# EMPIRICAL ANALYSIS OF ROUGH SET CATEGORICAL CLUSTERING TECHNIQUES BASED ON ROUGH PURITY AND VALUE SET

JAMAL UDDIN

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

# EMPIRICAL ANALYSIS OF ROUGH SET CATEGORICAL CLUSTERING TECHNIQUES BASED ON ROUGH PURITY AND VALUE SET

JAMAL UDDIN

A thesis submitted in
fulfillment of the requirement for the award of the
Doctor of Philosophy in Information Technology

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

AUGUST 2017

I would like to dedicate my this PhD thesis to my beloved parents whose sincere prayers make it possible for me to be a successful computer and mathematics researcher. May Allah always bless them.

## ACKNOWLEDGEMENT

Surely All praise is for Allah Almighty Who is the creator of this universe and Darood and Salam upon the Holy prophet Hazrat Muhammad (PBUH). Thanks to Allah Almighty Who enabled me to complete this research thesis with the continuous guidance and sincere cooperation of my kind supervisor Prof. Madya Dr Rozaida Binti Ghazali and my co-supervisor Prof. Dr. Mustafa Bin Mat Deris. In fact I learnt a lot from my worthy supervisors whose valuable suggestions, constructive comments, thought provoking ideas gave me this chance to make my research work a success with in a stipulated period. They are real role model and mentor for me.

I am also thankful to Allah Almighty Who bestowed upon me highly talented and sincere teachers, whose selfless guidance and tireless effort gave me opportunities and chances to complete this valuable research work. I extend my heartiest gratitude to my teachers especially the dean of faculty Prof. Madya Dr. Nazri Bin Mohd Nawi and Prof. Madya Dr Rathiah Binti Hashim for their sincere support, valuable comments and encouraging attitude. I am thankful to all my worthy teachers from the core of my heart they all remained cooperative honestly throughout my research work.

I will never forget the educational facilities and research oriented environment provided by the Faculty of Computer Science and Information Technology (FSKTM) and the Universiti Tun Hussien Onn Malaysia (UTHM). The sincere and continuous efforts of UTHM staff and administration to make available all modern and latest facilities to impart quality education in all fields are remarkable. It was their sincere efforts and approach that has made us able to learn information technology (IT) research and complete the research work under the guidance of able IT researchers, who are renowned in Malaysia and outside of the country in their field.

At last, it is a little sad to say that I could not get time for my old parents to serve them, but it is a fact that their prayers always remain with me and are still there with me throughout my education and research work. Their sincere prayers and advice helped me in every step of my life. Their prayers console me very much and I dedicate

this research work to them. I am really thankful to my sisters, brother, wife and kids whose best wishes and prayers always remain with me. I would also like to pay my best regard to my research fellow Rashid Naseem, my colleagues and all friends whose best wishes and prayers always remained with me throughout my research work. I would also like to thank Dr. Muhammad Imran who develop the uthmthesis LaTeX project for making the thesis writing process a lot easier for me.

## ABSTRACT

Clustering a set of objects into homogeneous groups is a fundamental operation in data mining. Recently, attention has been put on categorical data clustering, where data objects are made up of non-numerical attributes. The implementation of several existing categorical clustering techniques is challenging as some are unable to handle uncertainty and others have stability issues. In the process of dealing with categorical data and handling uncertainty, the rough set theory has become well-established mechanism in a wide variety of applications including databases. The recent techniques such as Information-Theoretic Dependency Roughness (ITDR), Maximum Dependency Attribute (MDA) and Maximum Significance Attribute (MSA) outperformed their predecessor approaches like Bi-Clustering (BC), Total Roughness (TR), Min-Min Roughness (MMR), and standard-deviation roughness (SDR). This work explores the limitations and issues of ITDR, MDA and MSA techniques on data sets where these techniques fails to select or faces difficulty in selecting their best clustering attribute. Accordingly, two alternative techniques named Rough Purity Approach (RPA) and Maximum Value Attribute (MVA) are proposed. The novelty of both proposed approaches is that, the RPA presents a new uncertainty definition based on purity of rough relational data base whereas, the MVA unlike other rough set theory techniques uses the domain knowledge such as value set combined with number of clusters (NoC). To show the significance, mathematical and theoretical basis for proposed approaches, several propositions are illustrated. Moreover, the recent rough categorical techniques like MDA, MSA, ITDR and classical clustering technique like simple K-mean are used for comparison and the results are presented in tabular and graphical forms. For experiments, data sets from previously utilized research cases, a real supply base management (SBM) data set and UCI repository are utilized. The results reveal significant improvement by proposed techniques for categorical clustering in terms of purity (21%), entropy (9%), accuracy (16%), rough accuracy (11%), iterations (99%) and time (93%).

# ABSTRAK

Pengelompokan satu set objek ke dalam kumpulan homogen adalah operasi asas dalam perlombongan data. Kebelakangan ini, perhatian banyak diberikan kepada pengelompokan data berasaskan kategori, iaitu objek data terdiri daripada atribut bukan berangka. Kebanyakan perlaksanaan teknik pengelompokan berasaskan kategori sedia ada adalah mencabar kerana sebahagiannya tidak dapat mengendalikan isu-isu ketidakpastian dan mempunyai masalah kestabilan. Dalam proses berurusan dengan data berasaskan kategori dan pengendalian ketidakpastian, teori set kasar telah menjadi mekanisme yang mantap dalam pelbagai aplikasi termasuk pangkalan data. Kategori set kasar berdasarkan teknik pengelompokan data seperti Teori-Informatik Bersandarkan Kekasaran (ITDR), Atribut Bersandarkan Maksimum (MDA) dan Atribut Signifikan Maksimum (MSA) telah mengatasi teknik-teknik terdahulu seperti Dwi-Kelompok (BC), Kekasaran Mutlak (TR) , Kekasaran Min-Min (MMR), dan Kekasaran Sisihan-Piawai (SDR). Kajian ini membentangkan kekangan dan isu-isu bagi teknik-teknik ITDR, MDA dan MSA ke atas set data tertentu di mana teknik-teknik ini gagal untuk memilih atau menghadapi kesukaran untuk memilih kelompok atribut yang terbaik. Selanjutnya, dua teknik alternatif yang dinamakan Pendekatan Ketulenan Kasar (RPA) dan Attribut Nilai Maksima (MVA) bagi mengkelompokkan data berasaskan kategori telah dicadangkan. Pembaharuan bagi kedua-dua teknik yang telah dicadangkan ini adalah berikut; mencadangkan definisi ketidakpastian baharu berdasarkan ketulenan bagi kekasaran pangkalan data hubungan, manakala MVA berbeza dengan teknik teori set kasar lain, yang mana teknik ini menggunakan pengetahuan domain seperti set nilai yang bergabung dengan beberapa kelompok (NoC) dalam memilih kelompok atribut yang terbaik. Bagi menunjukkan signifikasinya, asas matematik dan teori bagi pendekatan yang dicadangkan, beberapa cadangan telah digambarkan. Selain itu, teknik-teknik berasaskan kategori yang terkini seperti MDA, MSA, ITDR dan teknik pengelompokan klasik seperti K-mean asas digunakan sebagai perbandingan dan keputusan perbandingan dibentangkan di dalam

bentuk jadual dan grafik. Bagi kegunaan eksperimen, set data daripada kajian-kajian terdahulu digunakan seperti Supply Base Management (SBM) dan pangkalan data UC Irvine Machine Learning Repository (UCI). Keputusan menunjukkan prestasi bagi teknik yang dicadangkan adalah lebih baik dalam memilih atribut kelompok dan mengkelompokkan data berasaskan kategori dari segi ketulenan (21%), entropi (9%), lelaran (99%), masa (93%), ketepatan (16%), dan ketepatan kekasaran (11%).

# CONTENTS

**DECLARATION** ii

**DEDICATION** iii

**ACKNOWLEDGEMENT** iv

**ABSTRACT** vi

**ABSTRAK** vii

**CONTENTS** ix

**LIST OF TABLES** xii

**LIST OF FIGURES** xiv

**LIST OF SYMBOLS AND ABBREVIATIONS** xv

**LIST OF PUBLICATIONS** xvi

**CHAPTER 1 INTRODUCTION** **1**

1.1 Research background 1
1.2 Research Motivation 5
1.3 Research Objectives 6
1.4 Research Scope 7
1.5 Research Significance 7
1.6 Thesis Organization 8

**CHAPTER 2 LITERATURE REVIEW** **9**

2.1 Introduction 9
2.2 Cluster analysis 9
2.2.1 Probabilistic and Generative Models 10
2.2.2 Distance-Based Algorithms 11
2.2.3 Density and Grid-Based Methods 13
2.2.4 Software Model Clustering 13
2.2.5 Matrix Factorization and Co-Clustering 14
2.2.6 Related work on cluster analysis 14

2.3 Supplier base management (SBM) 21
2.4 Cluster validation 22
2.4.1 Unsupervised measures 23
2.4.2 Supervised measures 23
2.4.3 Relative measures 23
2.4.4 Related work on cluster validation 24
2.4.5 Accuracy 27
2.4.6 Entropy 28
2.4.7 Purity 28
2.4.8 Rough accuracy 29
2.4.9 Minimum iterations, respond time and Big O notation 29
2.5 Rough set theory 30
2.5.1 Information system 31
2.5.2 Indiscernibility relation 33
2.5.3 Set approximations 34
2.5.4 Related work on rough set theory 37
2.6 Rough categorical data clustering and related work 43
2.7 Comparative analysis of existing rough set categorical clustering techniques 47
2.7.1 The ITDR technique 49
2.7.2 Research questions on MDA and MSA techniques 50
2.8 Discussion: Scenario Leading to the Research Framework 63
2.9 Summary 68

