# NEW ROUGH SET BASED MAXIMUM PARTITIONING ATTRIBUTE ALGORITHM FOR CATEGORICAL DATA CLUSTERING

MUFTAH MOHAMED JOMAH BAROUD

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

SEPTEMBER 2022

## DEDICATION

Dedicated to my beloved family
To the souls of most precious persons in my life: my parents, to my brothers, sisters
and to my sincere wife and my sweetheart beautiful sons and daughters

# ACKNOWLEDGEMENT

Thanks to ALLAH, the Most Gracious, the Most Merciful, the Most Bountiful who gave me the courage and patience to accomplish this research work. Without his help and mercy, this would not have come into reality.

I would like to deeply express my gratitude to my Supervisor, AP. Dr. Siti Zaiton Mohd Hashim, my Co-Supervisor Prof. Dr. Siti Mariyam Shamsuddin and AP. Dr. Anazida Zainal for her fascinating guidance, encouragement, valuable comments goodness and humanity during my study. I would like to express my grateful to AP. Dr. Anazida Zainal for her support and for all the people who helped to complete the search even by a pray.

Last, but not the least, my greatest thanks from my heart to my family for giving the unlimited supports and patience to complete my study. I would never ever forget their sacrifice that they have done for me. Their prayers and support provided me with force and determination.

.

## ABSTRACT

Clustering a set of data into homogeneous groups is a fundamental operation in data mining. Recently, consideration has been put on categorical data clustering, where the data set consists of non-numerical attributes. However, implementing several existing categorical clustering algorithms is challenging as some cannot handle uncertainty while others have stability issues. The Rough Set theory (RST) is a mathematical tool for dealing with categorical data and handling uncertainty. It is also used to identify cause-effect relationships in databases as a form of learning and data mining. Therefore, this study aims to address the issues of uncertainty and stability for categorical clustering, and it proposes an improved algorithm centred on RST. The proposed method employed the partitioning measure to calculate the information system's positive and boundary regions of attributes. Firstly, an attributes partitioning method called Positive Region-based Indiscernibility (PRI) was developed to address the uncertainty issue in attribute partitioning for categorical data. The PRI method requires the positive and boundary regions-based partitioning calculation method. Next, to address the computational complexity issue in the clustering process, a clustering attribute selection method called Maximum Mean Partitioning (MMP) is introduced by computing the mean. The MMP method selects the maximum degree of the mean attribute, and the attribute with the maximum mean partitioning value is chosen as the best clustering attribute. The integration of proposed PRI and MMP methods generated a new rough set hybrid clustering algorithm for categorical data clustering algorithm named Maximum Partitioning Attribute (MPA) algorithm. This hybrid algorithm is an all-inclusive solution for uncertainty, computational complexity, cluster purity, and higher accuracy in attribute partitioning and selecting a clustering attribute. The proposed MPA algorithm is compared against the baseline algorithms, namely Maximum Significance Attribute (MSA), Information-Theoretic Dependency Roughness (ITDR), Maximum Indiscernibility Attribute (MIA), and simple classical K-Mean. In addition, seven small data sets from previously utilized research cases and 21 UCI repository and benchmark datasets are used for validation. Finally, the results were presented in tabular and graphical form, showing the proposed MPA algorithm outperforms the baseline algorithms for all data sets. Furthermore, the results showed that the proposed MPA algorithm improves the rough accuracy against MSA, ITDR, and MIA by 54.42%. Hence, the MPA algorithm has reduced the computational complexity compared to MSA, ITDR, and MIA with 77.11% less time and 58.66% minimum iterations. Similarly, a significant percentage improvement, up to 97.35%, was observed for overall purity by the MPA algorithm against MSA, ITDR, and MIA. In addition, the increment up to 34.41% of the overall accuracy of simple K-means by MPA has been obtained. Hence, it is proven that the proposed MPA has given promising solutions to address the categorical data clustering problem.

# ABSTRAK

Mengelompokkan set data ke dalam kumpulan homogen adalah operasi asas dalam perlombongan data. Baru-baru ini, fokus penyelidikan telah diberikan pada pengelompokan data kategori, di mana set data terdiri daripada atribut bukan angka. Walau bagaimanapun, melaksanakan beberapa algoritma pengelompokan kategori sedia ada adalah mencabar kerana sesetengahnya tidak dapat menangani ketidakpastian manakala yang lain mempunyai masalah kestabilan. Teori Set Kasar (RST) ialah alat matematik untuk menangani data kategori dan mengendalikan ketidakpastian. Ia juga digunakan untuk mengenal pasti hubungan sebabakibat dalam pangkalan data sebagai satu bentuk pembelajaran dan perlombongan data. Oleh itu, kajian ini bertujuan untuk menangani isu ketidakpastian dan kestabilan dalam pengelompokan kategori, dan ia mencadangkan algoritma yang lebih baik berkait dengan RST. Kaedah yang dicadangkan menggunakan ukuran pembahagian untuk mengira Kawasan positif dan sempadan atribut untuk sistem maklumat. Pertama, kaedah pembahagian atribut yang dipanggil Indiscernibility berasaskan Wilayah Positif (PRI) telah dibangunkan untuk menangani isu ketidakpastian dalam pembahagian atribut untuk data kategori. Kaedah PRI memerlukan kaedah pengiraan pembahagian berasaskan wilayah positif dan sempadan. Seterusnya, untuk menangani isu kerumitan pengiraan dalam proses pengelompokan, kaedah pemilihan atribut pengelompokan yang dipanggil Pemisahan Min Maksimum (MMP) diperkenalkan dengan mengira nilai min. Kaedah MMP memilih darjah maksimum atribut min dan atribut dengan nilai pembahagian min maksimum dipilih sebagai atribut pengelompokan terbaik. Penyepaduan kaedah PRI dan MMP yang dicadangkan menghasilkan algoritma pengelompokan hibrid set kasar baharu untuk algoritma pengelompokan data kategori yang dinamakan algoritma Atribut Pembahagian Maksimum (MPA). Algoritma hibrid ini ialah penyelesaian menyeluruh untuk ketidakpastian, kerumitan pengiraan, ketulenan kelompok dan ketepatan yang lebih tinggi dalam pembahagian atribut dan pemilihan atribut pengelompokan. Algoritma MPA yang dicadangkan dibandingkan dengan algoritma garis dasar, iaitu Atribut Kepentingan Maksimum (MSA), Kekasaran Ketergantungan Teoritik Maklumat (ITDR), Atribut Kebolehlihatan Maksimum (MIA) dan K-Mean klasik yang ringkas. Selain itu, tujuh set data kecil daripada kes penyelidikan yang digunakan sebelum ini dan 21 repositori UCI dan set penanda aras digunakan untuk pengesahan. Seterusnya, keputusan dibentangkan dalam bentuk jadual dan grafik, telah menunjukkan bahawa algoritma MPA yang dicadangkan mengatasi algoritma garis dasar untuk semua set data. Tambahan pula, keputusan menunjukkan bahawa algoritma MPA yang dicadangkan meningkatkan ketepatan kasar terhadap MSA, ITDR dan MIA sebanyak 54.42%. Oleh itu, algoritma MPA berjaya mengurangkan kerumitan pengiraan berbanding dengan MSA, ITDR dan MIA dengan 77.11% masa dan 58.66% lelaran minimum. Begitu juga, peratusan peningkatan yang ketara sehingga 97.35% diperhatikan untuk ketulenan keseluruhan oleh Algoritma MPA terhadap MSA, ITDR dan MIA. Di samping itu, peningkatan ketepatan sehingga 34.41% diperoleh daripada ketepatan keseluruhan K-means mudah oleh MPA. Oleh itu, terbukti bahawa MPA yang dicadangkan berpotensi memberikan penyelesaian yang lebih baik dalam menangani masalah pengelompokan data kategorikal.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| BC | - | Bi-Clustering |
| TR | - | Total Roughness |
| MMR | - | Min-Min Roughness |
| MSA | - | Maximum Significance Attribute |
| SDR | - | Standard-Deviation Roughness |
| MDA | - | Maximum Dependency Attribute |
| ITDR | - | Information-Theoretic Dependency Roughness |
| MIA | - | Maximum Value Attribute |
| SSDR | - | Standard Deviation Roughness |
| RST | - | Rough Set Theory |
| SBM | - | Supply Base Management |
| MMeR | - | Min-Mean-Roughness |
| U | | Universe of Objects |
| NoC | - | Number of Clusters |
| PRI | - | Positive Region based Indiscernibility |
| MMP | - | Maximum Mean Partitioning |
| MPA | | Maximum Partitioning Attributes |
| IND | - | Indiscernibility |

