustering algorithm for categorical attributes," *Information Systems*, vol. 25, no. 5, pp. 345-366, 2000.

[9] Herawan Tutut, Deris Mustafa Mat, and Abawajy Jemal H, "A rough set approach for selecting clustering attribute," *Knowledge-Based Systems*, vol. 23, no. 3, pp. 220-231, 2010.

[10] Jain Anil K, "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651-666, 2010.

[11] Mesakar Suchita S and Chaudhari MS, "Review Paper On Data Clustering Of Categorical Data," *International Journal of Engineering Research & Technology (IJERT)*, vol. 1, no. 10, pp. 1-18, 2012.

[12] Düntsch Ivo and Gediga Günther, "Rough set clustering," in *Handbook of Cluster Analysis*: Chapman and Hall/CRC, 2015, pp. 596-613.

[13] Park In-Kyoo and Choi Gyoo-Seok, "Rough set approach for clustering categorical data using information-theoretic dependency measure," *Information Systems*, vol. 48, pp. 289-295, 2015.

[14] Skowron Andrzej and Dutta Soma, "Rough sets: past, present, and future," *Natural computing*, vol. 17, no. 4, pp. 855-876, 2018.

[15] Cerrada Mariela, Sánchez René-Vinicio, Pacheco Fannia, Cabrera Diego, Zurita Grover, and Li Chuan, "Hierarchical feature selection based on relative dependency for gear fault diagnosis," *Applied Intelligence*, vol. 44, no. 3, pp. 687-703, 2016.

[16] Pawlak Zdzisław, *Rough sets: Theoretical aspects of reasoning about data*. Springer Science & Business Media, 2012.

[17] Pawlak Zdzisław and Skowron Andrzej, "Rudiments of rough sets," *Information sciences*, vol. 177, no. 1, pp. 3-27, 2007.

[18] Parmar Darshit, Wu Teresa, and Blackhurst Jennifer, "MMR: an algorithm for clustering categorical data using rough set theory," *Data & Knowledge Engineering*, vol. 63, no. 3, pp. 879-893, 2007.

[19] Mazlack Lawrence J, He Aijing, and Zhu Yaoyao, "A rough set approach in choosing partitioning attributes," in *Proceedings of the ISCA 13th International Conference (CAINE-2000*, 2000: Citeseer.

[20] Herawan Tutut, Yanto Iwan Tri Riyadi, and Deris Mustafa Mat, "Rough set approach for categorical data clustering," in *International Conference on Database Theory and Application*, 2009, pp. 179-186: Springer.

[21] Hassanein W and Elmelegy A, "An algorithm for selecting clustering attribute using significance of attributes," *International Journal of Database Theory & Application*, vol. 6, no. 5, pp. 53-66, 2013.

[22] Najeeb Shaima Miqdad Mohamed, "Finding the discriminative frequencies of motor electroencephalography signal using genetic algorithm," *Telkomnika*, vol. 19, no. 1, pp. 285-292, 2020.

[23] Uddin Jamal, Ghazali Rozaida, and Deris Mustafa Mat, "An empirical analysis of rough set categorical clustering techniques," *PloS one,* vol. 12, no. 1, 2017.
[24] Qin Yuchu, Niu Z, Chen F, Li B, and Ban Yifang, "Object-based land cover change detection for cross-sensor images," *International Journal of Remote Sensing,* vol. 34, no. 19, pp. 6723-6737, 2013.
[25] Rissino Silvia and Lambert-Torres Germano, "Rough set theory—fundamental concepts, principals, data extraction, and applications," in *Data mining and knowledge discovery in real life applications*: IntechOpen, 2009.
[26] Kumar Sharan, Jayadevappa D, and Shetty Mamata V, "A novel approach for segmentation and classification of brain MR images using cluster deformable based fusion approach," *Periodicals of Engineering and Natural Sciences,* vol. 6, no. 2, pp. 237-242, 2018.
[27] Wang Jiang, Zhu Cheng, Zhou Yun, Zhu Xianqiang, Wang Yilin, and Zhang Weiming, "From partition-based clustering to density-based clustering: Fast find clusters with diverse shapes and densities in spatial databases," *IEEE Access,* vol. 6, pp. 1718-1729, 2017.
[28] Nies Hui Wen, Zakaria Zalmiyah, Mohamad Mohd Saberi, Chan Weng Howe, Zaki Nazar, Sinnott Richard O, Napis Suhaimi, Chamoso Pablo, Omatu Sigeru, and Corchado Juan Manuel, "A Review of Computational Methods for Clustering Genes with Similar Biological Functions," *Processes,* vol. 7, no. 9, p. 550, 2019.
[29] Wang Ying and Zhang Nan, "Uncertainty analysis of knowledge reductions in rough sets," *The Scientific World Journal,* vol. 2014, 2014.
[30] Guan JW, Bell David A, and Liu DY, "The rough set approach to association rule mining," in *Third IEEE International Conference on Data Mining,* 2003, pp. 529-532: IEEE.
[31] Bi Yaxin, Anderson Terry, and McClean Sally, "A rough set model with ontologies for discovering maximal association rules in document collections," *Knowledge-Based Systems,* vol. 16, no. 5-6, pp. 243-251, 2003.
[32] Bell David A, Guan JW, and Liu DY, "Mining association rules with rough sets," in *Intelligent data mining*: Springer, 2005, pp. 163-184.
[33] Zhang Qinghua, Xie Qin, and Wang Guoyin, "A survey on rough set theory and its applications," *CAAI Transactions on Intelligence Technology,* vol. 1, no. 4, pp. 323-333, 2016.
[34] Qian Jin, Miao DQ, Zhang ZH, and Li W, "Hybrid approaches to attribute reduction based on indiscernibility and discernibility relation," *International Journal of Approximate Reasoning,* vol. 52, no. 2, pp. 212-230, 2011.
[35] Tian Zhongyan, Li Yuqian, Li Linlin, Liu Xiaotian, Zhang Haiqing, Zhang Xia, Qian Xinling, Zhou Wen, Jiang Jingjing, and Zhao Jingzhi, "Gender-specific associations of body mass index and waist circumference with type 2 diabetes mellitus in Chinese rural adults: The Henan Rural Cohort Study," *Journal of Diabetes and its Complications,* vol. 32, no. 9, pp. 824-829, 2018.
[36] Dempster Arthur P, Laird Nan M, and Rubin Donald B, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 39, no. 1, pp. 1-22, 1977.
[37] Pawlak Zdzisław, "Rough set approach to knowledge-based decision support," *European journal of operational research,* vol. 99, no. 1, pp. 48-57, 1997.
[38] Roiger Richard J and Geatz MW, "Data Mining: A Tutorial-based Primer, Pearson Education," *Inc: USA,* 2003.