**CHAPTER 3 RESEARCH METHODOLOGY 69**

3.1 Introduction 69
3.2 Proposed research framework 69
3.3 Comparison of rough set based proposed and existing techniques 71
3.4 Information-theoretic purity measure and Rough Purity Approach (RPA) 73
3.5 Maximum value attribute (MVA) 79
3.6 Summary 87

**CHAPTER 4 EXPERIMENTAL RESULTS AND DISCUSSIONS 88**

4.1 Introduction 88

4.2 Experimental setup 88
4.3 Experiments with rough purity approach (RPA) 89
4.3.1 Computational complexity in terms of time and iterations for RPA technique 90
4.3.2 Other evaluation measures for RPA technique 91
4.4 Experiments with maximum value attribue (MVA) 95
4.4.1 Effect of NoC on the purity and entropy of clustering 95
4.4.2 Small cases for MVA technique 100
4.4.3 Real and UCI data sets for MVA technique 107
4.4.4 Computational complexity in terms of time and iterations for MVA technique 109
4.4.5 Other evaluation measures for MVA technique 114
4.4.6 Comparison with Simple K Mean 115
4.5 Summary 117

**CHAPTER 5 CONCLUSION 119**

5.1 Accomplished objectives 120
5.1.1 Objective 1 120
5.1.2 Objective 2 121
5.1.3 Objective 3 121
5.2 Contributions of research 122
5.3 Threats to validity 124
5.4 Future Works 124

**REFERENCES 126**
**Vita** 140

## LIST OF TABLES


| | | |
|---|---|---|
| 2.1 | Summary of related work on cluster analysis | 20 |
| 2.2 | An information system | 32 |
| 2.3 | Information System of Flu Patients | 33 |
| 2.4 | Summary of related work on rough set theory | 42 |
| 2.5 | Summary of existing work on rough categorical data clustering | 48 |
| 2.6 | A Dengue Diagnosis Data Set | 54 |
| 2.7 | Dependency Degree of Attributes for Dengue Diagnosis Information System | 55 |
| 2.8 | Significance Degree of Attributes for Dengue Diagnosis Information System | 56 |
| 2.9 | Dependency Degree of Attributes for Lenses Data Set | 57 |
| 2.10 | Significance Degree of Attributes for Lenses Data Set | 57 |
| 2.11 | Suraj's LEMS Data Set | 58 |
| 2.12 | The attribute dependency degrees from Suraj's LEMS Data Set | 59 |
| 2.13 | The degree of significance of all attributes from Suraj's LEMS Data Set | 59 |
| 2.14 | Grzymala's Information System | 60 |
| 2.15 | The attribute dependency degrees from Grzymala's Information System | 60 |
| 2.16 | Pawlak's Car performance Data Set | 60 |
| 2.17 | The attribute significance degree from Pawlak's Car performance information system | 61 |
| 2.18 | Stores Characterization Data Set | 62 |
| 2.19 | Minimum Iterations and Respond Time for Store Data Set | 62 |
| 2.20 | Minimum Iterations and Time for Train Data Set | 63 |
| 2.21 | Strengths and limitations of rough categorical clustering techniques | 65 |
| 3.1 | Comparison of proposed and existing rough set based techniques | 74 |
| 3.2 | Student's enrollment qualification information system | 76 |
| 3.3 | MMP roughness of Table 3.2 | 78 |

| | | |
|---|---|---|
| 3.4 | Flu patients data set | 85 |
| 3.5 | Value Set Cardinality of Table 3.4 | 86 |
| 3.6 | Comparison of RPA and MVA techniques on Balloons data set | 87 |
| 4.1 | The discretized supplier data set | 90 |
| 4.2 | Time complexity of all techniques | 91 |
| 4.3 | Iterative complexity of all techniques | 92 |
| 4.4 | Suraj's Flu Patients Data (Suraj, 2004) | 96 |
| 4.5 | Stores Characterization Data Set (Pawlak, 1991) | 97 |
| 4.6 | Grzymala Data Set (Grzymala-busse, 2005) | 98 |
| 4.7 | Pawlak's Modified Data Set (Pawlak et al., 1995) | 99 |
| 4.8 | Pawlak's Modified Data Set | 100 |
| 4.9 | Dependency Degree of Attributes from Table 4.8 | 101 |
| 4.10 | Significance Degree of Attributes from Table 4.8 | 101 |
| 4.11 | Cardinality of Value Sets from Table 4.8 | 101 |
| 4.12 | Influenza Data Set | 102 |
| 4.13 | Dependency Degree of Attributes from Table 4.12 | 103 |
| 4.14 | Significance Degree of Attributes from Table 4.12 | 104 |
| 4.15 | Cardinality of Attributes Value Sets from 4.12 | 104 |
| 4.16 | Toys attitude data set | 105 |
| 4.17 | Dependency Degree of Attributes from Table 4.16 | 106 |
| 4.18 | Significance Degree of Attributes from Table 4.16 | 106 |
| 4.19 | Cardinality of Attributes Value Sets from Table 4.16 | 107 |
| 4.20 | Number of Clusters Obtained | 109 |
| 4.21 | Computational complexity comparison of techniques | 112 |
| 4.22 | Iterative complexity of all data sets | 113 |
| 4.23 | Time complexity of all data sets | 114 |
| 4.24 | Comparative performance of techniques for all data sets | 116 |

## LIST OF FIGURES


| | | |
|---|---|---|
| 2.1 | A rough set | 35 |
| 2.2 | The ITDR algorithm | 50 |
| 2.3 | The MDA algorithm | 51 |
| 2.4 | The MSA algorithm | 52 |
| 2.5 | Stores characterization data set evaluations | 62 |
| 2.6 | Train Data Set Evaluations | 63 |
| 2.7 | Scenario Leading to the Research Framework | 67 |
| 3.1 | Detail of research process | 70 |
| 3.2 | Proposed research framework | 72 |
| 3.3 | Algorithmic steps comparison | 73 |
| 3.4 | The RPA algorithm | 77 |
| 3.5 | The MVA algorithm | 80 |
| 4.1 | The accuracy of MDA, MSA, ITDR and RPA | 93 |
| 4.2 | The entropy of MDA, MSA, ITDR and RPA | 93 |
| 4.3 | The purity of MDA, MSA, ITDR and RPA | 93 |
| 4.4 | The rough accuracy of MDA, MSA, ITDR and RPA | 94 |
| 4.5 | Evaluation Performance of Surajs Flu Dataset | 96 |
| 4.6 | Evaluation Performance of Stores Characterization Data Set | 97 |
| 4.7 | Evaluation Performance of Grzymala's Data Set | 98 |
| 4.8 | Evaluation Performance of Pawlak's Modified Data Set | 99 |
| 4.9 | Pawlak's Modified Data Set Evaluation Graphs | 102 |
| 4.10 | Pawlak's Influenza Data Set Evaluation Graphs | 105 |
| 4.11 | Infant Toy Attitude Data Set Evaluation Graphs | 108 |
| 4.12 | Clusters visualization of balloons data set | 109 |
| 4.13 | Clusters visualization of soya been data set | 110 |
| 4.14 | Clusters visualization of lenses data set | 110 |
| 4.15 | Clusters visualization of balance scale data set | 111 |
| 4.16 | Comparison of MVA with K-mean technique | 117 |

# LIST OF SYMBOLS AND ABBREVIATIONS

| | | |
|---|---|---|
| RST | – | Rough Set Theory |
| MDA | – | Maximum Dependency Attribute |
| MSA | – | Maximum Significance Attribute |
| ITDR | – | Information Theory Dependency Roughness |
| RPA | – | Rough Purity Approach |
| MVA | – | Maximum Value Attribute |
| NoC | – | Number of Clusters |
| BC | – | Bi-Clustering |
| MMR | – | MinMin-Roughness |
| SDR | – | Standard Deviation Roughness |
| U | – | Universe of objects |
| SBM | – | Supply Base Management |

## LIST OF PUBLICATIONS

1. Jamal Uddin, Rozaida Ghazali, Mustafa Bin Mat Deris (2016), An Empirical Analysis of Rough Set Categorical Clustering Techniques , *PLoS ONE*, Accepted, DOI: 10.1371/journal.pone.0164803 (ISI Q1, IF=3.54)

2. Jamal Uddin, Rozaida Ghazali, Mustafa Bin Mat Deris, Rashid Naseem, Habib Shah (2016), A Survey on Bug Prioritization, *Artificial Intelligence Review, Springer*, PP 1-36, DOI: 10.1007/s10462-016-9478-6 (ISI Q2, IF=2.1)

3. Jamal Uddin, Rozaida Ghazali, Mustafa Bin Mat Deris, Tutut Herawan (2016), Does Number of Clusters Affects the Purity and Entropy of Clustering?, *International Conference on Soft Computing and Data Mining (SCDM)*, Vol 549, ISBN : 978-3-319-51279-2, Springer Conference.