## LIST OF SYMBOLS

| | | |
|---|---|---|
| $\|U\|$ | - | Cardinality of Objects |
| [x] | - | Equivalence Class |
| $V$ | - | Value |
| $\subseteq$ | - | Sub-Set |
| $\cup$ | - | Union |
| $\cap$ | - | Intersection |
| $\bar{S}$ | - | Upper Approximation |
| $\underline{S}$ | - | Lower Approximation |
| $\emptyset$ | - | Empty Set |
| U | - | Universe of Objects |
| POS | - | Positive Region |
| BND | - | Boundary Region |
| NGE | - | Negative Region |
| $\delta$ | - | Delta Function |
| ( ) | - | Value Function |
| / | - | Division |
| { , } | - | Set Brackets |
| { , } | - | The Absolute Value |
| $=$ | - | Equal |
| $\neq$ | - | Not Equal |
| $\in$ | - | Belongs To |
| $\notin$ | - | Not Belongs To |
| $\sum$ | - | Summation |
| $\leq$ | - | Less Than or Equal |
| $\rightarrow$ | - | Implication |
| $A$ | - | Attribute |
| $f$ | - | Function |
| O | - | Time Complexity |

## LIST OF APPENDICES

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

Since classification is the philosophy of classical rough set theory, i.e. rough set theory was used mainly to classify objects or to assign them to classes known as a posteriori (Komorowski *et al.*, 1999; Pramanik *et al.*, 2021). Therefore, this thesis focuses on application of rough set theory for data clustering (a priori), particularly, for categorical data clustering.

Data clustering is one of the basic tools available, to understand the structure of the data set (Mesakar and Chaudhari, 2012). The process of grouping a set of physical or abstract objects into classes of similar objects is known as clustering. Clustering algorithms play an important role in machine learning, data mining, information retrieval, web analytics, marketing, medical diagnostics, and pattern recognition. Clustering is often called unsupervised learning task because there is no class that shows the value of a prior clustering given from the data sample, which is the case in supervised learning. General definition of clustering could be "the process of organizing objects into groups whose members are similar in some ways". Therefore, the cluster is a collection of data objects that are similar to each other in a same and distinct cluster with objects in other clusters.

## 1.2 Background of Study

The amount of knowledge in the world cumulatively doubles approximately every 20 months. This information is required for decision-making process, resulting in a cumbersome process due to enormous data. To combat the increase in the volume of information, many tools have been developed in different fields, including retrieval, acquisition, storage and maintenance (Jensen, 2005). Besides, considering the data explosion, various organizations have developed a large volume of databases that can accommodate a large amount of valuable information. However, in recent years, these massive amounts of data in disparate structures have been rapidly overwhelming. Therefore, Database Management System (DBMS) and systematized databases are established (Öztürk, 1999).

An effective DBMS aids in the retrieval of information from a large data corpus. When dealing with large datasets, automated data summarization, pattern identification in raw data, and information extraction aid in enhancing managerial decisions. Scientific data, games data, software engineering data, personal data, digital media data, satellite sensing data, written reports data, medical data, commercial transactions data, virtual worlds data, world wide web repositories data, as well as surveillance video and photographs are just some of the types of data gathered on a regular basis (Figueiredo Filho *et al.*, 2014). Humans cannot effectively examine a large data size, and require a knowledge discovery process, especially Knowledge Discovery in the Databases (KDD) (Düntsch *et al.*, 2000). KDD is a multi-step process that can convert raw data into useful information. Upon conversion, the data are now nontrivial and implicit from the data in databases (Bagga and Singh, 2011).

The KDD process consists of stages that include collecting raw data that will lead to the creation of new knowledge, data selection, data transformation, data cleaning, evaluating patterns, data integration, knowledge representations, and data mining (Keerthi *et al.*, 2002). A data mining task is done to determine the nature of information discovered (Hu *et al.*, 2017). As a result, the best approach to learn about data mining is to get familiar with the types of roles or issues it can solve. Majority of data mining jobs may be classified as descriptive or predictive (Burgos *et al.*, 2018). Descriptive data mining tasks describe the general properties of the existing data, while predictive data mining tasks attempt to make predictions based on available data's inference.

There are pending issues that must be addressed, including data source, interface, mining methodologies, social, security and performance, before data mining can be developed into a conventional and trusted discipline (Rajalakshmi *et al.*, 2010). Data mining functionalities include association analysis, classification, clustering, characterization, discrimination, and prediction, etc. Clustering is the function that focuses on grouping data (objects) into clusters where identical objects are collected within the same cluster, while disparate ones belong to different clusters. There are several cogent reasons to cluster data, with the most important being the building of simpler and more understandable methods that are easily acted upon (Weiss and Davison, 2010). Cluster analysis is among the most extensively employed exploratory data analysis tasks in data mining, with applications in the fields of information retrieval, image processing, web applications and speech processing (Benabdellah *et al.*, 2019). The external validation indices measure the similarity between the output of the clustering algorithm and the unique partitioning of the dataset (Rodriguez *et al.*, 2019).

The different algorithms can be broadly classified into partitioning, hierarchical, density, grid and model-based algorithms (Fahad *et al.*, 2014; Wang *et al.*, 2018). Partitioning-based algorithms specify the initial groups by reallocating them towards a union and all clusters are determined promptly. In hierarchy-based clustering, depending on the medium of proximity the data is organized in a hierarchical manner. Similarly, density-based based algorithms separate the data objects based on their regions of density, boundary and connectivity. Grid based technique divides the space of the data objects into grids. Whereas, in model-based clustering techniques the fit between the given data and some (predefined) mathematical model is optimized (Ali *et al.*, 2017). Many domains like academic result analysis of institutions, machine learning, image mining, medical dataset, software engineering, bioinformatics, information retrieval and pattern recognition uses the core methodology of clustering (Aggarwal, 2014; Figueiredo Filho *et al.*, 2014).

