View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Periodicals of Engineering and Natural Sciences (PEN - International University of...

Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 4, December 2020, pp.2424-2439

Positive region: An enhancement of partitioning attribute based rough set for categorical data

[1]Muftah Mohamed Baroud, [2]Siti Zaiton Mohd Hashim, [3]Jamal Uddin Ahsan, [4]Anazida Zainal
[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), Malaysia
[3]Computer Science Department, IQRA National University, Peshawar KP Pakistan

ABSTRACT

Datasets containing multi-value attributes are often involved in several domains, like pattern recognition, machine learning and data mining. Data partition is required in such cases. Partitioning attributes is the clustering process for the whole data set which is specified for further processing. Recently, there are already existing prominent rough set-based approaches available for group objects and for handling uncertainty data that use indiscernibility attribute and mean roughness measure to perform attribute partitioning. Nevertheless, most of the partitioning attribute methods for selecting partitioning attribute algorithm for categorical data in clustering datasets are incapable of optimal partitioning. This indiscernibility and mean roughness measures, however, require the calculation of the lower approximation, which has less accuracy and it is an expensive task to compute. This reduces the growth of the set of attributes and neglects the data found within the boundary region. This paper presents a new concept called the Positive Region Based Dependency (PRD), that calculates the attribute dependency. In order to determine the mean dependency of the attributes, that is acceptable for categorical datasets, using a positive region-based mean dependency measure (PRD) defines the method. By avoiding the lower approximation, PRD is an optimal substitute for the conventional dependency measure in partitioning attribute selection. Contrary to traditional RST partitioning methods, the proposed method can be employed as a measure of data output uncertainty and as a tailback for larger and multiple data clustering. The performance of the method presented is evaluated and compared with the algorithmes of Information-Theoretical Dependence Roughness (ITDR) and Maximum Indiscernible Attribute (MIA).

Keywords: Clustering, Rough Set Theory, Performance, Partitioning Categorical Data, Attribute Dependency

Corresponding Author:

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

1. Introduction

One of the major fields in data mining research is guided or supervised techniques. To guide the search, the supervised techniques typically focus on domain knowledge. Met clustering is a technique that multiplies randomly chosen selections, features, and weights by running K-means [1-3]. However, the automatic use of domain knowledge requires computational context-sensitivity. Thus, methods include non-semantic heuristics instead of domain knowledge. The information supplied to these applications is merely the syntactic characteristics of the database. Therefore, more than one semantic domain can be linked to these techniques. Also, unsupervised data mining has a complex design issues-particularly computational complexity. To reduce the search space, a supervised search may use domain knowledge. Hence, this study proposes that consistency can be increased by gradually partitioning the data to decrease the intra-item dissonance in the resultant partitions [4]. This is based on the basis that dissonance can be found and removed within the sub-partitions by recursively partitioning the dataset. Besides, the coherence of the resulting partitions will be higher than that of

the original data collection. Pawlak's early RST can solve problems that involve the analysis of categorical data [5, 6], such as imprecise data [4, 7].

Several researchers in this field indicate where the RST has been applied to select the clustering attributes to handle uncertainty. The idea is that dissonance can be found and removed within the sub-partitions by recursively partitioning the dataset. Therefore, the consistency of the resulting partitions will be greater than the consistency of the initial dataset. After this heuristic reduction, partitioning is achieved on the attributes with less distinct values. Following this, the first attempt is to partition the two-valued attributes. There are many different ways to select the partitioning attributes using the proposed reductionism approach. [8] suggested two methods to pick the attributes for partitioning the clusters, namely; (1) the method of bi-clustering (BC) "based on the bi-valued attributes"; and (2) the total roughness (TR) method. Furthermore, the authors proposed that the BC technique is first tested to attain minimal dissonance inside the cluster. The clustering partitioning attributes can be selected based on one of three techniques. Likewise, compared with the set of the remaining attributes of the information system, the method is related to the average mean roughness of the attribute. Usually, the higher the overall roughness, the more effective the clustering partitioning attributes are chosen. There are however, three types of problems, namely: arbitrary, balanced and imbalanced partitioning, whether unbalanced or balanced.

[9] proposed an innovative algorithm known as "Min-Min Roughness" (MMR) to enhance the bi-clustering approach for multi-value feature clustering of data and to measure its approximation accuracy using the Marczeweski-Steinhaus metric to handle multi-value attributes equally. This is normally a sub-set of approximations in the universe [10, 11]. However, MMR is a TR complement that has a similar computational intricacy and accuracy as the TR method. The TR and MMR methods are considered to be able to select a clustering attribute with comparable performance. With this technique, the difficulty is to obtain the clustering partitioning attribute based on all other attributes. [12] proposed the Maximum Dependency Attribute (MDA) algorithm that employs partitions of attributes caused by the dependency attributes' measure. This measure is required to calculate the minimal and maximum approximations for the uncertainty attributes in categorical data. Furthermore, it considers the attribute dependencies that calculates similarity in terms of purity and computational complexity.

[13] proposed the Maximum Attribute Significance (MSA) technique that employs partitions of attributes caused by the significant attributes. The measure is required to calculate the minimal and maximum approximations for uncertainty attributes in categorical data. [14] proposed the algorithm known as "The Information-Theoretic Dependency Roughness" (ITDR) was based on the mean roughness of attributes compared with the collection of other attributes in the IS. The mean total roughness indicates higher accuracy. Typically, the partitions of attributes employed by the ITDR algorithm is based on the mean roughness attributes. This measure is needed to determine the lower and upper estimates in categorical data for uncertainty attributes [15].

The proposed Maximum Indiscernibility Attribute (MIA) technique typically employs a novel data partitioning approach based on the attribute' indiscernibility connection of showing the clusters received. The MIA used for data partitioning employs the partitioning of attributes induced by the measure of the indiscernibility relation. To compute the upper and lower estimate of uncertainty attributes, the measure is necessary for categorical data. This method helps to build the correlation between the partitioning of the upper and lower estimation cardinality of the indiscernibility attribute. The set-up includes the partitioning of objects induced by approximation sets, where a single attribute is substantially comparable to another attribute stimulated by others. The two techniques generate corresponding values of partitioning attributes. However, it is not desirable for the partitioning attribute to have a similar value as the selecting partitioning attribute because it is not possible to partition the objects [14, 16].

However, the RST techniques has several drawbacks. The first drawback is that the techniques ignores the uncertain attributes within the boundary region, which may include the information required to enhance the performance of attribute clustering [17-19]. This is a very challenging drawback because the lower approximation involves a attributes that may be not directly significant to the concept [20].The higher uncertainty among the approximation sets reduces the performance of the rough set clustering technique. An estimate of objects is therefore, one of the main problems in rough sets [21-24]. In other words, there is less uncertainty involved when complete information available. The inappropriate approximation of a set is conducted to estimate the upper and lower approximations, which aims to classify objects from categorical attributes that will decrease the growth of the attribute subset and increase the equivalence classes.

This paper therefore proposes an innovative method of partitioning, referred to as 'Positive Region Based Dependency' (PRD), which measures the mean dependency of the attribute without the lower approximation being used. To determine the mean dependency of the attributes, PRD describes the method using a positive region-based mean dependency measure that is appropriate regarding categorical groups of data. PRD may be an optimal substitute for the normal dependency measure in the choosing process of partitioning attributes by avoiding the lower approximation. The proposed partitioning method was employed to solve the inappropriate partitioning attribute(s) of categorical data. The proposed method can detect the calculated traces in the positive region method and identify the object partitioning strategy through the RST boundary region. The study is organised in two sections: Second part focuses using of RST in an information system, while Section 3 presents the analysis of two RST partitioning approaches used (ITDR and MIA). The drawbacks of the RST-based partitioning techniques are presented in Section 4, along with an explanatory small data set. The findings of the newly proposed approach have also been compared to the techniques of ITDR and MIA. The conclusions are presented in Section 5, along with the rationale for the hypothesis of the study.