# CHAPTER 1

# INTRODUCTION

## 1.1 Research background

In this present information age, it is believed that information prompts success and strength. The modern technologies like computers and satellites are collecting tremendous amounts of information for us. However, these huge amounts of data in disparate structures overwhelmed in recent years rapidly. Therefore, data base management system (DMBS) and organized data bases are developed (Zaïane, 1999). An efficient DMBS contributes towards effective retrieval of specific information from huge corpus of data. Dealing with huge collections of data, the needs such as automatic summarization of data, discovery of patterns in raw data and extraction of information helps in making better managerial choices. Different kinds of information are collected daily that includes scientific data, software engineering data, games, personal data, satellite sensing, digital media, text reports, business transactions, medical data, world wide web repositories, virtual worlds, surveillance video and pictures.

This enormous amount of data stored in databases, files and other repositories requires a powerful means for interpretation of such data, analysis and for the knowledge extraction that could help in decision-making. Knowledge discovery in databases (KDD) refer to the extraction of previously unknown but potentially useful information which is nontrivial and implicit from the data in databases (Zaïane, 1999). The data mining term being part of the knowledge discovery process is frequently used as synonyms for KDD. The KDD process includes steps like raw data collections leading to formation of new knowledge, data cleaning, data integration, data selection,

data transformation, data mining, pattern evaluation and knowledge representation. The data mining task that is employed determines the kind of information needed to be discovered. In general, there are two types of data mining tasks, that is descriptive and predictive tasks (Zaïane, 1999). Descriptive data mining tasks describe the general properties of the existing data, while the predictive data mining tasks attemps to make predictions based on inference on available data.

Many issues are still pending to be addressed like security, social, interface, mining methodologies, performance and data source before the data mining develops into a conventional and trusted discipline (Zaïane, 1999). The data mining functionalities include prediction, association analysis, classification, clustering, characterization and discrimination etc. The clustering is actually used to analyze accurately the data generated by different modern sources and has appeared as a powerful meta-learning tool. It is considered to be a concise model of the data in the absence of specific labeled information. In particular, the key objective of clustering is to categorize data into clusters so that similar objects are grouped in the same cluster according to specific metrics (Fahad *et al.*, 2014). The internal homogeneity and the external separation is considered by most researchers while describing a cluster (Xu & Wunsch, 2005; Norušis, 2011) i.e., similar objects in the same cluster while different objects in separate clusters.

The different clustering techniques can be broadly classified into partitioning, hierarchical, density, grid and model based approaches (Fahad *et al.*, 2014). Partitioning-based techniques 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 approaches separates 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 (Fahad *et al.*, 2014). 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 (Wong *et al.*, 2000; Dharmarajan & Velmurugan, 2013; Naseem *et al.*, 2013; Britto *et al.*, 2014; Aggarwal & Reddy, 2014).

The particular choice of a clustering technique also relies tremendously on specific data type. The different data types are textual, discrete sequences, time series, uncertain data, categorical and multimedia data (Aggarwal & Reddy, 2014). 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 Haimov *et al.* (1989); Wong *et al.* (2000); Shuanhu et al. (2004). 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, a number of clustering techniques have been developed (Huang, 1998; Guha, S.; Rastogi, 1999; Ganti & Ramakrishnan, 1999; Gibson & Kleinberg, 2000). Though, they contributed well to clustering process but they are not able to handle uncertainty (Herawan *et al.*, 2010a). In many cases where there is no sharp boundary between clusters, the uncertainty becomes an important real world issue.

Huang, Gupta and Kang (Huang, 1998; 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 techniques require multiple runs (Herawan *et al.*, 2010a). Zdzislaw Pawlak introduced rough set theory (RST) (Pawlak, 1991; Pawlak et al., 1995), 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 & 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 (Pawlak & Skowron, 2007).

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 (Düntsch & Gediga, 2015). 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 (Ngo & Nguyen, 2004) and rough-K-Means clustering (Peters, 2006) are the examples of set approximation methods. Despite of having satisfactory results, these methods have issues as they depend on several parameters and thresholds (Düntsch & Gediga, 2015). 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 (Düntsch & Gediga, 2015).

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 techniques to select clustering attribute are developed by Mazlack *et al.* (2000) which includes Total Roughness (TR) and Bi-Clustering (BC) . These techniques work on the accuracy of roughness (approximation accuracy average) in the RST. Later on, another rough categorical clustering approach named Min-Min Roughness (MMR) was proposed by Parmar *et al.* to improve previous techniques (Parmar *et al.*, 2007). Despite of MMR's better performance, issues like accuracy, computational complexity and purity are yet to be addressed.

In 2010, a technique based on the dependency of attributes was introduced by Herawan *et al.* (2010a) 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 & Elmelegy, 2013). Recently, Park and Choi introduced information-theoretic

dependency roughness (ITDR) technique (Park & Choi, 2015b) which finds the entropy roughness to choose the suitable clustering attribute. It is another rough clustering technique that uses the information-theoretic dependencies of categorical attributes in information systems.

## 1.2 Research Motivation

Today the world is full of data and every day 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 & 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 techniques 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 approach for clustering categorical data is using rough set theory (Park & Choi, 2015a). Originally the motivation and inspiration for this study came from exploring useful limitations and issues of existing rough categorical clustering techniques (Mazlack *et al.*, 2000; Parmar *et al.*, 2007; Herawan *et al.*, 2010a; Hassanein & Elmelegy, 2013; Park & Choi,

2015b). This research is conducted in order to come with more general, efficient and better rough categorical clustering techniques. The MDA, MSA and ITDR techniques outperformed their previous techniques such as BC, TR, MMR etc, however, they have certain issues like accuracy, purity, generalizability and computational complexity. On several data sets, these techniques fail or face difficulties in choosing the suitable clustering attribute. Some of the limitations are outlined:

1. MDA technique cannot perform well on data sets with attributes having zero or equal dependency value.
2. MSA technique also fails to select clustering attribute on data sets having attributes with zero or equal significance value.
3. ITDR techniques face issues like random attribute selection and integrity of classes due to presence of entropy measure.

Accordingly in this work, two rough set based categorical clustering techniques are proposed. The first one, information theoretic Rough Purity Approach (RPA) is introduced by establishing a new rough set metric of uncertainty which is rough purity for categorical data clustering. The proposed RPA technique relates the concept of information theoretic purity to rough sets. Considering the domain knowledge of the data set, the second technique Maximum Value Attribute (MVA) is proposed. Here, the rough value set of an attribute is combined with number of clusters. This technique chooses the suitable clustering attribute on basis of maximum number of clusters by an attribute. Several propositions and experiments on benchmark data sets demonstrate the significance, novelty and contribution of these proposed techniques to practical systems.

## 1.3 Research Objectives

The research objectives are listed as follows:

1. To propose a new rough set based categorical clustering technique Rough Purity Approach that takes into account the purity of attributes.
2. To propose another rough set based categorical clustering technique Maximum

Value Attribute that takes into account the value set of attributes combined with number of clusters.

3. To elaborate the performance of proposed techniques on real and benchmark datasets by comparing them with the recent baseline rough categorical clustering techniques like Maximum Dependency Attribute, Maximum Significant Attribute and Information Theoretic Dependency Roughness and classical K-mean clustering algorithm using accuracy, purity, rough accuracy, time and iterative complexity (Big O notation) and entropy.

## 1.4 Research Scope

This research only focuses on proposing two rough set theory based categorical clustering techniques named RPA and MVA. The proposed and existing MDA, MSA, ITDR and classical K-mean techniques are analyzed on several benchmark (UCI and KEEL repositories) and a real Supply Base Management (SBM) data set. The experimental results are evaluated using metrics like accuracy, purity, rough accuracy, number of iterations, respond time and entropy.

## 1.5 Research Significance

The system implementation is significant by two ways in this research. Firstly, information-theoretic purity is introduced as a new definition to measure the uncertainty using RST for categorical data clustering. Secondly, a domain knowledge about data like rough value set is utilized to develop another rough categorical clustering technique combined with internal evaluation measure like number of clusters. Both these approaches show significant improvement for clustering categorical data not only in terms of time and iterations but also in terms of accuracy, purity, entropy and rough accuracy.

## 1.6 Thesis Organization

The remaining thesis is arranged as below:

Chapter 2 discusses some fundamental concepts and overview of existing work on clustering the categorical data using RST. It comprises of an information system notion in rough relational database, an indiscernibility relation, set approximations and quality of approximations. This chapter discusses the literature review of existing work for cluster analysis, cluster validation, SBM, RST and rough categorical data clustering. Moreover, it also presents the analysis and limitations of some existing rough categorical data clustering techniques with the help of examples.

Chapter 3 discusses the proposed techniques of clustering the categorical data, named Rough Purity Approach (RPA) and Maximum Value Attribute (MVA). The notion of purity using rough set theory and the value set cardinality are presented. Moreover, the evaluation metrics used in this research are also defined. Several propositions and examples are illustrated to show the significance of proposed techniques.

Chapter 4 illustrates the results of experiments on proposed techniques. An empirical study on ten small, fifteen benchmark data sets and a real SBM data set demonstrates the better performance of proposed techniques. Moreover, they are compared with most recent and leading rough set-based categorical clustering techniques. All the experimental results are discussed and analyzed in detail by presenting them in form of tables and graphs.