The particular choice of a clustering algorithm also relies tremendously on specific data type. The different data types are textual, discrete sequences, time series, uncertain data, categorical and multimedia data (Kumar and Tripathy, 2009). There are several clustering techniques developed to combine objects of same characteristics, however the implementation of them is challenging due to certain issues like categorical data clustering, handling uncertainty, stability and efficiency issues. Different techniques for clustering data having only numerical values were proposed by (Zhou and Wu, 2008). Unlike numerical data, the multi-valued attributes known as categorical data have common values or common objects and association between both. To deal with categorical data, several clustering algorithms have been developed (Jiang and Liu, 2020). Though, they contributed well to clustering process, but they are not able to handle uncertainty (Pramanik *et al.*, 2021). In many cases where there is no sharp boundary between clusters, the uncertainty becomes an important real-world issue.

Huang, Gupta and Kang (Kim *et al.*, 2004) explored fuzzy sets to handle uncertainty in categorical data clustering. However, to attain the stability and to control the membership fuzziness these algorithms require multiple runs (Naouali *et al.*, 2020b). Pawlak had introduced rough set theory (RST) (Pawlak, 2012), a mathematical tool to deal with vagueness and uncertainty. Many researchers and practitioners are attracted towards RST by contributing essentially to the applications and development in the fields of artificial intelligence, decision support systems, machine learning, knowledge acquisition, decision analysis, pattern recognition, expert systems, cognitive sciences, inductive reasoning, and knowledge discovery from data bases (Pawlak and Skowron, 2007). Many interesting applications, the basic ideas of RST and its extensions can be found in several books, issues of the transactions on rough sets, special issues of other journals, international conferences, proceedings and tutorials (Li *et al.*, 2017). In general, and comparing to other clustering algorithms, the RST is selected in this research due to its simplicity, its capability to deal with uncertain and fuzzy information; it is completely data-driven that does not require any additional information such as fitness for the probability distribution, or function of membership, it does not need special measures such as consistency and distance measures, which resulted in high computational cost.

The RST is a viable system to deal with uncertainty in clustering process of categorical data. RST was originally a symbolic data analysis tool now being developed for cluster analysis (Zhou *et al.*, 2016). In rough categorical clustering, mainly the data set is expressed as the decision table by introducing a decision attribute. Most of these methods assume one or more given partitions of the data set aiming to find a cluster which best represents the data according to some predefined measure. Set approximation and reduct based methods are the two main ideas of the rough set model which are promising for applications. Tolerance rough set clustering (Mingoti and Matos, 2012) and rough-K-Means clustering (Peters and Skowron, 2007) are the examples of set approximation methods. Despite of having satisfactory results, these methods have issues as they depend on several parameters and thresholds (Koç and Koç, 2016). The reduct based methods either work as pre-processing tool or as a tool for cluster generation but the problem of time complexity has not been solved yet (Eskandari and Javidi, 2016).

In RST, a subset of universe can be represented in terms of equivalence classes as clustering of universe. Therefore, RST has been successfully applied for selecting best suitable clustering attribute. The pioneer algorithms to select clustering attribute are developed by (Mazlack *et al.*, 2000) which includes Total Roughness (TR) and Bi-Clustering (BC). These algorithms work on the accuracy of roughness (approximation accuracy average) in the RST. Later on, another rough categorical clustering algorithm named Min-Min Roughness (MMR) was proposed by Parmar *et al*. to improve previous algorithms (Parmar *et al.*, 2010). Despite of MMR's better performance, issues like accuracy, computational complexity and purity are yet to be addressed. In 2010, an algorithm based on the dependency of attributes was introduced by (Herawan and Mat Deris, 2009) named maximum dependency of attributes (MDA) which uses rough set information system for categorical data clustering. Hassanein and Elmelegy in 2013, proposed maximum significance of attributes (MSA) that utilized the RST concept of significance of attributes for selecting clustering attribute (Hassanein and Elmelegy, 2013). Moreover, Park and Choi introduced information-theoretic dependency roughness (ITDR) algorithm (Park and Choi, 2015) which finds the entropy roughness to select the suitable clustering attribute. It is another rough clustering algorithm that uses the information-theoretic dependencies of categorical attributes in information systems. Recently, Uddin *et al* in 2017 introduced an alternative algorithm named maximum indiscernible attribute (MIA) algorithm (Uddin *et al.*, 2017). for clustering categorical data using rough set indiscernible relations is proposed. The novelty of the proposed approach is based on the concept of indiscernibility relation combined with a number of clusters.

Today the world is full of data and everyday people encounter a large amount of information and they store or represent it as data for further analysis and management. One of the vital means in dealing with these data is to classify or group them into a set of categories or clusters. Rough Set Theory (RST) is a powerful mathematical tool proposed by Pawlak (Pawlak and Skowron, 2007) successfully applied to deal with vagueness and uncertainty in data analysis. The concept of rough set theory in this research work is utilized in terms of data in an information system.

Rough set theory has the ability of decision making in the presence of uncertainty and vagueness. Moreover, it can represent a subset of universe in terms of equivalence classes of partition of the universe. Obviously, every subset of attributes induces unique indiscernibility relation which is an equivalence relation and hence, induces unique clustering. This notion of indiscernibility is very attractive, since each indiscernible relation is also a sort of cluster. In this study, the indiscernibility is used as a measure of similarity without any distance function for clustering the objects.

Recently, the problem of clustering categorical data has received much attention in many fields from statistics to psychology. The categorical data unlike numerical data cannot be naturally ordered. Therefore, those clustering algorithms dealing with numerical data cannot be used to cluster categorical data. In addition, very less work has been done for clustering the categorical data. A well-known algorithm for clustering categorical data is using rough set theory (Park and Choi, 2015). Originally the motivation and inspiration for this study came from exploring useful limitations and issues of existing rough categorical clustering algorithms (Mazlack *et al.*, 2000; Parmar *et al.*, 2007; Herawan *et al.*, 2010; Hassanein and Elmelegy, 2013; Park and Choi, 2015; Uddin *et al.*, 2017). This research is conducted in order to come with more general, efficient and better rough categorical clustering algorithms. The MSA, ITDR and MIA algorithms outperformed their previous algorithms such as BC, TR, MMR etc.

Most rough set-based clustering algorithms consider two methods: (i) introducing a condition attribute based on which the dataset is divided to partition the objects, and (ii) evaluating the dataset lower and quality of approximations. All of the previous methods have issues with accuracy, purity, and computational complexity. The limitations and issues of MSA, ITDR, and MIA algorithms on several data sets where those algorithms fail to select or randomly select attributes or struggle to select their best clustering attribute (Naouali *et al.*, 2020a; Naouali *et al.*, 2020b; Salem *et al.*, 2021; Ye and Liu, 2021). Some of the limitations are listed.

1. Accuracy is an issue for MSA, ITDR, and MIA algorithms because they are all primarily determined by the cardinality of lower approximation of an attribute, and partitioning attribute based on approximation of sets on one attribute is highly similar to that induced by other attribute values.

2. The MSA algorithm cannot perform well on data sets with attributes of equal significance value.

3. The MIA algorithm fails to select the clustering attribute for data sets with attributes having an indiscernibility value of zero or equal to zero.

4. Due to the presence of purity measures, ITDR and MIA algorithms face issues like random attribute selection and integrity of clusters.

5. For MSA, ITDR, and MIA algorithms, computation complexity is still an outstanding issue due to the fact that all attributes are considered to be selected and the ever-increasing computing capabilities.