2. Fundamentals of RST

The fundamental concept of the RST described in [25] is explained in this part of the study. The lesser approximation of the universe U includes the entire objects contained in the group X categorically. The higher approximation of U, by comparison, consists of the whole objects that are likely to be found in X. However, the lack of information prevents classify of the entire potential objects, whereas the outstanding indistinguishable ones exist in the appositely termed "boundary region" [26]. The RST technique focuses on inducing the relationship between indiscernible objects to produce the approximation area and reduce concepts [27]. The RST technique is a subset region with two specific terms, namely; the negative, positive, and boundary regions. The positive section of a set comprises of all the elements in the set X. The set boundaries include all elements in X and contain all the objects that unclassifiable by using the complete information accessible to the set and its complements [28]. Furthermore, the boundary region is rarely crisp but exists as an occupied boundary region in every rough set when a diverse crisp is obtained. The motivation for RST stems from the importance of depicting the universe U with regards to the similar groups of a universe partition [29]. This can be formally described as follows: X boundary regions negative and positive are correspondingly determined based on a sub-set K of U and an invisible relationship IND as:

a. Positive region $POS_{(S)}(T) = \underline{S}(X)$ of set X in relation to S, T is the group of complete objects which can be characterised precisely as X using S, T, as defined X in relation to S, T.
b. Boundary region $BND_{(S)}(T) = \underline{S}(X) - \bar{S}(X)$ of set X in relation to S, T is the set of complete objects, potentially categorized as X using S, T, and possibly X in relation to S, T.
c. Negative region $NEG_{(S)}(T) = U - \bar{S}(X)$ is the set of complete objects which are definite when categorized as not X utilising S, T, which are not X with regards to S, T.
d. An attribute set $S, T \subseteq K$ is said to maintain a positive region except if it creates the same positive region as K does i.e., $POS_S(T)$. If S, T preserves the positive region affirmed by K, it must also preserve the boundary area well-defined by K. Considering Pawlak's rough set pattern, an attribute set $S, T \subseteq K$ that preserves both the region and boundary with the positive region potentially preserves the clustering quality.

The RST is a characteristic process of soft computing that was developed and rapidly applied after its establishment. The theory is characterized by the following physiognomies;

a. The RST configuration is well established and does not need forehand information.
b. It can be readily computed due to its simplicity.
c. It can process inexact, inadequate and ambiguous types of data.
d. It expresses the simplest reduction and attributes cores of the knowledge.
e. Approximate descriptions of ambiguous conceptions are described at various stages of granularity.
f. Specific and streamlined guidelines are created and applied to intelligent controls.

According to the abovementioned features, three peculiarities exist amongst the RST and other concepts of uncertainty. Firstly, the RST does not require previous information except for the data. Furthermore, the RST is comparatively independent when describing and addressing uncertainty. Lastly, the RST strongly complements the fuzzy mathematics, along with the theories of probability and evidence relative to other concepts or techniques of uncertainty. However, only discrete data can be processed through conventional categorical data clustering techniques, although most categorical data is recurrent. Therefore, the process of

managing categorical data utilizing the techniques of "Rough Set Theory" will remain valuable in the future for data clustering research issues.

2.1. Main concepts and key definitions

2.1.1. Definition 1. Information system (IS)

The information system is a quadruple, i.e. $I = (U, K, V, \varepsilon)$, where U defined as a group of objects which are non-empty that is finite. K is a fixed set of attributes that are non-empty. $V = \bigcup_{k \in K} V_k, V_k$ is the value set of the attributes $K, \varepsilon: U \times K \to V, \varepsilon(u, k) \in V_k$ for each $(u, k)\ U \times K$, which is identified as the information function [29]. Naturally, the information system can be termed a table that is part of an attribute valued system.

2.1.2. Definition 2. Indiscernibility relation

The indiscernibility relation $Ind(B)$ is a relation on U [29]. Given that the two objects $(x_i, x_j) \in U$are indiscernible by the attribute sets B in A, specifically if $a(x_i) = a(x_j)$ for each $a \in B$. Meaning, $((x_i, x_j) \in Ind(B))$ specifically if $V_a \in$ B, where $B \subseteq A,\ a(x_i) = a(x_j)$.

2.1.3. Definition 3. Equivalence classes

A group of objects x_i with an attribute set in B consisting of a uniform class $[x_i]_{Ind(B)}$ given $Ind(B)$, proposed by [30] is set to equivalence class $[x_i]_{Ind(B)}$. This is also called the basic set for B.

2.1.4. Definition 4. Approximation

The S-lower estimate of X is represented as $\underline{S}(X)$, whereas the S-upper estimate of X is represented as $\bar{S}(X)$ and expressed as:

$$\underline{S}(X) = \{x \in U | [x]_S \subseteq X\} \quad (1)$$

$$\bar{S}(X) = \{x \in U | [x]_S \cap X \neq \emptyset\} \quad (2)$$

The term $|X|$ represents the cardinality of the X set [29]. According to the initiation of uniform groups, the universe U can be categorized into three disjoint areas. Given the attribute subgroups S and T for K, the significant notions of the boundary regions, negative and positive are examined separately as the positive $POS_S(T)$, the boundary $BND_S(T)$, and negative $NEG_S(T)$ regions, which are highlighted in Definition 4. Although selected regions are empty, it can be inferred that an element $x \in POS_S(T)$ goes with K and that the element $x \in NEG_S(T)$ is not part of K. However, it cannot be ascertained whether or not the element $x \in BND_S(T)$ is part of K. The concept of rough set regions and rough set estimations are able to definitely be extended to the partition of the universe. According to the positive, negative, and boundary regions of S and T, the respective three disjoint regions of U, where $I(x)$ donates the objects' class imperceptible with x. Hence,
If the object $x \in POS(X)$, hence it is part of the target set X.
If the object $x \in BND(X)$ hence it is not part of the target set X.
If the object $x \in NEG(X)$ hence it is not ascertained if the object X is part of the target set X or not.

2.1.5. Definition 5. Total roughness

From the TR procedure, the average roughness of the attribute $a_i \in A$ for $a_j \in A$ where $i \neq j$, is represented as $Rough\ a_j\ (a_i)$, and evaluated as follows:

$$ai(a_i) = \frac{\sum_{k=1}^{|V(a_i)|} R_{a_j}(X | a_i = \beta_k)}{|V(a_i)|} \quad (3)$$

2.1.6. Information-theoretic dependency roughness (ITDR) algorithm

In this equation, $V(a_i)$ is an attribute set of values, $a_i \in A$. Let $Q = (U, F, V, \beta)$ be the upper and lower approximation, and suppose M and N are several subgroups of F and $M,\ N \neq \emptyset$. Hence, the ITDR is an attribute N on elements M. $M \Rightarrow_H N$ is clear from Eq. (4).

$$H(N_i | M_j) = \left\{ \begin{matrix} - \\ 1.0 \end{matrix} \sum_{j=1}^{n} R_j \log_2 |M_j \cap N_i| \ /\ |M_j|,\ |M_j| \cap N_i| > 0, |M_j \cap N_i| = 0 \right. \quad (4)$$

The ITDR technique proposed by [14] is more effective than earlier techniques such as Min Mean Roughness (MMeR) [31], Standard Deviation Roughness (SSDR) [32], "Min-Min Roughness (MMR)" [9], [32] in some situations.