Finally, Chapter 5 gives concluding remarks, accomplished objectives, contributions and future work.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Introduction

The previous chapter demonstrated that cluster analysis and rough set clustering techniques are widely utilized for numerical and categorical data in various forms. Accordingly, this chapter gives an overview of related work on cluster analysis, validation criteria, rough set theory and rough categorical data clustering.

This chapter comprises of nine sections. An overview of cluster analysis techniques and existing work on them are discussed in Section 2.2. The existing work in the field of supply base management is summarized in Section 2.3. The detail of cluster evaluation measures are described in Section 2.4. Similarly, Section 2.5 explains some preliminaries and related research work on rough set theory. Section 2.6 illustrates the overview of existing research on categorical data clustering. Section 2.7 presents the analysis of best recent rough categorical clustering techniques to explore their limitations. Section 2.8 discusses the scenario that leads to research framework. Section 2.9 summarizes the chapter.

## 2.2 Cluster analysis

Clustering is one of the most important unsupervised learning tasks in which the objects are divided into clusters so that similar objects are combined in the same cluster while dissimilar objects in separate clusters. Clustering is widely used in many fields, such as text mining (Naresh Kumar Nagwani, 2012), image analysis (Li

*et al.*, 2014), and bio-informatics (Deris *et al.*, 2015). The issue of data clustering has been widely answered in the machine learning and data mining literature. It has numerous applications to learning, summarization, target marketing and segmentation. Clustering is a concise model of the data in the absence of specific labels that can be referred to either as a generative model or summary. The basic problem of clustering as illustrated by Charu Aggarwal and Chandan Reddy is partitioning of set of data points into possible similar groups (Aggarwal & Reddy, 2014). The variations in this problem definition may be significant depending on specific data type used and type of model utilized like generative and distance-based models.

In many domains such as health care (Abawajy *et al.*, 2015, 2016; Chowdhury *et al.*, 2016), businesses (Cameron *et al.*, 2006), science (e.g., environmental data analysis) (Astel *et al.*, 2007), information security (Abawajy *et al.*, 2014) and software maintenance (Liao *et al.*, 2012), the clustering methods are utilized to support data-driven decision making. The application areas in which the clustering is required are collaborative filtering (Xue *et al.*, 2005), customer segmentation (Mudambi, 2002), data summarization (Jain, 2010), dynamic trend detection (Kontostathis *et al.*, 2004) , multimedia data analysis (Guha *et al.*, 2000), biological data analysis (Chen *et al.*, 2002), intermediate step for other fundamental data mining problems (Berry, 2004; Warren Liao, 2005; Liao *et al.*, 2012) and social network analysis (Ahn *et al.*, 2007).

A wide variety of cluster analysis techniques is employed to address the clustering problems. Moreover, the data preprocessing phase requires dedicated techniques like feature selection or dimensionality reduction methods (Banitaan, 2013; Abawajy *et al.*, 2015). Several good surveys and books have elaborated the clustering issues (Berry, 2004; Warren Liao, 2005; Liao *et al.*, 2012; Aggarwal & Reddy, 2014). The commonly used clustering techniques are illustrated subsequently.

### 2.2.1 Probabilistic and Generative Models

The modeling of data from a generative process is the main idea of probabilistic models. Firstly, assuming a particular form of the generative model like Gaussian

model and estimating model parameter by using algorithm like Expectation Maximization (EM) (Berry, 2004). The available data set is utilized to find the parameters such that they need maximum likelihood set. Later on, for the underlying data points, the generative probabilities (or fit probabilities) of this model are estimated. It is based on the assumption that the data is generated by a mixture of underlying probability dis- tributions. Anomalies have very low fit probabilities while the data points which fit the distribution well will have high fit probabilities.

Generative model tries to know the underlying process of generated cluster that is the reason, it is one of the most fundamental clustering methods (Biernacki *et al.*, 2006; Zivkovic, 2004). Several useful connections between generative models and other clustering methods are present in terms of mixture parameters or prior probabilities (Zhong & Ghosh, 2005). For instance, the exceptional case in which each earlier probability is fixed and all mixture components are expected to have similar radius along all dimensions, leads to soft version of the k-mean algorithm (Jain, 2010).

### 2.2.2 Distance-Based Algorithms

Several special types of generative algorithms are reduced to distance-based algorithms. A distance function within the probability distribution is often used by generative models especially in mixture components for example the mean of mixture of Gaussian distribution generates data probabilities as euclidean distance. Therefore, the Gaussian distribution with generative model can have a close relationship with the k-means algorithm. Thus, many distance-based algorithms can be presented as simplifications or reductions of generative models. Distance-based methods are often attractive due to its ease and simplicity of implementation in wide range of environments. Generally, the distance-based algorithms are of two types; flat and hierarchical.

In flat clustering, the data is separated into number of clusters in one attempt, normally using partitioning representatives. The selection of a distance function and partitioning representative is important as it predicts the performance of underlying

algorithm. The commonly used partitioning techniques are, k-Means (Voges *et al.*, 2002; Peters, 2006; Jain, 2010; Sripada, 2011; Prabha & Visalakshi, 2014), k-Medians (Guha *et al.*, 2000, 2003; Babcock *et al.*, 2003; Har-Peled & Mazumdar, 2004; Cardot *et al.*, 2012) and k-Medoids (Purwitasari *et al.*, 2015; Khatami *et al.*, 2015; Zhou & Mu, 2016). It should be noted that, the k-Means clustering method is one of the most classical, extensively and commonly adopted method due to its simple practical implementations. Despite of drawing from original data set, the K-Means utilizes euclidean distance and forms partitioning representative as a function of the underlying data. In k-Medians methods, instead of mean the median is used to form the partitioning representative along each dimension. The median is usually less sensitive to extreme values of data, hence the k-Medians is more stable to outliers and noise. Whereas, the partitioning representative is sampled from original data in k-Medoids methods. These techniques are helpful in particular where arbitrary objects need to be clustered without considering functions of these objects.

In hierarchical clustering, the clusters are shown as hierarchy using dendogram at different levels of granularity (Maqbool & Babri, 2007; Wang *et al.*, 2010). Depending upon creation in top-down or bottom-up style, hierarchical clustering representations may be either agglomerative or divisive methods. A bottom-up approach is adopted in agglomerative methods where it is initiated with individual data points and sequentially combined with clusters by making a tree-like structure (Feng & Seok, 2011). Different options are there to combine these clusters, which give several trade offs between efficiency and quality. These options include for example centroid-linkage, all-pairs linkage, single-linkage and sampled-linkage clustering. The distance between the centroids is utilized in centroid-linkage whereas the average over all pairs are used in all-pairs linkage (Aggarwal & Reddy, 2014). Single-linkage clustering utlizes the smallest distance between different pairs of points whereas sampled linkage calculates the average distance by sampling data points in the two clusters (Xu & Wunsch, 2005; Miyamoto & Takumi, 2012). The variations in all of these techniques have the drawback of chaining, that is bigger clusters are biased by nature to have nearer distances. Hence, it attracts sequentially larger number of points. Similarly,

a top-down approach is performed in divisive methods to partition the data points successively hence making a tree-like structure (Karaboga & Ozturk, 2011). For performing the partitioning at every step, any flat clustering algorithm can be helpful. Divisive partitioning shows better flexibility in terms of both the level of balance in the different clusters and the hierarchical tree structure.

### 2.2.3 Density and Grid-Based Methods

Density and grid-based methods try to search the data space at high levels of granularity, hence they are two closely related classes (Fahad *et al.*, 2014). At any particular point, the density in data space is defined either in terms of an estimated kernel density or number of data points in a predefined volume of its locality. Grid-based methods are a particular type of density-based methods where the individual regions are explored and converted to grid-like structure of the data space (Warren Liao, 2005). In the post-processing phase, grid like structures are mainly convenient due to better ease in combining the various dense blocks. Such structures may be utilized for high-dimensional methods, as the lower dimensional grids describe clusters on subsets of dimensions.

### 2.2.4 Software Model Clustering

Model-based methods optimize the fit among some predefined mathematical models and given data. They are based on supposition that the data is extracted by a mixing of the underlying probability distributions. Moreover, it helps in automatically finding the number of clusters on the basis of classical statistics. It takes into account the noise (outliers) and hence producing a robust clustering method. The two main approaches based on model-based method include neural network and statistical approaches Fahad *et al.* (2014). MCLUST Xu & Wunsch (2005) and Expectation Maximization Christopher D. Manning & Schütze (2009) are likely the best model-based algorithms whereas, others include neural network approaches (SOM) Xu & Wunsch (2005) and conceptual clustering (COBWEB) Ahmad & Dey (2007).

### 2.2.5 Matrix Factorization and Co-Clustering

Matrix factorization (Li *et al.*, 2014; Chang & Peng, 2012; Bozcan & Bener, 2013) and co-clustering methods (Sun *et al.*, 2010; Wu *et al.*, 2011) are also frequently used methods. They are normally utilized for data that is shown as sparse non negative matrices. Moreover, these methods can be generalized to other types of matrices as well. However, the actual attraction of these methods is the extra interpretability inherent in non negative matrix factorization methods. Thus, in the underlying data, a data point may be expressed as a non negative linear combination of the concepts. Co-clusterings are closely related to non negative matrix factorization methods in a way that they cluster the columns and rows of a matrix at once (Gong & Zhang, 2016; Li *et al.*, 2016).