6. Due to the presence of objects of different classes within a cluster, ITDR and MIA cluster purity remain an issue for cluster validity.

## 1.3 Problem Statement

However, one of the main research problems of rough sets is set approximation; existing algorithms struggle to select or fail to select or randomly select their best clustering attribute during the clustering process; and the other is data analysis algorithms. The initial data partitioning influences the quality of the final rough set categorical clustering (Salem *et al.*, 2021; Sun *et al.*, 2011; Zhang *et al.*, 2016; Zhang *et al.*, 2018). To address these issues and problems, it is necessary to propose more appropriate methods-based algorithms to partition the attributes and select the best clustering attribute.

Solving or mitigating these problems of the RST can lead to enhance its performance. Therefore, this study proposed a variety of solutions-based RST algorithms as well as RST itself. For RST- based algorithms, the research has expanded to include the RST in combination with two methods such as attribute partitioning and attribute selection. For the RST itself, this research proposed several extensions, definitions, and proofs to RST to overcome the problem of the approximation of sets and ignoring the attributes in the boundary region.

This thesis arose from the discovery of useful limitations and existing issues in categorical clustering algorithms while searching for an efficient algorithm for categorical data clustering. However, because the main algorithms for categorical data clustering based on rough set theory are relatively new, a robust clustering algorithm that can also handle uncertainty in categorical data clustering is required.

Accordingly in this work, two rough set based categorical clustering methods are proposed. Positive Region Indiscernibility (PRI) for attribute partitioning, and Maximum Mean Partitioning (MMP) for attribute selection, to improve RST categorical clustering algorithms. Furthermore, a proposed RST categorical clustering algorithm, Maximum Partitioning Attribute (MPA), which takes maximal mean partitioning measures into account, necessitates calculating the positive and boundary regions of attributes in an information system. Several propositions and experiments on benchmark data sets show the significance, novelty and contribution of these proposed methods and algorithms to practical systems.

## 1.4 Research Aim and Objectives

The main aim of the research is to propose an enhanced rough set based categorical clustering algorithm using the integration of the attribute partition and attribute selection method. The categorical attributes in RST boundary region are evaluated and the candidate attribute is chosen to reconstruct the positive region that could enhance the performance of RST clustering. For this purpose, the following research objectives are developed:

i. To propose rough set-based attributes partitioning method, Positive Region, based Indiscernibility (PRI), that includes the positive and boundary regions in attributes to reduce the similarity attributes value for selecting partitioning attribute and increasing accuracy of approximation sets.

ii. To propose rough set-based attribute selecting method, Maximum Mean Partitioning (MMP), that speed up selection of the best clustering attributes in order to reduce computational complexity (Iteration and Time).

iii. To propose rough set based categorical clustering algorithm, Maximum Partitioning Attributes (MPA), by integrating PRI and MMP methods, that combines the partitioning attributes with best clustering attribute selected to evaluate their performance and increase cluster purity.

iv. To validate the performance of proposed methods and algorithm on real and benchmarked datasets by comparing them with recent baseline rough categorical clustering algorithms including Maximum Significant Attribute (MSA), Information Theoretic Dependency Roughness (ITDR), Maximum Indiscernibility Attribute (MIA), and classical K-mean clustering algorithms in terms of computational complexity (time and iteration), and purity.

## 1.5 Research Questions

The following research questions have been constructed based on the objectives above:

i. How to address inappropriate attribute partitioning in order to reduce the value set of similarity attributes and increase accuracy?

ii. How can the difficulty to select or failure to select a clustering attribute be addressed in order to reduce computational complexity (requiring fewer iterations and delivering a better response)?

iii. How can cluster validity estimation algorithms for categorical data clustering be improved to maximize cluster purity?

## 1.6 Research Scope and Assumptions

The research falls into the domains of data mining and clustering and aims to develop RST-based categorical clustering methods, namely Positive Region Indiscernibility (PRI) and Maximum Mean Partitioning (MMP) for partitioning and attribute selection, to enhance the RST categorical clustering algorithms. Moreover, a proposed RST categorical clustering algorithm, Maximum Partitioning Attributes (MPA) is also introduced to find a better cluster validity estimation algorithm for the categorical clustering process. The relevant propositions are illustrated to prove the correctness and effectiveness of the proposed algorithms. Twenty-one (21) from the UCI-repository and seven (7) small categorical datasets are considered for experimentation and validation of proposed methods and algorithm.

A real-world supply base management (SBM) dataset is also considered in the experiments. Three existing RST-based categorical clustering algorithms, Maximum Significance Attribute (MSA), Information Theoretic Dependency Roughness (ITDR) and Maximum Indiscernible Attribute (MIA), are used for comparison with proposed PRI, MMP methods, and MPA algorithms in terms of rough accuracy, purity, number of iterations, and response time. Finally, the proposed MPA algorithm is compared to the classical simple K-Mean algorithm on 10 datasets to test and evaluate its performance.

## 1.7 Research Hypothesis Development

The following research hypothesis have been constructed based on the objectives and questions above:

Ho - Null Hypothesis
Ha - Alternate Hypothesis

**Hypothesis 1:**

**Ho -** There is no significant relationship between attributes partitioning and accuracy performance.

**Ha -** There is significant relationship between attributes partitioning and accuracy performance.

**Hypothesis 2:**

**Ho:** There is no significant relationship between faster attributes selection and computational complexity (Iteration and Time) performance.

**Ha:** There is significant relationship between faster attributes selection and computational complexity (Iteration and Time) performance.

**Hypothesis 3:**

**Ho -** There is no significant relationship between number of cluster and purity performance.

**Ha -** There is significant relationship between number of cluster and purity performance.

## 1.8 Research Significance

There are three phases' implications for this thesis. Firstly, a union positive and boundary regions-based dependency measure induces an alternative definition for assessing uncertainty using a rough set for categorical data clustering. Second, an alternative method for selecting a clustering attribute-based rough set is proposed. To settle the increasing computing capabilities, a better selection targeting process was used to select the maximal value of a mean dependency degree as a clustering attribute. Third, domain knowledge on data like rough value set is utilized to develop a RST categorical clustering algorithm integrating the previous methods, and nm cluster purity measurement and validation are presented. All the proposed algorithms show significant improvement for clustering categorical data, not only in terms of accuracy and cluster purity, but also in terms of time taken and number of iterations. Furthermore, an application of the proposed methods and algorithm for clustering supplier chain management is presented. Discussion and analysis of the results of the proposed method and algorithms will be provided in detail later.

## 1.9 Thesis Organization

The remaining chapters of the research are organised as follows:

Chapter 2 or literature review discusses some fundamental concepts and overview of existing works on clustering categorical data using RST. It comprises of an information system notion in rough relational database, an indiscernibility relation, set approximations and rough set based categorical clustering algorithm. Moreover, it also presents analysis, limitations, and examples of some existing rough for clustering categorical data algorithms.

Chapter 3 presents the research methodology. The suggested clustering-based methods for categorical data, namely Positive Region-based Indiscernibility (PRI), Maximum Mean Partitioning (MMP) and Maximum Partitioning Attributes (MPA) methods, are discussed. Aside from that, basic info on partitioning, attribute selection and categorical clustering algorithm using RST and set cardinality value are also discussed. The evaluation metrics applied in this study are also described. Multiple suggestions and instances are provided to indicate the significance of suggested algorithms and approaches.