2.1.7. Indiscernibility of attributes (MIA) algorithm

There are three major steps in the MIA technique to determine the indiscernibility of attributes. The first procedure is to compute each collection value of an attribute using lower and upper estimates. In the $Q = (U, F, V, \beta)$ information system, 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 attribute [33]. Eq. (5) is adapted to establish cardinality of the indiscernibility of attributes.

$$Card\left(Ind\ (T)\right) = |Ind(T)| \tag{5}$$

Let T be the subset of A, where two elements $(x, y \in U)$ are seen to be T-indiscernible. The indiscernibility of the attributes set $T \subseteq A$ in S if $\delta(x,t) = \delta(y,t)$ for each $t \in T$ cardinality of the indiscernibility correlation for an available attribute present in the number of clusters, which depicts the amount of the determinable clusters in the attribute as shown in Eq. (5).

2.1.8. Dependency on attributes

Suppose the dependency of the attribute $S = (U, A, V, F)$ is the information system and let a_i and a_j signify the subset of A. The dependency attributes a_i and a_j in degree k $(0 < k < 1)$ is denoted by $a_i \Rightarrow_{a_j}$. The degree k is given by [6] as:

$$K\gamma_{a_i} = \frac{\sum_{X \in U/a_j} a_i(x)}{|U|} \tag{6}$$

2.2. Application of the ITDR technique

A small-sized dataset Animal World Dataset [9], was used to apply the ITDR technique, as shown in Table 1, [34]. There are 9 mortals and 9 explicit elements: Teeth, Feet, Eye, Hair, Eat, Fly, Feather, Swim, and Milk. The elements 'Hair', 'Eye', 'Feather', 'Milk', 'Fly', and 'Swim' comprise 2 traits. The element 'Teeth' has 3 traits and various aspects have 4 qualities. Hence, the mean roughness of attributes of subsets of U can be determined based on the row of the attributes "Hair, Teeth, Eye, Feet, Eat, Milk, Fly, and Swim". Dataset in the illustrative examples of Table 1 can be considered [12]. Likewise, the mean roughness of the subsets of U related to separate attributes can be determined from other attributes using equation. (3). According to [14], the attributes 'Hair', 'Feather' and 'Milk' are identical and have a similar partitioning attribute value of 0.10. However, if the attribute's average roughness $a_i \in A$ for the attributes' group of $A - \{a_i\}$ can be measured, then the values of the ITDR technique can be obtained, as presented in Table 2. As the Table 1 shows, a higher degree of partitioning attribute cannot be determined. On the other hand, the ITDR technique leads to the following problem: after calculating the attribute's average roughness $a_i \in A$ for the attributes' set of $A - \{a_i\}$, the ITDR technique's value is unable to preserve the original decision. Thus, it can be deduced that the modified ITDR technique does not apply to all types of datasets.

Table 1. Animal world dataset [9]

Row(s)	Hair	Teeth	Eye	Feather	Feet	Eat	Milk	Fly	Swim
Tiger	Y	Pointed	Forward	N	Claw	Meat	Y	N	Y
Cheetah	Y	Pointed	Forward	N	Claw	Meat	Y	N	Y
Giraffe	Y	Blunt	Side	N	Hoof	Grass	Y	N	N
Ostrich	Y	Blunt	Side	N	Hoof	Grass	Y	N	N
Zebra	N	N	Side	Y	Claw	Grain	N	N	N
Penguin	N	N	Side	Y	Web	Fish	N	N	Y
Albatross	N	N	Side	Y	Claw	Grain	N	Y	Y
Eagle	N	N	Forward	Y	Claw	Meat	N	Y	N
Viper	N	Pointed	Forward	N	N	Meat	N	N	N

Y: Yes, N: No

Table 2. Estimate of average roughness for every attribute based on the ITDR technique

Attribute(s)	Average Roughness of Attribute(s)	Average
Hair Rough	$a_j(a_1)$, j = 2,3,4,5,6,7,8,9 (0, 0.2959,0.0620,0,0,0,0.2176,0.2959)	0.10
Teeth Rough	$a_j(a_2)$, j = 1,3,4,5,6,7,8,9 (0.2467,0.3069,0.2643,0,0.0426,0.2467,0.5445,0.3751)	0.25
Eye Rough	$a_j(a_3)$, j = 1,2,4,5,6,7,8,9 (0.2959,0,0.2058,0,0,0.2959,0.1540,0.2959)	0.15
Feather Rough	$a_j(a_4)$, j = 1,2,3,5,6,7,8,9 (0.0620,0,0.2058,0,0,0.0620,0.1309,0.2959)	0.10
Feet Rough	$a_j(a_5)$, j 1,2,3,4,6,7,8,9 (0.5950,0.2775,0.5368,0.5368,0.1540,0.5950,0.7436,0.5368)	0.49
Eat Rough	$a_j(a_6)$, j 1,2,3,4,5,7,8,9 (0.5048,0.2310,0.4781,0.4293,0.1273,0.5048,0.5706,0.5123)	0.41
Milk Rough	$a_j(a_7)$, j 1,2,3,4,5,6,8,9 (0, 0,0.2959,0.0620,0,0,0.2176,0.2959)	0.10
Fly Rough	$a_j(a_8)$, j 1,2,3,4,5,6,7,9 (0.2545,0.1540,0.3700,0.1540,0.2545,0.0770,0.2545,0.3700)	0.23
Swim Rough	$a_j(a_9)$, j 1,2,3,4,5,6,7,8 (0.2959,0.0676,0.2959,0.2959,0,0,0.2959,0.1540)	0.17

2.3. Application of the MIA algorithm

A small-sized dataset of the enrolment qualifications of students was used to apply the MIA algorithm [12]. It comprises 8 items (m = 8) with 7 categories of qualities (n = 7), namely: English, Degree, Experience, Mathematics, Statistics, Programming, and IT.

Table 3. Data system for the enrolment qualifications of students [12]
G: Good, M: Medium, B: Bad

Degree(s)	English	Experience	IT	Mathematic(s)	Programming	Statistic(s)
PhD	G	M	G	G	G	G
PhD	M	M	G	G	G	G
MSc	M	M	M	G	G	G
MSc	M	M	M	G	G	M
MSc	M	M	M	M	M	M
MSc	M	M	M	M	M	M
BSc	M	G	G	M	M	M
BSc	B	G	G	M	M	G

The indiscernibility of the subset(s) attributes of U can be obtained according to the attribute 'Degree' for further attributes (Experience, English, Mathematics, Statistics, Programming, and IT). The dataset in Table 3 was this considered [12]. In addition, the indiscernibility of the subsets of U can be determined from each attribute for other attributes using equation. (5). According to [15], the attributes 'Degree' and 'English' are identical that have a similar partitioning attribute value of 3. Moreover, the attributes 'IT', 'Mathematics', 'Programming', and 'Statistics' are identical with a partitioning attribute value of 2. However, the determination of the attribute's average roughness $a_i \in A$ for the attributes' set of $A - \{a_i\}$ results in the values of the MIA technique shown in Table 4.