The literature overview of existing work on cluster analysis is discussed in subsequent paragraphs.

### 2.2.6 Related work on cluster analysis

Clustering algorithms accurately analyze the enormous amount of data generated by modern applications and they are developed as powerful meta-learning tools. Many clustering algorithms have been introduced by researchers for different application domains (Xu & Wunsch, 2005; Norušis, 2011; Dharmarajan & Velmurugan, 2013; Fahad *et al.*, 2014). Such algorithms create high impact in their clustering result quality. The existing work in the area of cluster analysis is summarized as below.

Rousseeuw (1987) proposes partitioning techniques as graphical display. Each cluster is represented as silhouette on the basis of comparison of its separation and tightness. The silhouette presents location of objects in cluster or somewhere in between clusters. The whole clustering is shown by an overview of data configuration, allowing an appreciation of relative quality and by combining the silhouettes into a single plot. The average silhouette width may choose a suitable number of clusters and can evaluate the clustering validity.

Cluster analysis is used to segment radar signals in scanning land and marine objects Haimov *et al.* (1989). Ninety radar signatures are digitally recorded during the probing of ten different land and marine objects with a pulsed coherent Doppler radar. Their spectra are evaluated on the basis of the Marple algorithm for auto regressive model fitting. The method consists in representing the radar signatures as points in four-dimensional (4-D) space and identifying the obtained clusters of 4-D points with the observed objects. The cluster analysis is carried out assuming that the classification parameters have different clustering lengths.

The explanation of a decision support approach to development (D) planning and large-scale Research (R) is presented by Mathieu & Gibson (1993). A quantitative model is used based on analytical tools. Results of the model are used to determine the number of R and D program areas, the technological focus of each R and D program area and the relative allocation of resources to the R and D program areas. The decision support approach developed by them supports, rather than replaces, the judgment of the R and D planner by using a graphic display of the relative position of technology clusters and by using an interactive and iterative approach to problem solving.

The k-means algorithm is extended by Huang (1998) to numeric, categorical and mixed domains values. The k-modes algorithm deals with categorical objects by replacing the means of clusters with modes, utilizing a dissimilarity measure. Moreover, it uses a frequency-based method in the clustering process to update modes. This all minimizes the clustering cost function. These extensions of the k-modes algorithm allow categorical data clustering like k-means. The definition of a combined dissimilarity measure used by k-prototypes algorithm by further integrating the k-modes and k-means. It is also capable to cluster mixed, numeric and categorical attribute objects.

Anquetil & Lethbridge (1999) studied some clustering algorithms and other parameters to establish whether and how they could be used for software re-modularization. They explored the aspects of the clustering activity. Abstract descriptions chosen for the entities to cluster, metrics computing coupling between the entities and clustering algorithms. The experiments were conducted on few public

domain systems. Among other things, they confirmed the importance of a proper description scheme of the entities being clustered. They listed few good coupling metrics to use and characterize the quality of different clustering algorithms. They also proposed novel description schemes not directly based on the source code and advocated better formal evaluation methods for the clustering results.

In the same year, Ganti & Ramakrishnan (1999) generalized the cluster definition for numerical attributes and introduced a novel formalization of a categorical clustering. They described a very fast summarization-based algorithm called CACTUS that discovered exact clusters in the data. CACTUS has two important characteristics. First, the algorithm requires only two scans of the data set, and hence is very fast and scalable. The experiments on a variety of data sets show that CACTUS outperforms previous work by a factor of 3 to 10. Second, CACTUS performs a subspace clustering of the data hence finds clusters in subsets of all attributes.

Moreover, Wong *et al.* (2000) worked on automatic segmentation of tissues in dynamic PET studies using cluster analysis. There proposed tool potentially replaces the manual ROI delineation. Considering the case of segmentation of dynamic lung data, this approach is validated by simulated phantom study to evaluate their performance.

The correct number of clusters in a data set is estimated by Shuanhu et al. (2004). The developed clustering algorithm identifies the natural clusters by handling the complexities of gene data specifically. Moreover, it is tested on real gene changes in yeast cell cycle. The assignment of genes to clusters and basic patterns of gene expression are well explained through previous research. The efficiency of thier proposed algorithm can be witnessed by the comparative analysis with other clustering algorithms.

Similarly, Xu & Wunsch (2005) conducted the survey of clustering algorithms required for applications like bio-informatics, salesman problem, benchmark data sets, machine learning, computer science and statistics. Some related preliminaries, cluster validation, and proximity measures are also illustrated. They conclude by summarizing their review with research trends and exploring several significant issues

for cluster algorithms. According to them, despite of several successful cluster analysis applications, due to the presence of inherent uncertain factors still different open issues remain to be solved. These issues have already got attention and requires more intensive efforts from extensive disciplines.

Some refinements of rough k-means clustering were illustrated by Peters (2006). They analyzed the rough cluster algorithm developed by Lingras et al. (Lingras, 2002) for web mining along with numerical stability, objective function, stability of clusters etc. A comparatively better rough cluster algorithm is proposed by Peters (2006) based on this analysis. The proposed algorithm is applied to gene expression, forest and synthetic data.

Meanwhile, Maqbool & Babri (2007) presented the review of hierarchical clustering in the modularization and software architecture recovery perspective which is related to software model clustering. The in depth analysis of the performance of different distance and similarity measures utilized for software clustering are provided. Similarly, they also analyzed several eminent clustering algorithms specifically studied their clustering process in terms of multiple criteria. Their outcomes show that during a clustering process, the arbitrary decisions affect the algorithm result quality. At last, the recently proposed clustering algorithms are analyzed with argument that different clustering approaches have apparently close similarities. Four legacy software systems in the software domain are selected for experimentation purpose to illustrate the characteristics and working of these prominent clustering algorithms.

To assess cluster analysis application in marketing, an empirical study was conducted by Michael N. Tuma, Sören W. Scholz (2009). They examined the dealing of marketing researchers towards some of the general usage issues. They analyzed that in marketing research since 2000, almost 200 journal articles have been published where cluster analysis was empirically utilized. The outcome of this empirical analysis reveals that new methods are rarely developed and misconceptions still abound. The researchers of marketing field are trying to follow the same procedures as were adopted in past. Moreover, higher standards and better teaching is required in data exploration. They also explores in thbis study that marketing researchers often not describes the

clustering technique they used.

Similarly, Naseem *et al.* (2010) explored the limitations of Jaccard measure for finding appropriate similarity between entities. Accordingly, they came up with a novel similarity measure that handled these limitations. In software systems, the better performance of proposed similarity measure can be seen from the experimental results. Later on, they combined more than one similarity measures to propose Cooperative Clustering Technique (CCT) (Naseem *et al.*, 2013) for hierarchical clustering. They presented an analysis of well-known measures. Secondly, they presents a cooperative clustering approach for two types of well-known agglomerative hierarchical software clustering algorithms, for binary as well as non-binary features. Third, to evaluate the proposed CCT, they conducted modularization experiments on five software systems. Their analysis identifies certain cases that reveal weaknesses of the individual similarity measures. The experimental results supported their hypothesis that these weaknesses may be overcome by using more than one measure, as their CCT produces better modularization results for test systems in which these cases occur. They concluded that CCTs are capable of showing significant improvement over individual clustering algorithms for software modularization.

To discuss the various application areas of partition based clustering algorithms like k-Means, k-Medoids, Fuzzy C-Means, Dharmarajan & Velmurugan (2013) conducted a survey. According to them, the k-Means algorithm is very consistent when compared and analyzed with the other two algorithms. Further, it stamps its superiority in terms of its lesser execution time. From this survey, it is identified that the applications of innovative and special approaches of clustering algorithms principally are for medical domain. From the various applications by several researchers, particularly, the performance of k-Means algorithm is well suited. Most of the researchers are finding that the k- means algorithm is more suitable than other algorithms in their field.

In 2014, Fahad *et al.* (2014) introduced concepts and algorithms related to clustering by conducting a concise survey of existing (clustering) algorithms as well as providing a comparison of both theoretical and empirical perspectives. From a

theoretical perspective, they developed a categorizing framework based on the main properties pointed out in previous studies. They empirically conducted extensive experiments where they compared the most representative algorithm from each of the categories using a large number of real (big) data sets. The effectiveness of the candidate clustering algorithms is measured through a number of internal and external validity metrics, stability, run time, and scalability tests. In addition, they highlighted the set of clustering algorithms that are the best performing for big data.

In the same year, Britto *et al.* (2014) provided an intuitive introduction to cluster analysis. Their targeting audience were both scholars and students in Political Science. Methodologically, they used basic simulation to illustrate the underlying logic of cluster analysis and they replicated data from Coppedge, Alvarez and Maldonado (2008) to classify political regimes according to Dahls (1971) polyarchy dimensions: contestation and inclusiveness. They hoped to help novice scholars to understand and employ cluster analysis in empirical research of political science.

Recently, Aldana-Bobadilla & Kuri-Morales (2015) benchmarked their method relative to the better results theoretically. They utilized best performer techniques like Bayes classifier for normally distributed data and multilayer perceptron network for otherwise. Since in supervised classifications, the elements of classes are known as priori therefore they outperform non-supervised techniques. Moreover, they presented comparatively that the proposed method is effective against supervised one which clearly shows the superiority of proposed approach.