Chapter 4 portrays the outcomes of studies on recommended PRI method. Empirical research on three small UCI-repository benchmark datasets demonstrates the performance of the recommended method. Furthermore, outcomes from this study are compared with results from the latest and prominent rough set algorithms for clustering categorical data. All the experimental outcomes are deliberated and examined in detail by illustrating them in graph and tabulation forms.

Chapter 5 provides the outcomes of the research on recommended MMP method. Empirical research on three small UCI-repository benchmark datasets demonstrates the performance of the suggested method. Moreover, outcomes of this study are compared with results from latest and prominent rough set algorithms for clustering categorical data. All the experimental outcomes are deliberated and examined in detail by depicting them in the forms of graph and tabulation.

Chapter 6 analyses the outcomes of experiments on the suggested MPA algorithm. Empirical research on UCI-repository benchmark datasets and a real SBM dataset portrays the performance of the suggested algorithm. Comparison with the latest and prominent rough set algorithms for clustering categorical data will also be implemented. All the experimental outcomes are deliberated and examined in detail by depicting them in graph and tabulation forms.

Finally, Chapter 7 provides closing remarks, recommendations, and suggestions for future works.

# REFERENCES


Aborujilah, A., Musa, S., Shahzad, A., Nazri, M., and Alsharafi, A. (2013). Flooding Based DoS Attack Feature Selection Using Remove Correlated Attributes Algorithm. Paper presented at the 2013 International Conference on Advanced Computer Science Applications and Technologies, 93-96.

Abu-Donia, H. (2013). New rough set approximation spaces. Paper presented at the Abstract and Applied Analysis.

Aggarwal, C. C. (2014). Data classification: algorithms and applications: CRC press.

Aggarwal, C. C., and Clustering, C. R. D. (2014). Algorithms and Applications: CRC Press Taylor and Francis Group.

Aggarwal, C. C., and Reddy, C. K. (2014). Data clustering. Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra.

Ahmad, A., and Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering, 63(2), 503-527.

Ali, D. S., Ghoneim, A., and Saleh, M. (2017). Data clustering method based on mixed similarity measures. Paper presented at the International Conference on Operations Research and Enterprise Systems, 192-199.

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information retrieval, 12(4), 461-486.

Amini, A., Wah, T. Y., and Saboohi, H. (2014). On density-based data streams clustering algorithms: A survey. Journal of Computer Science and Technology, 29(1), 116-141.

Anderberg, M. R. (1973). The broad view of cluster analysis. Cluster analysis for applications, 1(1), 1-9.

Azar, A. T., Inbarani, H. H., and Devi, K. R. (2017). Improved dominance rough set-based classification system. Neural Computing and Applications, 28(8), 2231-2246.

Bagga, S., and Singh, G. (2011). Three Phase Iterative Model of KDD. International Journal of Information Technology, 4(2), 695-697.

Beale, M. H., Hagan, M. T., and Demuth, H. B. (2010). Neural network toolbox™ user's guide. The MathWorks.

Beaubouef, T., Petry, F. E., and Arora, G. (1998). Information-theoretic measures of uncertainty for rough sets and rough relational databases. Information Sciences, 109(1-4), 185-195.

Bell, D. A., Guan, J., and Liu, D. (2005). Mining association rules with rough sets. In Intelligent data mining (pp. 163-184): Springer.

Benabdellah, A. C., Benghabrit, A., and Bouhaddou, I. (2019). A survey of clustering algorithms for an industrial context. Procedia computer science, 148, 291-302.

Bezdek, J. C., and Pal, N. R. (1998). Some new indexes of cluster validity. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 28(3), 301-315.

Bi, Y., Anderson, T., and McClean, S. (2003). A rough set model with ontologies for discovering maximal association rules in document collections. Knowledge-Based Systems, 16(5-6), 243-251.

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., and Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Computers & Electrical Engineering, 66, 541-556.

Bryant, D., Danziger, P., and Pettersson, W. (2015). Bipartite 2-factorizations of complete multipartite graphs. Journal of Graph Theory, 78(4), 287-294.

Bi, Y., Anderson, T., and McClean, S. (2003). A rough set model with ontologies for discovering maximal association rules in document collections. Knowledge-Based Systems, 16(5-6), 243-251.

Caruso, G., Gattone, S. A., Fortuna, F., and Di Battista, T. (2017). Cluster analysis as a decision-making tool: a methodological review. Paper presented at the International Symposium on Distributed Computing and Artificial Intelligence, 48-55.

Cheng, Y. C. (2013). School effectiveness and school-based management: A mechanism for development: Routledge.

Chowdhury, M., Abawajy, J., Kelarev, A., and Jelinek, H. (2016). A clustering-based multi-layer distributed ensemble for neurological diagnostics in cloud services. IEEE Transactions on Cloud Computing.

Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., and Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), e01802.

Dai, J., Wang, W., Xu, Q., and Tian, H. (2012). Uncertainty measurement for interval-valued decision systems based on extended conditional entropy. Knowledge-Based Systems, 27, 443-450.

Danziger, P. (2010). Big o notation. Source internet: http://www. scs. ryerson. ca/mth110/Handouts/PD/bigO. pdf, Retrieve: April.

Das, S., Abraham, A., and Konar, A. (2008). Automatic kernel clustering with a multi-elitist particle swarm optimization algorithm. Pattern recognition letters, 29(5), 688-699.

Davey, J., and Burd, E. (2000). Evaluating the suitability of data clustering for software remodularisation. Paper presented at the Proceedings Seventh Working Conference on Reverse Engineering, 268-276.

Ding, B., Zheng, Y., and Zang, S. (2009). A new decision tree algorithm based on rough set theory. Paper presented at the 2009 Asia-Pacific Conference on Information Processing, 326-329.

Dubois, C., Quinif, Y., Baele, J.-M., Barriquand, L., Bini, A., Bruxelles, L., et al. (2014). The process of ghost-rock karstification and its role in the formation of cave systems. Earth-Science Reviews, 131, 116-148.

Düntsch, I., Gediga, G., and Nguyen, H. S. (2000). Rough set data analysis in the KDD process. Paper presented at the Proc. of IPMU, 220-226.

Dutta, M., Mahanta, A. K., and Pujari, A. K. (2005). QROCK: A quick version of the ROCK algorithm for clustering of categorical data. Pattern Recognition Letters, 26(15), 2364-2373.

Düntsch, I., and Gediga, G. (2015). Rough set clustering. In Handbook of Cluster Analysis (pp. 596-613): Chapman and Hall/CRC.

Eskandari, S., and Javidi, M. M. (2016). Online streaming feature selection using rough sets. International Journal of Approximate Reasoning, 69, 35-57.

Everitt, B. S., and Dunn, G. (2001). Principal components analysis. Applied multivariate data analysis, 48-73.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., et al. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing, 2(3), 267-279.

Figueiredo Filho, D. B., da Rocha, E. C., da Silva Júnior, J. A., Paranhos, R., da Silva, M. B., and Duarte, B. S. F. (2014). Cluster analysis for political scientists. Applied Mathematics, 2014.

Ganti, V., Gehrke, J., and Ramakrishnan, R. (1999). CACTUS—clustering categorical data using summaries. Paper presented at the Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 73-83.

Gibson, D., Kleinberg, J., and Raghavan, P. (2000). Clustering categorical data: An approach based on dynamical systems. The VLDB Journal, 8(3-4), 222-236.