Table 4. Maximal indiscernibility degree of the attributes

Attribute(s)	Indiscernibility Degree	MIA
Degree	3	—
English	3	—
Experience	2	—
IT	2	—
Mathematics	2	—
Programming	2	—
Statistics	2	—

Based on Table 3, a higher degree of partitioning attribute cannot be determined. Hence, the MIA technique results in the following problem: after calculating the attribute's average roughness $a_i \in A$ for the attributes' class $A - \{a_i\}$, the magnitude of the MIA method not able to sustain the initial outcome. Thus, the adapted MIA technique does not apply to the entire class of datasets. Based on the findings in Tables 2 and 4 as well as the contingent on quality, there are several drawbacks with the RST algorithms. The first drawback is the mean roughness and the degree of indiscernibility relation cardinality of the two algorithms. Parenthetically, the ITDR and MIA algorithms are primarily focused on calculating the minimal and maximal approximation sets for the subsets of U according to the significance, mean roughness, and indiscernibility relation cardinality degree. This shows the similarity attribute partitioning values, which results in similar attributes values with higher uncertainty and less accuracy. The second drawback is that the inappropriate partitioning attribute can hamper selection of the partitioning attribute and clustering centre to split the attributes for categorical data clustering. However, this cannot always be achieved since decreasing the accuracy causes higher uncertainty. Hence, the Positive Region based Dependency (PRD) Partitioning Method is introduced in the next section to address the partitioning attribute problems.

3. Proposed positive region based mean dependency (PRD) method

Assuming the pair $IS = (U, AT)$ of non-empty, fixed sets U and AT (U is the universe of the object and AT is a attributes group) is presented. Furthermore, every attribute is defined as a function where $K \rightarrow V_{at}$ (V_{at} is denoted as the set attribute value at the domain of AT . The pair $IS = (U, AT)$ is termed an IS [33], which is comparable to the concept of classification propounded by [35]. Typically, an IS may be described by tables with rows and columns of data categorised by objects and attributes, respectively. For the pair (x, at), assuming $x \in U$ and at $at \in AT$ describes the specific admission specified as the value $at(x)$ value in the table.

The initial method PRD was founded on indiscernibility as described by the relations of equivalence class denoted by (U, IND). The term U is the fixed set whereas $IND \subseteq U \times U$ is the relation of equivalence of U. Typically, an indiscernibility expression describes the partition of the objects of universe.

In the past, numerous generalizations of the technique have been presented but the majority are founded on coverings instead of partitions in general, the approach of dominance-based rough group, fuzzy and rough mixed groups among others are rough set approaches such as similarity notable variants [36-38].

The partitioning-based methods of clustering are suitable for all kinds of data. To use these techniques, however it is important to have prior knowledge of the number of clusters. The primary factor that needs to be considered when selecting the best partitioning attributes is the unique partitioning objects of the attribute set. The objective is to divide the data points into K partitions in partitioning-based clustering methods, where each partition represents one class. The task of handling uncertainty is attributed to the boundary region with the rough set regions. The higher the uncertainty degree of the rough set is related with the large size of the boundary area, [22, 39]. The idea is based on the magnitude of the attribute subset of the original attributes S, T and the number of equivalence groups, which will produce a larger positive region [40, 41].

Therefore, the unique partitioning attribute method uses the mean dependency measure (MD) to compute the positive area of the categorical attributes. The high accuracy can be easily achieved for the MDM since the positive area increases whereas the boundary region decreases. The unique partitioning attribute method consists of two main steps: (1) the RST boundary region in addition to the union attributes; (2) degree of mean of dependency attributes. In this study, Positive Region based Dependency (PRD), measures the attribute dependency. In order to determine the mean dependency of the attributes, which is acceptable for categorical datasets, (PRD) describe the method using a positive region-based dependency measure (PRD), by avoiding the lower approximation. This method means that a balanced way to partition the attributes was accomplished.

The crisp partitioning attributes are the optimal condition, where no dissonance exists in the values. The uncertainty degree can therefore be detected and scaled utilising the following roughness and accuracy:

a. Identify the most distant (less dissonance) from others to minimise dissonance.
b. Identify the feasible partitioning that satisfies the suggested Positive Region based Dependency (PRD) method which a rough partition since the Positive Region based Dependency (PRD) technique reflects the rough partitioning, different forms of partitioning can result in different measures.
c. Compare the Positive Region based Dependency (PRD) method can identify a balanced approach to partition.
d. Consider a quick selection of partitioning attributes to determine the clustering centre.

3.1. RST positive boundary regions in addition to the union attributes

The partitioning attribute method takes advantage of the RST boundary region. The method helps to measure the partition instigated by a certain attribute subgroup. The Positive Region based Dependency measure (PRD) method can generate several significant characteristics, such as effective characterization of uncertainty information in the boundary region. The theory of the positive area was suggested by Pawlak in [42] and employed to compute the importance of the attribute's status in the table of chosen decision. Although the attribute's notion decline by the positive area was invented by J.W. Grzymala-Busse as described in the literature [43, 44]. The equivalent algorithm neglects the extra computation necessary to select optimal attributes. Hence, the study proposes the Positive Region based Dependency measure (PRD) method for data partition that maintains the boundary region of target decision making. [45] extended the positive region partition to compute the selected attribute in the background of the rough set. Due to the uniformity of concepts and guidelines of the techniques, the method of [46] is considered demonstrative. This partitioning method is a new partitioning effort for objects splitting using to selecting a partitioning attribute.

3.1.1. Definition 6

The two elements $x, y \in U$ are reportedly S-indiscernible (i.e., indiscernible due to the set of attributes $S \subseteq K$ in I but only when $f(x,a) = f(y,a)$, for all $a \in K$. Every subset of K prompts an exceptional relation of indiscernibility. Typically, a relation of n indiscernibility prompted by the attributes set K represented by $IND(S)$ is a relation of equivalence. The partition of X prompted by $IND(S)$ is represented by U/S, and the corresponding class in the partition U/S comprising $x \in U$ is represented by $[x]_S$. The partition of U prompted by the positive and boundary regions is represented by $POS_S(T)$ and $BND_S(T)$ of a region.

3.1.2. Definition 7

The partitioning measure offers an extra method to analyse data. The attribute "T" is completely reliant on the attribute "S" if "S" exclusively governs the value of "T". Officially, in the decision system $I = (U, K, V, \varepsilon)$, with S, T as the subsets of K, then the attribute "T" is dependent on the attribute "S" by degree "r" which is calculated using Eq. (7). The term $r = \delta(S,T)$ represents the fraction, which the samples on the universe U are separated into the S, T positive and S, T boundary regions approximately or certainly. The positive region is the combination of the entire corresponding modules in $[x]_S$, contained in subsets of the objective set. Since $NEG_S(T) = \emptyset$,$POS_S(T) \cap BND_S(T) = \emptyset$ and $POS_S(T) \cup BND_S(T) = U$, it could be necessary to consider only $POS_S(T)$. As will be shown later when the notions are spread wide in a probabilistic version, it is essential to deliberate on the two regions (with regards to $S, T \subseteq K$). It is also known that K is the largest element subgroup of the original attributes and the smallest equivalence group attribute may be higher than the relative positive component region [40, 47]. Here, "r" stipulates the elements' ratio which could be positively included in a partition prompted by S, T i.e., U/S, U/T is termed the partitioning attributes measure:

$$r = \delta(S,T) = \frac{|POS_S(T) \cup BND_S(T)|}{|U|} \tag{7}$$

The unification of the entire equivalence groups known as positive region in $[x]_S$ that is enclosed in the subsets of the objective set. If $r = 1$, "T" is fully reliant on S; for $0 < r < 1$, "T" is partly reliant on "S"; and for $r = 0$, "T" is independent of "S". It is evident that when $r = 1$, i.e., "T" is reliant on "S", then $IND_I(S) \subseteq IND_I(T)$. In simple terms, $\frac{U}{S}$ is finer than $\frac{U}{T}$.