Shelly *et al.* (2016) introduces a new strategy for earthquake focal mechanisms using waveform-correlation-derived relative polarities and cluster analysis. They addressed the limitation of small subset of located events by reliable focal mechanisms in microseismicity analyses. They presented framework for determining robust focal mechanisms for entire populations of very small events. They used cluster analysis to group events with similar patterns of polarities across the network. Their research work aims to address a fundamental gap in typical micro earthquake studies.

The existing work on rough categorical data clustering is summarized in Table 2.1.

Table 2.1: Summary of related work on cluster analysis

| Paper | Proposed Technique | Compared Techniques | Evaluation Metrics | Data Sets/ Application Area |
|---|---|---|---|---|
| Shelly et al. (2016) | New strategy for cluster analysis | Network-determined mechanisms | Polarity, Correlation | Focal mechanism |
| Aldana-Bobadilla & Kuri-Morales (2015) | Clustering Based on Entropy (CBE) | K-means, fuzzy c-means, Bayes classifier, Multilayer perceptron | Effectiveness | Synthetic Gaussian and non-Gaussian datasets, UCI datasets |
| Britto et al. (2014) | Agglomeration methods | K-means | Inclusiveness, contestation | Political Science |
| Fahad et al. (2014) | Taxonomy and empirical analysis | Classical clustering algorithms | Stability, runtime, and scalability tests | MHORD), MHIRD, SHORD, SHIRD, SPFDS, DOSDS, SPDOS, WTP, DARPA, ITD B Big data sets |
| Dharmarajan & Velmurugan (2013) | Survey | Partition based Clustering Algorithms | Number of clusters | Medical data sets |
| Naseem et al. (2010) | Cooperative clustering technique | Agglomerative, LIMBO, Wcombined | MoJoFM measure, arbitrary decisions | Object oriented software systems, Mozilla |
| Tuma et al. (2009) | Empirical study | Several clustering methods | Segmentation Variables, Number of clusters | Marketing research |
| Maqbool & Babri (2007) | Combined and Weighted Algorithms | Agglomerative approaches | Arbitrary decisions, Number of clusters | Open source software systems written |
| Peters (2006) | Refined rough cluster algorithm | Rough cluster algorithm | Objective function,stability | Synthetic, forest and gene data. |
| Xu & Wunsch (2005) | Survey | Several clustering algorithms | Percentage error, Accuracy | Iris, Mushroom, Salesman problem, Bioinformatics. |
| Shuanhu et al. (2004) | Self-Splitting and Competitive Learning | OPTOC | Number of clusters | Gene Expression Data |
| Wong et al. (2000) | Segmentation and phantom study | Manual ROI | Average mean squared error, time | PET Images, lung data |
| Ganti et al. (1999) | CACTUS | STIRR | Similarity, time | Real and synthetic datasets |
| Anquetil & Lethbridge (1999) | Software Re-modularization | Complete, single, weighted | Precision, Recall, Cohesion, Coupling, Similarity | gcc, Linux, Mosaic and real world legacy system |
| Huang (1998) | Extented k-Means and k-modes | k-Means and k-modes | Accuracy, run time, standard deviation | Soybean disease and credit approval |
| Mathieu & Gibson (1993) | Decision support approach | Average linkage, Centroid, Ward's | Growth rate, Gamma frequency | Large scale R & D planning. |
| Haimov et al. (1989) | Fine-classification procedure | Cluster classification | Spectra | Land and marine object |
| Rousseeuw (1987) | Silhouettes | Fuzzy clustering | Average silhouette width, Number of clusters | Ruspini |

## 2.3 Supplier base management (SBM)

Nowadays, the flexibility, business environment, uncertainty, globalization, customer behavior, security and interdependencies among the various factors of supply chain define the market trends. The non-core activities are outsourced whereas, organizations are currently focusing on core competencies. It motivates the importance on supplier based management and dependence of companies on their suppliers. Supply base rationalization, supplier evaluation and development are the three main categories of supplier base management practices. The suppliers evaluate their companies using different supplier selection techniques and models besides supporting the decisions relating to supplier selection. The supplier selection methods include analytical hierarchical process, linear weighting models, mathematical programming models, outranking, expert systems, total cost of ownership, case based reasoning, data envelopment analysis and portfolio analysis (Darshit et al., 2010). The identification and elimination of non capable suppliers in terms of meeting the companys needs comes in the category of supply base rationalization or supply base optimization. This approach results in a group of suppliers that are capable to meet the services and product requirements of the purchasing organization. The needs of buying firm can be fulfilled by developing and managing the performance of suppliers (Krause *et al.*, 1998).

The specifications are replaced with finished deliverables by referring to supply chain through a network of dependent and integrated systems. This term is used commonly in academia and industry. Moreover, the integrated information and materials regarding the product flow from suppliers to end users are managed by the supply chain. It is also useful for improving the inventories cost, time to market and customer satisfaction. Hence, these complexities which is not an easy task are managed by supply chain management. The complexity is dependent on prevailing characterized circumstances because of certain market followings like collaboration, uncertainty, continuously changing business environment, flexibility, globalization, security and customer behavior (Darshit et al., 2010).

According to literature, in whole procedure of transformation to finished deliverables, implication as disturbance and exemption results in delaying significantly customer delivery and are costly. These exemptions are classified into categories like process, output and input by Xu *et al.* (2003). The process associated exemptions are disturbances appearing in the manufacturing system whereas the output associated exemptions are related to order changes from customers. The input associated exemptions includes partial and delayed deliveries etc by suppliers.

According to Zsidisin & Ellram (2003), the lean supply chains becomes more weaker due to supply side disturbance and are fragile because it results from Six Sigma, lean systems and Just-In-Time etc. They also concluded the inbound supply disruptions like delayed launches, late deliveries, stockouts/lost sales and unplanned downtime. This may result in lower market, revenue and high price. The supply chains are now more weak to external disturbance as claimed by Christopher and Peck Darshit et al. (2010). Though, from a cost and quality management perspective, the single sourcing may be advantageous but in terms of resilience it could be dangerous. Moreover, they suggests an alternative supply source.

Despite of valuable energies spent on supplier base management, the enlarged size of the supplier pool is one of challenging problem that need to be addressed. Darshit et al. (2010) applied MMR whereas Herawan *et al.* (2010c) uses MDA technique for supply base management. To date, very limited data mining approaches like clustering are utilized for arranging a large number of suppliers into similar small manageable groups.

## 2.4 Cluster validation

Clusters obtained as a result of clustering process must be assessed to evaluate their quality. Cluster validation or cluster evaluation is important and should be a part of every clustering process. A key motivation is that each clustering approach results in its own kind of clusters in a data set due to several possible cluster combinations. Hence, these clustering algorithms can be judged by evaluating their performance

comparatively. The evaluation may be carried out by using external, relative and internal assessments (Maqbool & Babri, 2007). The evaluation metrics or measures are used to judge different aspects of clustering. They are also named as unsupervised, supervised and relative measures. The definitions of them are taken from the book on data mining by Tan *et al.* (2006).

### 2.4.1 Unsupervised measures

In the absence of external information, if the goodness of a clustering structure is measured then it is unsupervised measurement Tan *et al.* (2006). Sum of Square Error (SSE) is an example of it. The unsupervised cluster validity measures are further divided into cohesion and separation mainly. Custer cohesion (compactness, tightness) finds how much the objects in a cluster are closely related. On the other hand, the cluster separation (isolation) finds how much a cluster is well-separated or distant from other clusters. Due to the reason that, only the information present in the data set is used by unsupervised measures hence, they are also known as internal indices.

### 2.4.2 Supervised measures

The clustering algorithm discovers a clustering structure matching some external structure to an extent that is measured using supervised measures Tan *et al.* (2006). Entropy is a type of supervised index that finds how good the developed clusters are matching the external class labels. Though, the supervised metrics are using the information not available in the data set hence, they are also known as external indices.

### 2.4.3 Relative measures

Different clusterings or clusters are compared in relative cluster evaluation measures Tan *et al.* (2006). It is either an unsupervised or supervised evaluation utilized for the reason of comparison. Hence, they are actually for particular use of such measures but

not a separate cluster evaluation measure like either SSE or entropy measures can be utilized for comparison of two K-means clusterings.

Now, an overview of different cluster measures utilized by researchers are presented in subsequent paras.

### 2.4.4 Related work on cluster validation

Several aspects of cluster validation include: determining the clustering tendency, comparing and evaluating the results to determine the better clustering combination etc. The overview of some existing research that utilizes the cluster validation measures are summarized subsequently.

In year 2000, Davey & Burd (2000) described the investigation of a technique for re-modularizing legacy software for cluster analysis. They took into account the data cohesion as an influencing factor to the re-modularization process and compared and contrasted this with calling structure analysis. A number of different cluster analysis techniques were chosen for evaluation. The authors develop a tool to perform this cluster analysis with two main aims; to provide a way of evaluating the chosen techniques and to provide a usable method of generating a re-modularization of a software system. The techniques evaluated techniques produced modularization of varying quality. However, they thought that cluster analysis is a valuable and useful approach to software re-modularization that is worth further investigation.