Gondek, D., and Hofmann, T. (2007). Non-redundant data clustering. Knowledge and Information Systems, 12(1), 1-24.

Grace, G. H., and Desikan, K. (2015). Experimental estimation of number of clusters based on cluster quality. arXiv preprint arXiv:1503.03168.

Guan, J., Bell, D. A., and Liu, D. (2003). The rough set approach to association rule mining. Paper presented at the Third IEEE International Conference on Data Mining, 529-532.

Guha, S., Rastogi, R., and Shim, K. (1998). CURE: an efficient clustering algorithm for large databases. ACM Sigmod record, 27(2), 73-84.

Guha, S., Rastogi, R., and Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. Information systems, 25(5), 345-366.

Haimov, S., Michalev, M., Savchenko, A., and Yordanov, O. (1989). Classification of radar signatures by autoregressive model fitting and cluster analysis. IEEE Transactions on Geoscience and Remote Sensing, 27(5), 606-610.

Han, J., Kamber, M., and Pei, J. (2011). Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 5(4), 83-124.

Hassanein, W., and Elmelegy, A. (2013). An algorithm for selecting clustering attribute using significance of attributes. International Journal of Database Theory & Application, 6(5), 53-66.

He, Z., Xu, X., and Deng, S. (2008). k-ANMI: A mutual information based clustering algorithm for categorical data. Information Fusion, 9(2), 223-233.

Herawan, T., Deris, M. M., and Abawajy, J. H. (2010). A rough set approach for selecting clustering attribute. Knowledge-Based Systems, 23(3), 220-231.

Herawan, T., and Mat Deris, M. (2009). Rough set theory for selecting clustering attribute. Paper presented at the AIP Conference Proceedings, 331-338.

Hu, Z., Bodyanskiy, Y. V., Tyshchenko, O. K., and Samitova, V. O. (2017). Possibilistic fuzzy clustering for categorical data arrays based on frequency prototypes and dissimilarity measures. International Journal of Intelligent Systems and Applications, 9(5), 55.

Huang, S.-Y. (1992). Intelligent decision support: handbook of applications and advances of the rough sets theory (Vol. 11): Springer Science & Business Media.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3), 283-304.

Hui, T., Short, T., Hong, W., Suen, T., Gin, T., and Plummer, J. (1995). Additive interactions between propofol and ketamine when used for anesthesia induction in female patients. The Journal of the American Society of Anesthesiologists, 82(3), 641-648.

Jabbar, A., and Zainudin, S. (2014). Water cycle algorithm for attribute reduction problems in rough set theory. Journal of Theoretical and Applied Information Technology, 61(1), 107-117.

Jensen, R. (2005). Combining rough and fuzzy sets for feature selection. Citeseer.

Jia, X., Rao, Y., Shang, L., and Li, T. (2020). Similarity-based attribute reduction in rough set theory: a clustering perspective. International Journal of Machine Learning and Cybernetics, 11(5), 1047-1060.

Jiang, Z., and Liu, X. (2020). A Novel Consensus Fuzzy K-Modes Clustering Using Coupling DNA-Chain-Hypergraph P System for Categorical Data. Processes, 8(10), 1326.

Karol, S., and Mangat, V. (2013). Evaluation of text document clustering approach based on particle swarm optimization. Open Computer Science, 3(2), 69-90.

Kaufman, L., and Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis: John Wiley & Sons.

Kavzoglu, T. (2009). Increasing the accuracy of neural network classification using refined training data. Environmental Modelling & Software, 24(7), 850-858.

Keerthi, S. S., Ong, C. J., Siah, K. B., Lim, D. B., Chu, W., Shi, M., et al. (2002). A machine learning approach for the curation of biomedical literature: KDD Cup 2002 (task 1). ACM SIGKDD Explorations Newsletter, 4(2), 93-94.

Kim, D.-W., Lee, K. H., and Lee, D. (2004). Fuzzy clustering of categorical data using fuzzy centroids. Pattern recognition letters, 25(11), 1263-1271.

Koç, D. İ., and Koç, M. L. (2016). Fuzzy viscometric analysis of polymer–polymer miscibility based on fuzzy regression. Chemometrics and Intelligent Laboratory Systems, 157, 58-66.

Komorowski, J., Pawlak, Z., Polkowski, L., and Skowron, A. (1999). Rough sets: A tutorial. Rough fuzzy hybridization: A new trend in decision-making, 3-98.

Krızek, P. (2008). Feature selection: stability, algorithms, and evaluation. Ph. d. thesis, Czech Technical University in Prague.

Kumar, P., and Tripathy, B. (2009). MMeR: an algorithm for clustering heterogeneous data using rough set theory. International Journal of Rapid Manufacturing, 1(2), 189-207.

Kumar, R., and Sawant, K. (2010). On the design of inscribed triangle non-concentric circular fractal antenna. Microwave and Optical Technology Letters, 52(12), 2696-2699.

Knote, R., Janson, A., Söllner, M., and Leimeister, J. M. (2019). Classifying smart personal assistants: an empirical cluster analysis. Paper presented at the Proceedings of the 52nd Hawaii international conference on system sciences.

Lakshmi, B. J., Shashi, M., and Madhuri, K. (2017). A rough set based subspace clustering technique for high dimensional data. Journal of King Saud University-Computer and Information Sciences.

Lange, M. D., Jia, X., Parisot, S., Leonardis, A., Slabaugh, G., and Tuytelaars, T. (2020). Unsupervised model personalization while preserving privacy and scalability: An open problem. Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14463-14472.

Leung, Y., Fischer, M. M., Wu, W.-Z., and Mi, J.-S. (2008). A rough set approach for the discovery of classification rules in interval-valued information systems. Available at SSRN 2898125.

Li, M., Deng, S., Wang, L., Feng, S., and Fan, J. (2014). Hierarchical clustering algorithm for categorical data using a probabilistic rough set model. Knowledge-Based Systems, 65, 60-71.

Li, Z., Xie, N., and Gao, N. (2017). Rough approximations based on soft binary relations and knowledge bases. Soft Computing, 21(4), 839-852.

Liang, J., Wang, J., and Qian, Y. (2009). A new measure of uncertainty based on knowledge granulation for rough sets. Information Sciences, 179(4), 458-470.

Lichman, M. (2013). UCI machine learning repository, 2013.

Liu, H., and Ban, X.-j. (2015). Clustering by growing incremental self-organizing neural network. Expert Systems with Applications, 42(11), 4965-4981.

Lu, L., Li, G. Y., Swindlehurst, A. L., Ashikhmin, A., and Zhang, R. (2014). An overview of massive MIMO: Benefits and challenges. IEEE journal of selected topics in signal processing, 8(5), 742-758.

Mac Parthaláin, N., and Shen, Q. (2009). Exploring the boundary region of tolerance rough sets for feature selection. Pattern recognition, 42(5), 655-667.

Mandal, P., and Ranadive, A. (2019). Fuzzy multi-granulation decision-theoretic rough sets based on fuzzy preference relation. Soft Computing, 23(1), 85-99.

Manning, E. (2009). Relationscapes: Movement, Art.

Mazlack, L. J., He, A., and Zhu, Y. (2000). A rough set approach in choosing partitioning attributes. Paper presented at the Proceedings of the ISCA 13th International Conference (CAINE-2000.