3.2. Degree of mean dependency attributes

The second part represents the possible way of calculating the level of dependency on the factors. The dependency measure for the partitioning element is introduced to optimize information from the positive boundary region. Figure 1 shows the rough set and the three related disjoint regions that comprise the set.

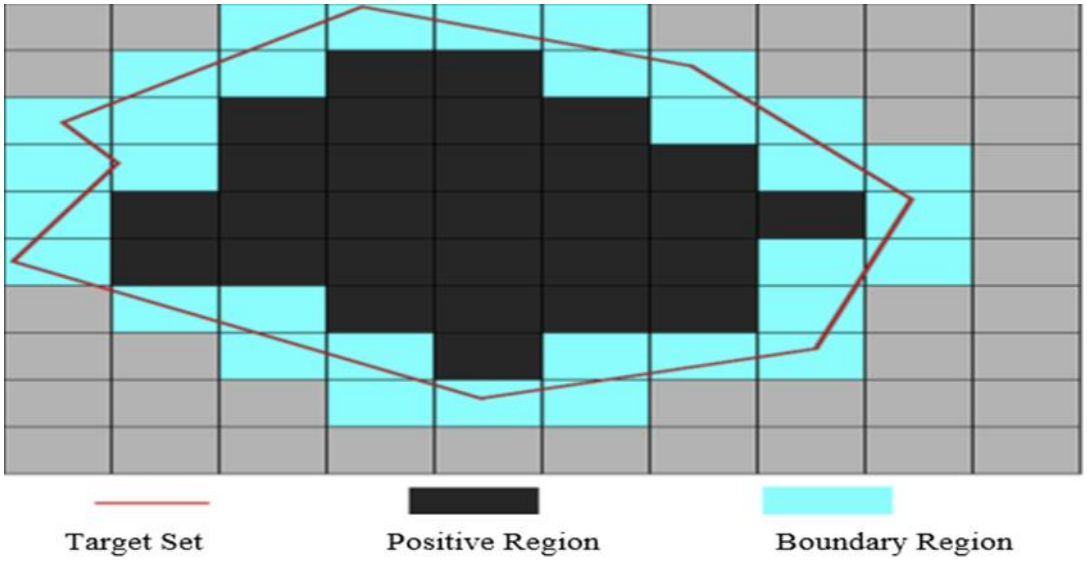

Fig 1. A rough set-in rough region space

The ambiguity of a rough set is instigated by its boundary area., the higher uncertainty degree is connected with the larger boundary area. The roughness defines the ambiguity of a rough set. The level and completeness of knowledge for a given objective subset are specified by roughness and accuracy. Incidentally, the roughness and accuracy show the number of elements in each approximation along with the potential for application in the uncertainty assessment of the boundary region [48].

3.2.1. Definition 8

The irregularity of any rough set [49] is given by an IS $I = (U, K, V, \varepsilon)$ for any objective subset $X \subseteq U$ and the attribute subset $S \subseteq K$. Hence, the irregularity of set X for S is defined by the relation:

$$\sigma_S(X) = 1 - \frac{|\underline{S}(X)|}{|\overline{S}(X)|} \tag{8}$$

The term $X \neq \emptyset$ (if $X = \emptyset$, then $\sigma_S(X) = 0$); $|0|$represents the cardinality of a fixed set. If X is the merger of selected classes which are equivalent of U, then $\sigma_S(X) = 1$. Therefore, set X is precise (crisp) for S. Besides, if X is not a union of some equivalence classes of U, then $\sigma_S(X) < 1$, and the set X is rough for S. This verifies the higher accuracy of the roughness of subset $X \subseteq U$. Therefore, the greater the roughness, the greater the accuracy of the partitioning clustering attributes.

[50] proposed the concept of soft computing (SC), which is defined as a key intelligent technological system for the future. Over the years, RST has been widely researched and implemented to solve numerous practical difficulties. The fundamental standard of RST is to permit the inexact, vague or undiscernible approximate explanations to substitute the exact explanations of an initial problem. Hence, a robust and inexpensive solution could be recognized to better organize the actual systems [51]. The subdivision K's dependency degree on the partition U/S and U/T, can be quantified using the positive region. Hence, [29] suggested the following estimate. Therefore, it is observed that according to the hypothesis, the partitioning U/S is finer than U/T, which gives $IND(S) \subseteq IND(T)$. This equation is the definition of the partitioning measure, which is given as follows:

$$\mathrm{r} = \delta(\mathrm{S}, \mathrm{T}) = \frac{|POS_S(T) \cup BND_S(T))|}{|\mathrm{U}|} = \frac{\sum_{\mathrm{i}=0}^{\mathrm{m}} |\mathrm{S_i}|}{|\mathrm{U}|} \tag{9}$$

Therefore, $r = \delta(S, T)$ represents the fraction of the samples on the universe U separated as S, T positive and S, T boundary regions approximately or certainly. The positive region serves as a merger of the entire equivalence classes in $[x]_S$, which is confined by subsets of the objective set. If $r = 1$, "T" is completely reliant on S; for $0 < r < 1$, "T" is partly reliant on "S"; and for $r = 0$, "T" is independent of "S". It is obvious that

when $r = 1$, i.e., "T" is reliant on "S", then $IND_I(S) \subseteq IND_I(T)$. In uncertain relations, U/S is greater than U/T.

PRD Method

Input: Dataset without partitioning **Output:** Partitioning attributes

1. Compute the equivalence classes using indeclinability relation on each attribute.
2. Calculate the positive region for the uncertainty objects in categorical attributes by taking advantage of the RST boundary region of attribute concerning all and where
3. To partition the attributes, apply the mean dependency measure of each attribute within the degree of attributes.

End.

Fig 2. Partitioning attributes method steps

4. Experiments

For the experiments, a small-sized dataset defining the appearance of objects was considered [52], as presented in Table 5.

Table 5. Information scheme for the appearance of objects

U/A	Shape	Colour	Area
1	Circle	Red	Big
2	Circle	Red	Small
3	Triangle	Blue	Small
4	Triangle	Green	Small
5	Circle	Blue	Small

As observed, the dataset has five categorical attributes (n = 3), namely: (1) Shape (Shape), (2) Color (Color), (3) Area (Area). Two attributes have two distinct meanings (l = 2), and one attribute (m = 3) was taken into consideration.

4.1. Computations

The positive boundary region, relative Mean dependency, and subsets of U were obtained according to the attribute 'Shape' as it relates to other attributes (Color, and Area) from an IS of the appearance of the objects [52]. The values acquired for the proposed method are summarized in Table 6.

4.2. Equivalence classes

With reference to Table 5 and based on each attribute, there are several equivalent classes of U prompted by the exclusive indiscernibility relation of each presented attribute. $U/Shap$ = {{1,2}, {3,4}, {5}}, $U/Color$ = {{1,2}, {3,5}{4}}, $U/Area$ = {{1}, {2, 3, 4, 5}}.