The entropy based metrics to find the cluster heterogeneity have been utilized for a long time. Clustering the categorical data using entropy based metric was presented by Li & Ogihara (2004). They illustrated that in the prescribed framework of probabilistic clustering models this entropy based metric can be obtained. Later on, based on dissimilarity coefficients, they developed a link between the approach and criterion. Similarly, to search the partitions that minimizes the criterion, they introduced an iterative Monte-Carlo method. The effectiveness of proposed method is proved through conducting several experiments.

In document datasets, Zhao & KARYPIS (2004) presents their study of

## REFERENCES


Abawajy, J., Kelarev, A., Chowdhury, M. U., & Herbert, F. J. (2016). Enhancing predictive accuracy of cardiac autonomic neuropathy using blood biochemistry features and iterative multitier ensembles. *IEEE Journal of Biomedical and Health Informatics*, *20*(1), 408–415.

Abawajy, J. H., Kelarev, A., & Chowdhury, M. (2014). Large iterative multitier ensemble classifiers for security of big data. *IEEE Transactions on Emerging Topics in Computing*, *2*(3), 352–363.

Abawajy, J. H., Kelarev, A. V., & Chowdhury, M. (2015). Multistage approach for clustering and classification of ECG data. *Computer Methods and Programs in Biomedicine*, *112*(3), 720–730.

Aggarwal, C., & Reddy, C. (2014). *Data Clustering: Algorithms and Applications*. CRC Press Taylor & Francis Group.

Aggarwal, C., & Yu, P. (2009). A Survey of Uncertain Data Algorithms and Applications. *IEEE Transactions on Knowledge and Data Engineering*, *21*(5), 609–623.

Agresti, A. (2007). *An introduction to categorical Data Analysis*, vol. 2.

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

Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., & Jeong, H. (2007). Analysis of Topological Characteristics of Huge Online Social Networking Services. In *Proceedings of the 16th International Conference on World Wide Web*. pp. 835–844.

Aldana-Bobadilla, E., & Kuri-Morales, A. (2015). A clustering method based on the maximum entropy principle. *Entropy*, *17*(1), 151–180.

Amigó, E., Gonzalo, J., Artiles, J., & Verdejo, F. (2009). A comparison of extrinsic

clustering evaluation metrics based on formal constraints. *Information Retrieval*, *12*(4), 461–486.

Anaraki, J., & Eftekhari, M. (2013). Rough set based feature selection: A Review. In *Proceedings of 5th Conference on Information and Knowledge Technology (IKT)*. pp. 301–306.

Anquetil, N., & Lethbridge, T. C. (1999). Experimen ts with Clustering as a Soft w are Remodularization Method. In *Proc. Sixth Working Conf. Reverse Eng.,*. pp. 235–255.

Astel, A., Tsakovski, S., Barbieri, P., & Simeonov, V. (2007). Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water Research*, *41*(19), 4566–4578.

Babcock, B., Datar, M., Motwani, R., & O'Callaghan, L. (2003). Maintaining variance and k-medians over data stream windows. *Proceedings of the twenty second ACM symposium on Principles of database systems*, pp. 234–243.

Banitaan, S. (2013). TRAM : An Approach for Assigning Bug Reports using their Metadata. pp. 215–219.

Bean, C., & Kambhampati, C. (2008). Autonomous clustering using rough set theory. *International Journal of Automation and Computing*, *5*(1), 90–102.

Beaubouef, T., Petry, F. E., & Arora, G. (1998). Information-theoretic measures of uncertainty for rough sets and rough relational databases. *Journal of Information Sciences*, *5*.

Berry, M. W. (2004). *Survey of Text Mining : Clustering, Classification, and Retrieval*.

Biernacki, C., Celeux, G., & Govaert, G. (2006). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725.

Bozcan, O., & Bener, A. B. (2013). Handling missing attributes using matrix factorization. In *Proceedings of 2nd International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE)*. IEEE. pp. 49–55.

Britto, D., Filho, F., Carvalho, E., Alexandre, J., Paranhos, R., Batista, M., Sofia, B., & Duarte, F. (2014). Cluster Analysis for Political Scientists. *Applied Mathematics*, *5*(August), 2408–2415.

Cameron, a. C., Gelbach, J. B., & Miller, D. L. (2006). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics*, *90*(3), 414–427.

Cardot, H., Cénac, P., & Monnez, J.-M. (2012). A fast and recursive algorithm for clustering large datasets with k-medians. *Computational Statistics and Data Analysis*, *Volume 56,*(Issue 6), 1434–1449.

Chang, H. T., & Peng, H. W. (2012). Facial Image Prediction Using Exemplar-based Algorithm and Non-negative Matrix Factorization. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*. pp. 1–4.

Chen, G., Jaradat, S., Banerjee, N., Tanaka, T., Ko, M., & Zhang, M. (2002). Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data. *STATISTICA SINICA*, *12*(1), 241–262.

Chen, L.-F., & Tsai, C.-T. (2016). Data mining framework based on rough set theory to improve location selection decisions: A case study of a restaurant chain. *Tourism Management*, *53*, 197–206.

Chowdhury, M., Abawajy, J., Kelarev, A., & Jelinek, H. (2016). A Clustering-Based Multi-Layer Distributed Ensemble for Neurological Diagnostics in Cloud Services. *IEEE Transactions on Cloud Computing*, *4*(2), 1–1.

Christopher D. Manning, P. R., & Schütze, H. (2009). *Introduction to Information Retrieval*.

Darshit et al., P. (2010). A clustering algorithm for supplier base management. *International Journal of Production Research*, *48*(13), 3803–3821.

Davey, J., & Burd, E. (2000). Evaluating the suitability of data clustering for software remodularisation. In *Proceedings of Seventh Working Conference on Reverse Engineering*. IEEE Comput. Soc. pp. 268–276.

Deris, M. M., Abdullah, Z., Mamat, R., & Yuan, Y. (2015). A new limited tolerance

relation for attribute selection in incomplete information systems. In *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. IEEE. pp. 964–970.

Dharmarajan, A., & Velmurugan, T. (2013). Applications of partition based clustering algorithms: A survey. *Proceedings of IEEE International Conference on Computational Intelligence and Computing Research*.

Düntsch, I., & Gediga, G. (2015). Rough set clustering. Tech. rep., Brock University Department of Computer Science Rough, Ontario, Canada.

Fahad, A., Alshatari, N., Tari, Z., Alamri, A., Khalil, I., Zomoya, A., Foufou, S., & Bauras, A. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis. *IEEE Transactions on Emerging Topics in Computing*, *2*(3), 1–13.

Feng, H., Chen, Y., Ni, Q., & Huang, J. (2014). A New Rough Set Based Classification Rule Generation Algorithm (RGI). In *Proceedings of International Conference on Computational Science and Computational Intelligence*. Ieee. pp. 380–385.

Feng, J., & Seok, H. (2011). Applying agglomerative hierarchical clustering algorithms to component identification for legacy systems. *Information and Software Technology*, *53*(6), 601–614.

Ganti, V., & Ramakrishnan, J. G. R. (1999). CACTUS Clustering Categorical Data Using Summaries. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 73–83.

Gao, Y., Zhang, X., Wu, L., Yin, S., & Lu, J. (2017). Resource basis, ecosystem and growth of grain family farm in China: Based on rough set theory and hierarchical linear model. *Agricultural Systems*, *154*(May), 157–167.

Garcia, H. V., & Shihab, E. (2014). Characterizing and Predicting Blocking Bugs in Open Source Projects Categories and Subject Descriptors. In *Proceedings of the 11th Working Conference on Mining Software Repositories*. pp. 72–81.

Gibson, D., & Kleinberg, J. (2000). Clustering categorical data : an approach based on dynamical systems. *The VLDB Journal*, *8*, 222–236.

Gong, X., & Zhang, G. (2016). Non-Negative Matrix Co-Factorization for Weakly Supervised Image Parsing. *IEEE Signal Processing Letters*, *9908*(c), 1–1.

Grzymala-busse, J. W. (2005). Rough Set Theory with Applications to Data Mining. *Real World Applications of Computational Intelligence*, *Volume 179*, pp 221–244.

Guha, S., Meyerson, A., Mishra, N., Motwani, R., & OCallaghan, L. (2003). Clustering data streams: Theory and practice. *Knowledge and Data Engineering, IEEE Transactions on*, *15*(3), 515–528.

Guha, S., Mishra, N., Motwani, R., & O'Callaghan, L. (2000). Clustering data streams. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*. IEEE Comput. Soc. pp. 359–366.

Guha, S.; Rastogi, R. K. S. (1999). ROCK : A Robust Clustering Algorithm for Categorical. In *Proceedings., 15th International Conference on Data Engineering,*. pp. 512 – 521.

Haimov, S., Michalev, M. A., & Savchenko, A. (1989). Classification of radar signatures by autoregressive model fitting and cluster analysis. *IEEE Transactions on Geoscience and Remote Sensing*, *21*(5), 606–610.

Har-Peled, S., & Mazumdar, S. (2004). Coresets for k-Means and k-Median Clustering and their Applications. *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pp. 291—-300.

Hassanein, W., & Elmelegy, A. (2013). An Algorithm for Selecting Clustering Attribute using Significance of Attributes. *International Journal of Database Theory & Application*, *6*(5), 53–66.