Mesakar, S. S., and Chaudhari, M. (2012). Review Paper On Data Clustering Of Categorical Data. International Journal of Engineering Research & Technology, 1(10).

Mingoti, S. A., and Matos, R. A. (2012). Clustering algorithms for categorical data: a Monte Carlo study. Int J Stat Appl, 2, 24-32.

Mogotsi, I. (2010). Christopher d. manning, prabhakar raghavan, and hinrich schütze: Introduction to information retrieval: Springer.

Naouali, S., Ben Salem, S., and Chtourou, Z. (2020a). Clustering categorical data: A survey. International Journal of Information Technology & Decision Making, 19(01), 49-96.

Naouali, S., Salem, S. B., and Chtourou, Z. (2020b). Uncertainty mode selection in categorical clustering using the rough set theory. Expert Systems with Applications, 158, 113555.

Nies, H. W., Zakaria, Z., Mohamad, M. S., Chan, W. H., Zaki, N., Sinnott, R. O., et al. (2019). A Review of Computational Methods for Clustering Genes with Similar Biological Functions. Processes, 7(9), 550.

Öztürk, M. (1999). Tefsirde Zâhir-Bâtın Düalizmi ya da Tasavvufî Aşırı Yorum. İslâmiyât Dergisi, 3, 101-120.

Pacheco, F., Cerrada, M., Sánchez, R.-V., Cabrera, D., Li, C., and de Oliveira, J. V. (2017). Attribute clustering using rough set theory for feature selection in fault severity classification of rotating machinery. Expert Systems with Applications, 71, 69-86.

Park, I.-K., and Choi, G.-S. (2015). Rough set approach for clustering categorical data using information-theoretic dependency measure. Information Systems, 48, 289-295.

Parmar, D., Wu, T., and Blackhurst, J. (2007). MMR: an algorithm for clustering categorical data using rough set theory. Data & Knowledge Engineering, 63(3), 879-893.

Parmar, D., Wu, T., Callarman, T., Fowler, J., and Wolfe, P. (2010). A clustering algorithm for supplier base management. International Journal of Production Research, 48(13), 3803-3821.

Pawlak, Z. (1982). Rough sets. International journal of computer & information sciences, 11(5), 341-356.

Pawlak, Z. (1995). Vagueness and uncertainty: a rough set perspective. Computational intelligence, 11(2), 227-232.

Pawlak, Z. (1997). Rough set approach to knowledge-based decision support. European journal of operational research, 99(1), 48-57.

Pawlak, Z. (2002). Rough sets and intelligent data analysis. Information sciences, 147(1-4), 1-12.

Pawlak, Z. (2012). Rough sets: Theoretical aspects of reasoning about data (Vol. 9): Springer Science & Business Media.

Pawlak, Z., Grzymala-Busse, J., Slowinski, R., and Ziarko, W. (1995). Rough sets. Communications of the ACM, 38(11), 88-95.

Pawlak, Z., and Skowron, A. (2007). Rudiments of rough sets. Information sciences, 177(1), 3-27.

Peters, J. F., and Skowron, A. (2007). Zdzisław Pawlak life and work (1926-2006). Information Sciences, 177(1), 1-2.

Prabha, K. A., and Visalakshi, N. K. (2014). Improved particle swarm optimization based k-means clustering. Paper presented at the 2014 International Conference on Intelligent Computing Applications, 59-63.

Pramanik, A., Sarkar, S., Maiti, J., and Mitra, P. (2021). RT-GSOM: Rough tolerance growing self-organizing map. Information Sciences, 566, 19-37.

Preeti, L., Magesh, K., Rajkumar, K., and Karthik, R. (2011). Recurrent aphthous stomatitis. Journal of oral and maxillofacial pathology: JOMFP, 15(3), 252.

Rafsanjani, M. K., Varzaneh, Z. A., and Chukanlo, N. E. (2012). A survey of hierarchical clustering algorithms. The Journal of Mathematics and Computer Science, 5(3), 229-240.

Rajalakshmi, M., Purusothaman, T., and Pratheeba, S. (2010). Collusion-free privacy preserving data mining. International Journal of Intelligent Information Technologies (IJIIT), 6(4), 30-45.

Reddy, H. V., Agrawal, P., and Raju, S. V. (2013). Data labeling method based on cluster purity using relative rough entropy for categorical data clustering. Paper presented at the 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 500-506.

Rissino, S., and Lambert-Torres, G. (2009). Rough set theory—fundamental concepts, principals, data extraction, and applications. In Data mining and knowledge discovery in real life applications: IntechOpen.

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F., et al. (2019). Clustering algorithms: A comparative approach. PloS one, 14(1), e0210236.

Roiger, R. J. (2017). Data mining: a tutorial-based primer: Chapman and Hall/CRC.

Roiger, R. J., and Geatz, M. (2003). Data Mining: A Tutorial-based Primer, Pearson Education. Inc: USA.

Sagi, O., and Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 8(4), e1249.

Saitta, L., and Neri, F. (1998). Learning in the “real world”. Machine learning, 30(2-3), 133-163.

Salamo, M., and Lopez-Sanchez, M. (2011). Rough set based approaches to feature selection for case-based reasoning classifiers. Pattern Recognition Letters, 32(2), 280-292.

Salem, S. B., Naouali, S., and Chtourou, Z. (2021). A rough set based algorithm for updating the modes in categorical clustering. International Journal of Machine Learning and Cybernetics, 1-22.

Salem, S. B., Naouali, S., and Sallami, M. (2017). A computational cost-effective clustering algorithm in multidimensional space using the manhattan metric: application to the global terrorism database. Paper presented at the ICMLA, 14th.

San, O. M., Huynh, V.-N., and Nakamori, Y. (2004). An alternative extension of the k-means algorithm for clustering categorical data. International journal of applied mathematics and computer science, 14, 241-247.

Senan, N., Ibrahim, R., Nawi, N. M., Yanto, I. T. R., and Herawan, T. (2011). Rough set approach for attributes selection of traditional Malay musical instruments sounds classification. Paper presented at the International Conference on Ubiquitous Computing and Multimedia Applications, 509-525.

Soliman, O. S., Hassanien, A. E., and El-Bendary, N. (2011). A rough clustering algorithm based on entropy information. paper presented at the soft computing models in industrial and environmental applications, 6th International Conference SOCO 2011, 213-222.

Sripada, S. C., and Rao, M. S. (2011). Comparison of purity and entropy of k-means clustering and fuzzy c means clustering. Indian journal of computer science and engineering, 2(3), 343-346.

Stańczyk, U., and Zielosko, B. (2020). Heuristic-based feature selection for rough set approach. International Journal of Approximate Reasoning, 125, 187-202.

Suhirman, S. (2017). An application of rough set theory to cluster student assessment at universities. JIKO (Jurnal Informatika dan Komputer), 2(1).

Sun, Z., Ye, Y., Deng, W., and Huang, Z. (2011). A cluster tree method for text categorization. Procedia Engineering, 15, 3785-3790.

Tian, Z., Li, Y., Li, L., Liu, X., Zhang, H., Zhang, X., et al. (2018). 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, 32(9), 824-829.

Tripathy, B., and Ghosh, A. (2011). SDR: An algorithm for clustering categorical data using rough set theory. Paper presented at the 2011 IEEE Recent Advances in Intelligent Computational Systems, 867-872.