4.3. Application of proposed method

The expression specified below identifies the positive region $POS_S(T)$ of "U/T" in the context of "S" $POS_S(T) = \cup_{X \in U/T} \underline{S}(X)$. The mean dependency of the attribute S, T on k is expressed as $S, T \Rightarrow rk$. The positive region of portion U/S (denoted as $POS_S(T)$ is set off by the entire objects that can be exclusively categorized to block the partition U/S through S. However, the negative region (denoted as $NEG_S(T)$ is a collection of objects that cannot be categorised by the partition U/S. Lastly, the boundary region (defined as $BND_S(T)$ is an objects' group of that cannot be categorized as such. The following degrees are obtained by using the same approach:

$$r = \delta(S, T) = \frac{|POS_S(T) \cup BND_S(T))|}{|U|} = \frac{\sum_{i=0}^{m} |S_i|}{|U|}$$

1- Shape, Color
S1= (1,2,5) T1= (1,2)

S2= (3,4) T2= (3, 5)
T3= (4)

Relative to the partition, there exists a complete dependence of the region of the attribute on the Color of the attribute ', i.e. {Color}) $\Longrightarrow$ r $-$ 1{Shape}, and the term r is computed as follows:

Color $\Rightarrow$ Shape, $r = \delta(Color, Shape) = \frac{|\{3\}+\{4\}|}{|1,2,3,4,5|} = \frac{2}{5} = 0.4$

For the attribute's dependency {Color}) $\Longrightarrow$ k $-$ 1{Shape}, the unit value is introduced as r $= \frac{2}{5}$, since the area of the two objects can be exclusively computed by using the attribute 'Area'.

2- Color, Area

S1= (1) T1= (1,2)
S2= (2,3,4,5) T2= (3,5)
T3= (4)

Relative to the partition, there exists a complete dependence of attribute region on the attribute 'Aera', i.e. {Area}) $\Longrightarrow$ r $-$ 1{Color}, and the term r is calculated as follows: $Area \Rightarrow$ Color, $r = \delta(Area, Color) = \frac{|\{1\}|}{|1,2,3,4,5|} = \frac{1}{5}$ = 0.2. For the dependence attribute {Area}) $\Longrightarrow$ k $-$ 1{Color}, the value of the degree is introduced by r $= \frac{1}{5}$, since the area of two objects can be distinctively ascertained by using the attribute 'Shape'.

3- Shape, Color

S1= (1,2,5) T1= (1,2)
S2= (3,4) T2= (3,5)
T3= (4)

Relative to the partition, there is a complete dependence on attribute area on attribute 'Aera', i.e. {Area}) $\Longrightarrow$ r $-$ 1{Shape}, and the term r is calculated as follows: Color $\Rightarrow$ Shape, $r = \delta(Color, Shape) = \frac{|\{0\}|}{|1,2,3,4,5|} = \frac{0}{5}$ = 0. For the dependence attribute {Color}) $\Longrightarrow$ k $-$ 1{Shape}, the degree value is presented as r $= \frac{0}{5}$, as two objects' area could be identified uniquely by utilizing the attribute 'Area'.

$$r = \delta(S,T) = \frac{|POS_S(T) \cup BND_S(T))|}{|U|} = \frac{\sum_{i=0}^{m} |S_i|}{|U|}$$

$$POS_S(T) = \cup_{X \in U/T} \underline{S}(X)$$

$$NEG_S(T) = U - \cup_{X \in U/T} \bar{S}(X)$$

$$BND_S(T) = \cup_{X \in U/T} \bar{S}(X) - \cup_{X \in U/T} \underline{S}(X)$$

The set of attributes S is completely contingent on the class of attributes T, represented by $S \Longrightarrow_k T$, when the entire attributes from S are exclusively resolved by values of attributes from S. T, which relies on S in the degree of r as represented by equation. (7). This equation is the definition of the partitioning measure of the mean dependency degree. The mean dependency's level may be identified utilising Eq. (9). The results gained from employing the proposed technique to the dataset in Table 5 (An IS of objects' appearance in the small dataset [34], are presented in Table 6.

Table 6. Results obtained using the proposed method based on the dataset in Table 5

Attribute(s)	**Degree of $POS_S(T) \cup BND_S(T)$ based Dependency**		**PRD**
Shape	**Colour**	**Area**	
	0.4	0.2	0.3
Colour	Shape	Area	
	0.2	0.2	0.2
Area	Shape	Colour	
	0.4	0.1	0.25

Based on the proposed method in this study, a test was performed and the results presented in Table 6. Thus, the attribute 'Shape' is having a higher mean attribute degree (0.3) compared with the other four attributes. Therefore, the degree of PRD method does not have similar attribute values to the animal world dataset [9] in Table 1, which is comparable to the ITDR results in Table 2. The proposed method was also tested and compared

to the performance of the procedure using this dataset and the findings are presented in Table 7. The attribute 'Eat' was found to have a high mean degree of attributes (0.48) compared to the other attributes. Therefore, the novel proposed method (which is based on the degree of partitioning attribute) does not result in similar attribute values. Using equations. (7) and (9), the accuracy of the partitioning attributes and uncertainty for the datasets in Tables 2 and 7 were determined and the findings are illustrated in Figure 3.

Table 7. Findings gained utilising the proposed method based on the dataset in Table 1

Row(s)	**Degree of $POS_S(T) \cup BND_S(T)$ based Dependency**								**PRD**
Hair	Teeth	Eye	Feather	Feet	Eat	Milk	Fly	Swim	
	0	0	0.4	0	0	1	0.4	0	0.23
Teeth	Hair	Eye	Feather	Feet	Eat	Milk	Fly	Swim	
	0.6	0.5	0.5	0.4	1	0.6	0.5	0	0.46
Eye	Hair	Teeth	Feather	Feet	Eat	Milk	Fly	Swim	
	0	0	0.4	0	0.4	0	0	0	0.1
Feather	Hair	Teeth	Eye	Feet	Eat	Milk	Fly	Swim	
	0.4	1	0	0	0.5	1	0.4	0	0.42
Feet	Hair	Teeth	Eye	Feather	Eat	Milk	Fly	Swim	
	0.4	0.4	0.5	0.5	0.5	0.5	0.5	0.5	0.41
Eat	Hair	Teeth	Eye	Feather	Feet	Milk	Fly	Swim	
	1	0.5	1	0.5	0.5	0.5	0.3	0.3	0.48
Milk	Hair	Teeth	Eye	Feather	Feet	Eat	Fly	Swim	
	1	0.6	0	0.4	0.4	0.5	0	0	0.37
Fly	Hair	Teeth	Eye	Feather	Feet	Eat	Milk	Swim	
	0.4	0.5	0	0	0.4	0.3	0.2	0	0.22
Swim	Hair	Teeth	Eye	Feather	Feet	Eat	Milk	Fly	
	0	0	0	0	0	0	0	0	0

The accuracy of the partitioning attributes using the proposed PRD method is 0.90, which is higher than the ITDR technique (0.80). In the meantime, the proposed method has lower uncertainty, as evidenced by the higher degree of accuracy required compared with the ITDR technique. Hence, the proposed method was tested against the dataset in Table 3 (IS of the enrolment qualifications of students from [12] and compared with the results obtained from the MIA technique (Table 4). The results are presented in Table 8. The results showed that the attribute 'Degree' had the highest mean degree of attributes (0.542) among the six attributes. The attribute values were not the same using the proposed method. Based on Tables 4 and 8, and equations. (3) and (8), the accuracy of the partitioning attributes and uncertainty for the third dataset were determined and the results are shown in Figure 3. In this case, the accuracy of the MIA technique is 0.75, whereas the proposed method PRD, is slightly higher at 0.85. However, the proposed method has lower uncertainty compared to the MIA technique as indicated by the higher degree of accuracy required. From Tables 2, and 4, comparing the performance of the partitioning attributes of the ITDR and MIA techniques was done through test.

Table 8. Results obtained based on the dataset in Table 4

Attribute(s)	**Degree of $POS_S(T) \cup BND_S(T)$ based dependency**						**PRD**
Degree	English	Experience	IT	Math's	Programming	Statistics	
	0.5	1	0.5	0.5	0.5	0.25	0.542
English	Degree	Experience	IT	Math's	Programming	Statistics	
	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Experience	Degree	English	IT	Math's	Programming	Statistics	
	0.25	0	0.25	0.25	0.25	0	0.166
IT	Degree	English	Experience	Math's	Programming	Statistics	
	0.5	0.5	0.5	0.25	0.25	0.25	0.375
Math's	Degree	English	Experience	IT	Programming	Statistics	
	0	0	0.25	0	1	0	0.208
Programming	Degree	English	Experience	IT	Math's	Statistics	
	0	0	0.25	0	1	0	0.208
Statistics	Degree	English	Experience	IT	Math's	Programming	
	0	0.25	0	0	0	0	0.25

Furthermore, the results show that the techniques use a conventional rough set of attribute partitioning. Likewise, for partitioning attribute clustering, a partitioning-based clustering technique is used. This technique is based on partitioning the equivalent classes by different attribute measures and requires the same calculation of objects in the uncertainty of the attributes. This approach enables the selection of a few partitioning attributes that designate the inferior achievement of the technique. The evidence is based on the results of Tables 7 and 8 from comparative tests of the performance of the procedure utilising 3 test cases.

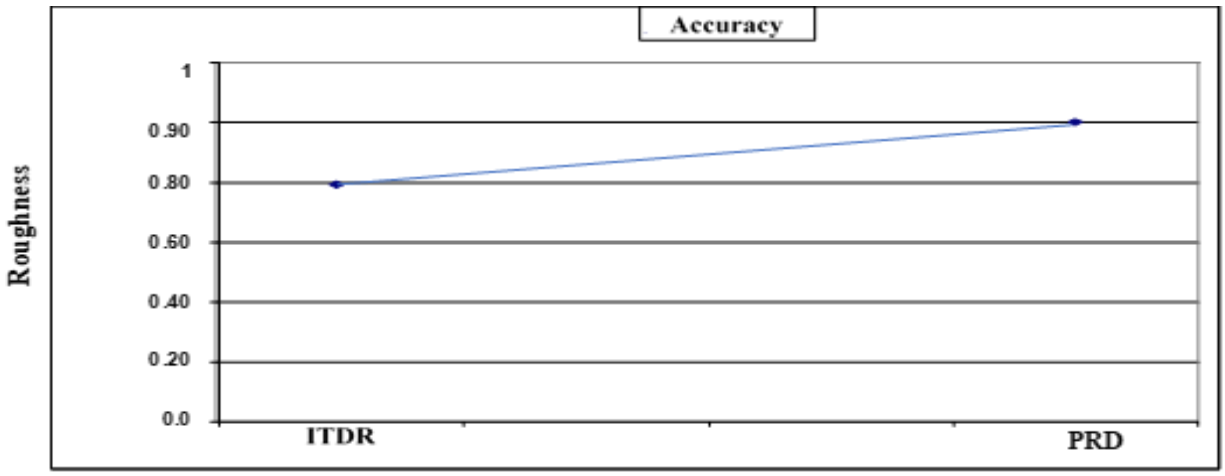

Fig 3. Accuracy of the ITDR technique and PRD method

The proposed method revealed different high partitioning attributes of reliance for the traits and therefore, this method delivers quality attribute partitioning. The proposed method PRD is a better partitioning attribute clustering method based on the quality of the results presented. When accounting for the proposed method, only a single assessment is made to determine the highest value and the most salient characteristic feature equivalency by different traits. Hence, the most salient attribute should be recorded based on the value of the suggested technique. In addition, the suggested technique has enhanced the partitioning attribute values, which effectively correspond to the splitting attribute and determines the best selecting partitioning attribute. Since the values of the ITDR, MIA techniques cannot usually preserve the original decision, the modified partitioning method can apply to all types of datasets.

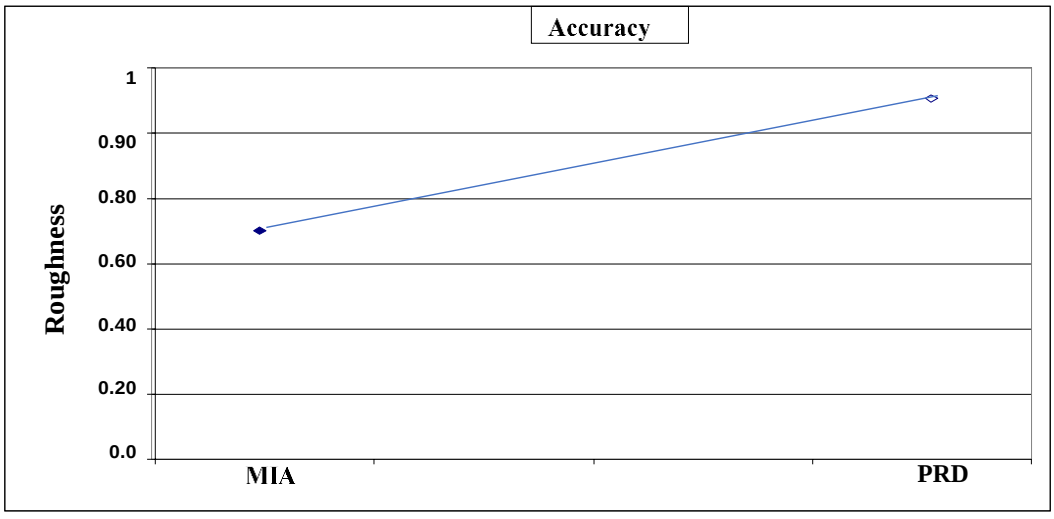

Fig 4. Accuracy of the MIA technique and PRD method

5. Conclusion

An improved RST-based procedure for data partition, partitioning attribute, and multi-value attribute resolution along with data indiscernibility was proposed in this study. In selected problems, such as greater uncertainty and lower accuracy, current algorithms such as ITDR and MIA were applied to partition the attributes. The

constraints of the two ITDR and MIA approaches were therefore calculated on the basis of the mean dependency of all attribute measures. This measure is required for the positive region to be determined for attributes of uncertainty in categorical data. This is achieved by using the RST boundary region to partition objects and to solve the inappropriate partitioning of categorical data. The Positive Boundary Regions using the mean Dependency measure method (PRD) (based on unique partitioning attributes) attained the highest accuracy. Hence, the proposed technique results in lower uncertainty and more available knowledge. In addition, the proposed method effectively clustered the various categorical small datasets. Lastly, the experimental results suggest that the technique can overcome previous constraints related to several experiments performed using small benchmark and UCI datasets.

References

[1] S. Kumar, D. Jayadevappa, and M. V. Shetty, "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.

[2] R. Caruana, M. Elhawary, N. Nguyen, and C. Smith, "Meta clustering," in *Sixth International Conference on Data Mining (ICDM'06)*, 2006: IEEE, pp. 107-118.

[3] S. R. A. Ahmed, I. Al Barazanchi, Z. A. Jaaz, and H. R. Abdulshaheed, "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.

[4] L. J. Mazlack, "Softly focusing on data," in *18th International Conference of the North American Fuzzy Information Processing Society-NAFIPS (Cat. No. 99TH8397)*, 1999: IEEE, pp. 700-704.

[5] Z. Pawlak, "Rough classification," *International Journal of Man-Machine Studies,* vol. 20, no. 5, pp. 469-483, 1984.

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

[7] S. L. M. Belaidan, L. Y. Yee, N. A. Abd Rahman, and K. S. Harun, "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.