Tripathy, B., Goyal, A., Chowdhury, R., and Patra, A. S. (2017). MMeMeR: An algorithm for clustering heterogeneous data using rough set theory. International Journal of Intelligent Systems and Applications, 9(8), 25.

Uddin, J., Ghazali, R., and Deris, M. M. (2017). An empirical analysis of rough set categorical clustering techniques. PloS one, 12(1).

Van der Walt, C. M., and Barnard, E. (2006). Data characteristics that determine classifier performance.

Vidhya, K., and Geetha, T. (2017). Rough set theory for document clustering: A review. Journal of Intelligent & Fuzzy Systems, 32(3), 2165-2185.

Wang, J., Zhu, C., Zhou, Y., Zhu, X., Wang, Y., and Zhang, W. (2017). From partition-based clustering to density-based clustering: Fast find clusters with diverse shapes and densities in spatial databases. IEEE Access, 6, 1718-1729.

Wang, P., Wu, P., Wang, J., Chi, H.-L., and Wang, X. (2018). A critical review of the use of virtual reality in construction engineering education and training. International journal of environmental research and public health, 15(6), 1204.

Wang, W., Gao, W., Wang, C., and Li, J. (2013). An improved algorithm for CART based on the rough set theory. Paper presented at the 2013 Fourth Global Congress on Intelligent Systems, 11-15.

Wang, Y., and Zhang, N. (2014). Uncertainty analysis of knowledge reductions in rough sets. The Scientific World Journal, 2014.

Wang, Z., Yue, H., and Deng, J. (2019). An Uncertainty Measure Based on Lower and Upper Approximations for Generalized Rough set Models. Fundamenta Informaticae, 166(3), 273-296.

Watt, C., Mitchell, S., and Salewski, V. (2010). Bergmann's rule; a concept cluster? Oikos, 119(1), 89-100.

Weiss, G. M., and Davison, B. D. (2010). Data mining. Paper presented at the TO APPEAR IN THE HANDBOOK OF TECHNOLOGY MANAGEMENT, H. BIDGOLI (ED.).

Wong, K.-P., Feng, D., Meikle, S. R., and Fulham, M. J. (2002). Segmentation of dynamic PET images using cluster analysis. IEEE Transactions on nuclear science, 49(1), 200-207.

Wu, J., Hassan, A. E., and Holt, R. C. (2005). Comparison of clustering algorithms in the context of software evolution. Paper presented at the 21st IEEE International Conference on Software Maintenance (ICSM'05), 525-535.

Wu, J., Xiong, H., and Chen, J. (2009). Adapting the right measures for k-means clustering. Paper presented at the Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 877-886.

Xia, S., Li, W., Wang, G., Gao, X., Zhang, C., and Giem, E. (2020). LRA: an accelerated rough set framework based on local redundancy of attribute for feature selection. arXiv preprint arXiv:2011.00215.

Yang, K., and Zhang, N. (2013). Structure and Cluster Analysis on Microblog User’s Relationship Networks. Complex Systems and Complexity Science, 2.

Yanto, I. T. R., Ismail, M. A., and Herawan, T. (2016). A modified Fuzzy k-Partition based on indiscernibility relation for categorical data clustering. Engineering Applications of Artificial Intelligence, 53, 41-52.

Yao, J., and Herbert, J. P. (2007). Web-based support systems with rough set analysis. Paper presented at the International Conference on Rough Sets and Intelligent Systems Paradigms, 360-370.

Yao, Y. (1996). Two views of the theory of rough sets in finite universes. International journal of approximate reasoning, 15(4), 291-317.

Yao, Y. (2003). Probabilistic approaches to rough sets. Expert systems, 20(5), 287-297.

Ye, T., and Liu, B. (2021). Uncertain hypothesis test with application to uncertain regression analysis. Fuzzy Optimization and Decision Making, 1-18.

YIN, B., and HE, S.-h. (2008). Fuzzy K-Prototypes clustering based on particle swarm optimization [J]. Computer Engineering and Design, 11.

Zadeh, L. A. (1996). Fuzzy sets. In Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh (pp. 394-432): World Scientific.

Zaki, M. J., and Peters, M. (2005). CLICKS: Mining subspace clusters in categorical data via K-partite maximal cliques. Paper presented at the 21st International Conference on Data Engineering (ICDE'05), 355-356.

Zhang, L., Li, Y., Sun, C., and Nadee, W. (2013). Rough set based approach to text classification. Paper presented at the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 245-252.

Zhang, Q., Xie, Q., and Wang, G. (2016). A survey on rough set theory and its applications. CAAI Transactions on Intelligence Technology, 1(4), 323-333.

Zhang, Q., Zhao, F., and Yang, J. (2018). The uncertainty analysis of vague sets in rough approximation spaces. IEEE Access, 7, 383-395.

Zhou, H.-B., Wang, J.-L., Jiang, W., Lu, G.-H., Aguiar, J., and Liu, F. (2016). Electrophobic interaction induced impurity clustering in metals. Acta Materialia, 119, 1-8.

Zhou, L., and Wu, W.-Z. (2008). On generalized intuitionistic fuzzy rough approximation operators. Information Sciences, 178(11), 2448-2465.

# List of publications

## Indexed Journal with Impact Factor

Baroad, M. M., Hashim, S. Z. M., Ahsan, J. U., and Zainal, A. (2021). Efficient Scheme to Measure the Crispiness of the Partitioning in Data Mining. Big Data https://www.liebertpub.com/loi/big (Accepted), (Indexed by SCOPUS and SCIE), **(SJR Q2. IF:3.05).**

Baroad, M. M., Hashim, S. Z. M., Ahsan, J. U., and Zainal, A. (2022). An Algorithm to Exploring Rough Set Boundary Region for Categorical data Clustering. International Journal of Speech Technology, https://www.springer.com/journal/10772 (Accepted), (Indexed by SCOPUS and ISI), **(SJR Q2. IF:2.03).**

## Indexed Journal without Impact Factor

Baroad, M. M., Hashim, S. Z. M., Ahsan, J. U., and Zainal, A. (2020). Positive region: An enhancement of partitioning attribute based rough set for categorical data. Periodicals of Engineering and Natural Sciences, 8(4), 2424-2439. http://pen.ius.edu.ba/index.php/pen/article/view/1745. (Indexed by SCOPUS), **(SJR Q2. IF:0).**

Baroud, M. M., Hashim, S. Z. M., Ahsan, J. U., Zainal, A., and Khalaff, H. (2020). Fast attribute selection based on the rough set boundary region. Periodicals of Engineering and Natural Sciences, 8(4), 2575-2587, http://pen.ius.edu.ba/index.php/pen/article/view/1761. (Indexed by SCOPUS), **(SJR Q2. IF:0).**

**Indexed Conference Proceedings**

Baroud, M. M. J., Hashim, S. Z. M., Zainal, A and Jamilah Ahmad. (2019). A New Algorithm-based Rough Set for Selecting Clustering Attribute in Categorical Data. 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE explorer https://ieeexplore.ieee.org/document/9074483. **(Indexed by SCOPUS)**

**Non-Indexed Conference Proceedings**

Baroud, M. M. J., Hashim, S. Z. M., and Zainal, A. (2019). Rough Set-Algorithm for Clustering Categorical Data Using Mean Attribute (MMA) Dependency Based Measure. 8th International Conference on Advances in Computing, Electronics and Communication, https://www.seekdl.org/conferences/paper/details/10067.html.