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

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

[10] Y. Yao, "Two views of the theory of rough sets in finite universes," *International journal of approximate reasoning,* vol. 15, no. 4, pp. 291-317, 1996.

[11] Y. Yao, "Information granulation and rough set approximation," *International Journal of Intelligent Systems,* vol. 16, no. 1, pp. 87-104, 2001.

[12] T. Herawan, M. M. Deris, and J. H. Abawajy, "A rough set approach for selecting clustering attribute," *Knowledge-Based Systems,* vol. 23, no. 3, pp. 220-231, 2010.

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

[14] I.-K. Park and G.-S. Choi, "Rough set approach for clustering categorical data using information-theoretic dependency measure," *Information Systems,* vol. 48, pp. 289-295, 2015.

[15] J. Uddin, R. Ghazali, and M. M. Deris, "An empirical analysis of rough set categorical clustering techniques," *PloS one,* vol. 12, no. 1, 2017.

[16] H. Qin, X. Ma, J. M. Zain, N. Sulaiman, and T. Herawan, "A Mean Mutual Information Based Approach for Selecting Clustering Attribute," in *International Conference on Software Engineering and Computer Systems*, 2011: Springer, pp. 1-15.

[17] S. Eskandari and M. M. Javidi, "Online streaming feature selection using rough sets," *International Journal of Approximate Reasoning,* vol. 69, pp. 35-57, 2016.

[18] Z. Lu, Z. Qin, Y. Zhang, and J. Fang, "A fast feature selection approach based on rough set boundary regions," *Pattern Recognition Letters,* vol. 36, pp. 81-88, 2014.

[19] L. Zhang, Y. Li, C. Sun, and W. Nadee, "Rough set based approach to text classification," in *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2013, vol. 3: IEEE, pp. 245-252.

[20] S. Rissino and G. Lambert-Torres, "Rough set theory—fundamental concepts, principals, data extraction, and applications," in *Data mining and knowledge discovery in real life applications*: IntechOpen, 2009.
[21] Z. Wang, H. Yue, and J. Deng, "An Uncertainty Measure Based on Lower and Upper Approximations for Generalized Rough set Models," *Fundamenta Informaticae,* vol. 166, no. 3, pp. 273-296, 2019.
[22] P. Mandal and A. Ranadive, "Fuzzy multi-granulation decision-theoretic rough sets based on fuzzy preference relation," *Soft Computing,* vol. 23, no. 1, pp. 85-99, 2019.
[23] Y. Wang and N. Zhang, "Uncertainty analysis of knowledge reductions in rough sets," *The Scientific World Journal,* vol. 2014, 2014.
[24] Q. Zhang, Q. Xie, and G. Wang, "A survey on rough set theory and its applications," *CAAI Transactions on Intelligence Technology,* vol. 1, no. 4, pp. 323-333, 2016.
[25] Z. Pawlak, "Vagueness and uncertainty: a rough set perspective," *Computational intelligence,* vol. 11, no. 2, pp. 227-232, 1995.
[26] V. Torra, "Hesitant fuzzy sets," *International Journal of Intelligent Systems,* vol. 25, no. 6, pp. 529-539, 2010.
[27] Z. Y. Xu *et al.*, "A Correlation Analysis Model of Fault Location of Distribution System Based on RS-IA Data Mining," in *Applied Mechanics and Materials*, 2017, vol. 863: Trans Tech Publ, pp. 345-354.
[28] P. C. Xuyen, D. S. Truong, and N. T. Tung, "AN INFORMATION-THEORETIC METRIC BASED METHOD FOR SELECTING CLUSTERING ATTRIBUTE," *PROCEEDING of Publishing House for Science and Technology,* 2017.
[29] Z. Pawlak, *Rough sets: Theoretical aspects of reasoning about data*. Springer Science & Business Media, 2012.
[30] Z. Pawlak, "Rough set approach to knowledge-based decision support," *European journal of operational research,* vol. 99, no. 1, pp. 48-57, 1997.
[31] P. Kumar and B. Tripathy, "MMeR: an algorithm for clustering heterogeneous data using rough set theory," *International Journal of Rapid Manufacturing,* vol. 1, no. 2, pp. 189-207, 2009.
[32] B. Tripathy and A. Ghosh, "SDR: An algorithm for clustering categorical data using rough set theory," in *2011 IEEE Recent Advances in Intelligent Computational Systems*, 2011: IEEE, pp. 867-872.
[33] Z. Pawlak, "Rough sets," *International journal of computer & information sciences,* vol. 11, no. 5, pp. 341-356, 1982.
[34] S. Maitrey and Y. K. Gupta, "Data Mining—A Tool for Handling Huge Voluminous Data," in *Applications of Machine Learning*: Springer, 2020, pp. 177-188.
[35] J. Barwise and J. Seligman, *Information flow: the logic of distributed systems*. Cambridge University Press, 1997.
[36] J. Kacprzyk and W. Pedrycz, *Springer handbook of computational intelligence*. Springer, 2015.
[37] S. Vluymans, Y. Saeys, C. Cornelis, A. Teredesai, and M. De Cock, "Fuzzy Rough Set Prototype Selection for Regression," in *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2015: IEEE, pp. 1-8.
[38] P. Skowron, "What do we elect committees for? A voting committee model for multi-winner rules," *arXiv preprint arXiv:1611.06858,* 2016.
[39] J. Yu, X. Zhang, Z. Zhao, and W. Xu, "Uncertainty measures in multigranulation with different grades rough set based on dominance relation 1," *Journal of intelligent & fuzzy systems,* vol. 31, no. 2, pp. 1133-1144, 2016.
[40] Y. Qu, C. Shang, Q. Shen, N. Mac Parthaláin, and W. Wu, "Kernel-based fuzzy-rough nearest neighbour classification," in *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, 2011: IEEE, pp. 1523-1529.
[41] A. Skowron, A. Jankowski, and S. Dutta, "Interactive granular computing," *Granular Computing,* vol. 1, no. 2, pp. 95-113, 2016.
[42] S.-Y. Huang, *Intelligent decision support: handbook of applications and advances of the rough sets theory*. Springer Science & Business Media, 1992.
[43] J. Grzymala-Busse, "An algorithm for computing a single covering," *Managing Uncertainty in Expert Systems,* p. 66, 1991.
[44] J. W. Grzymala-Busse, "LERS-a system for learning from examples based on rough sets," in *Intelligent decision support*: Springer, 1992, pp. 3-18.
[45] Q. Hu, Z. Xie, and D. Yu, "Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation," *Pattern recognition,* vol. 40, no. 12, pp. 3509-3521, 2007.

[46] X. Hu and N. Cercone, "Learning in relational databases: a rough set approach," *Computational intelligence*, vol. 11, no. 2, pp. 323-338, 1995.
[47] A. Skowron and S. Dutta, "Rough sets: past, present, and future," *Natural computing*, vol. 17, no. 4, pp. 855-876, 2018.
[48] N. Xie, M. Liu, Z. Li, and G. Zhang, "New measures of uncertainty for an interval-valued information system," *Information Sciences*, vol. 470, pp. 156-174, 2019.
[49] G. Wang, "Rough set theory and knowledge discovery," *Xi'an Jiaotong University Press, Xi'an,* 2001.
[50] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338-353, 1965.
[51] Q. Zhang, "Research on hierarchical granular computing theory and its application [D]," *Southwest Jiaotong University, Chengdu, China*, 2009.
[52] D. Mining, "A Tutorial-based Primer," ed: Addison Wesley, Boston, 2003.