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

Herawan, T., Ghazali, R., Tri, I., Yanto, R., & Deris, M. M. (2010b). Rough Set Approach for Categorical Data Clustering 1. *International Journal of database theory and Application*, *3*(1), 179–186.

Herawan, T., Tri, I., Yanto, R., & Deris, M. M. A. T. (2010c). ROSMAN : ROugh Set approach for clustering Supplier base MANagement. *Biomedical Soft*

*Computing and Human Sciences*, *16*(2), 105–114.

Hodge, V. J., & Austin, J. I. M. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, *22*, 85–126.

Huang, A. (2008). Similarity measures for text document clustering. *Proceedings of the Sixth New Zealand Computer Science Research Student Conference*, (April), 49–56.

Huang, Z. (1998). Extensions to the k -Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, *2*, 283–304.

Hunter, M. G., & Peters, G. (2012). Rough Sets: Selected Methods and Applications in Management and Engineering. *Advanced Information and Knowledge Processing*, pp. 129–138.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666.

Jia, X., Shang, L., Zhou, B., & Yao, Y. (2016). Generalized attribute reduct in rough set theory. *Knowledge-Based Systems*, *91*, 204–218.

Jyoti (2013). Clustering categorical data using rough set: A Review. *International Journal of Advanced Research in IT and Engineering*, *2*(12), 30–37.

Karaboga, D., & Ozturk, C. (2011). A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Applied Soft Computing*, *11*(1), 652–657.

Kent (2008). Cluster Validation. Tech. rep.

Khatami, A., Mirghasemi, S., Khosravi, A., & Nahavandi, S. (2015). A New Color Space Based on K-Medoids Clustering for Fire Detection. *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015*, pp. 2755–2760.

Kim, D.-w., Lee, K. H., & Lee, D. (2004). Fuzzy clustering of categorical data using fuzzy centroids. *Pattern Recognition Letters*, *25*(11), 1263–1271.

Komorowski, J., Polkowski, L., & Skowron, A. (1999). Rough sets: A tutorial. *Rough fuzzy*, pp. 2–8.

Kontostathis, A., Galitsky, L. M., Pottenger, W. M., Roy, S., & Phelps, D. J. (2004). *A survey of emerging trend detection in textual data mining*.

Krause, D. R., Handfield, R. B., & Scannell, T. V. (1998). An empirical investigation of supplier development: reactive and strategic processes. *Journal of Operations Management*, *17*(1), 39–58.

Kumar, P., & Tripathy, B. (2009). MMeR an algorithm for clustering heterogeneous data using rough set theory. *International Journal Rapid Manufacturing*, *1*(2).

Leibniz., G. W. (1989). *Discourse on Metaphysics*.

Lenič, M., Povalej, P., & Kokol, P. (2005). Impact of Purity Measures on Knowledge Extraction in Decision Trees. In *Foundations and Novel Approaches in Data Mining*, May 2016, pp. 229–242. Berlin/Heidelberg: Springer-Verlag.

Leung, Y., Fischer, M. M., Wu, W. Z., & Mi, J. S. (2008). A rough set approach for the discovery of classification rules in interval-valued information systems. *International Journal of Approximate Reasoning*, *47*(2), 233–246.

Li, J., Bioucas-Dias, J. M., Plaza, A., & Liu, L. (2016). Robust Collaborative Nonnegative Matrix Factorization for Hyperspectral Unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, *9*(9), 4267–4279.

Li, L., Yang, J., Zhao, K., Xu, Y., Zhang, H., & Fan, Z. (2014). Graph Regularized Non-negative Matrix Factorization By Maximizing Correntropy. *JOURNAL OF COMPUTERS*, *9*(11), 2570–2579.

Li, T., & Ogihara, M. (2004). Entropy-Based Criterion in Categorical Clustering. In *Proceedings of the 21st International Conference on Machine Learning, Banff, Canada*.

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications - A decade review from 2000 to 2011. *Expert Systems with Applications*, *39*(12), 11303–11311.

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

Lingras, P. (2002). Rough set clustering for Web mining. *2002 IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE'02. Proceedings (Cat. No.02CH37291)*, *2*, 1039–1044.

MacQueen, J. B. (1967). K means some methods for classification and analysis of multivariate observations. *5th Berkeley Symposium on Mathematical Statistics and Probability 1967*, *1*(233), 281–297.

Maqbool, O., & Babri, H. A. (2007). Hierarchical clustering for software architecture recovery. *IEEE TRANSACTIONS ON SOFTWARE ENGINEERING,*, *33*(11), 759–780.

Mathieu, R. G., & Gibson, J. E. (1993). A methodology for large-scale R&D planning based on cluster analysis. *IEEE Transactions on Engineering Management*, *40*(3), 283–292.

Mazlack, L. J., He, A., & Zhu, Y. (2000). A Rough Set Approach in Choosing Partitioning Attributes. In *Proceedings of the ISCA 13th, International Conference, CAINE*. pp. 1–6.

Michael N. Tuma, Sören W. Scholz, R. D. (2009). The Application Of Cluster Analysis In Marketing Research. *Business Quest*.

Miyamoto, S., & Takumi, S. (2012). Hierarchical clustering using transitive closure and semi-supervised classification based on fuzzy rough approximation. In *2012 IEEE International Conference on Granular Computing*. IEEE. pp. 359–364.

Mohebi, E., & Sap, M. (2009). Rough Set Based Clustering of the Self Organizing Map. *2009 First Asian Conference on Intelligent Information and Database Systems*, (1), 82–85.

Mudambi, S. (2002). Branding importance in business-to-business markets. Three buyer clusters. *Industrial Marketing Management*, *31*(6), 525–533.

Naresh Kumar Nagwani, S. V. (2012). CLUBAS: An Algorithm and Java Based Tool for Software Bug Classification Using Bug Attributes Similarities. *Journal of Software Engineering and Applications*, *05*(06), 436–447.

Naseem, R., Maqbool, O., & Muhammad, S. (2010). An Improved Similarity Measure for Binary Features in Software Clustering. In *Second International Conference on Computational Intelligence, Modelling and Simulation*.

Naseem, R., Maqbool, O., & Muhammad, S. (2013). Cooperative clustering for

software modularization. *The Journal of Systems & Software*, *86*(8), 2045–2062.

Ngo, C. L., & Nguyen, H. S. (2004). A Tolerance Rough Set Approach. *Knowledge Discovery in Databases*, pp. 515–517.

Norušis, M. (2011). Cluster Analysis. In *Statistical Procedures Companion*, pp. 361–391.

P. Danziger (2015). Big O Notation. Tech. rep.

Park, I. K., & Choi, G. S. (2015a). A variable-precision information-entropy rough set approach for job searching. *Information Systems*, *48*, 279–288.

Park, I.-k., & Choi, G.-s. (2015b). Rough set approach for clustering categorical data using information-theoretic dependency measure. *Information Systems*, *48*, 289–295.

Parmar, D., Wu, T., & Blackhurst, J. (2007). MMR: An algorithm for clustering categorical data using Rough Set Theory. *Data & Knowledge Engineering*, *63*(3), 879–893.

Pawlak, Z. (1991). *Rough Sets Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers.

Pawlak, Z. (1995). Vagueness and uncertainty: A Rough Set Perspective. *Computational Intelligence*, *11*(2), 227–232.

Pawlak, Z. (1996). Rough sets and data analysis. In *Proceedings of Asian Fuzzy Systems Symposium on Soft Computing in Intelligent Systems and Information Processing.*. IEEE. pp. 1–6.

Pawlak, Z., & Skowron, A. (2007). Rudiments of rough sets. *Information Sciences*, *177*(1), 3–27.

Pawlak et al., Z. (1995). Rough sets. *Communications of the ACM*, *38*(11), 88–95.

Peters, G. (2006). Some refinements of rough k-means clustering. *Pattern Recognition*, *39*(8), 1481–1491.

Prabha, K., & Visalakshi, N. (2014). Improved Particle Swarm Optimization Based K-Means Clustering. *Intelligent Computing Applications*.

Purwitasari, D., Fatichah, C., Arieshanti, I., & Hayatin, N. (2015). K-medoids

algorithm on Indonesian Twitter feeds for clustering trending issue as important terms in news summarization. *Proceedings of 2015 International Conference on Information and Communication Technology and Systems, ICTS 2015*, pp. 95–98.

Qamar, U. (2013). A Rough-Set Feature Selection Model for Classification and Knowledge Discovery. *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pp. 788–793.

Rahman, M. N. a., Lazim, Y. M., & Mohamed, F. (2011). Applying Rough Set Theory in Multimedia Data Classification. *International Journal on New Computer Architectures and Their Applications (IJNCAA)*, *1*(3), 683–693.

Ramani, G. (2013). Rough set with Effective Clustering Method. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(2), 1163–1167.

Reddy, H. V., Viswanadha Raju, S., & Agrawal, P. (2013). Data labeling method based on cluster purity using relative rough entropy for categorical data clustering. In *Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. pp. 500–506.

Rissino, S., & Lambert-torres, G. (2009). Rough Set Theory Fundamental Concepts , Principals , Data Extraction , and Applications. In Julio Ponce and Adem Karahoca (Ed.) *Data Mining and Knowledge Discovery in Real Life Applications*, pp. 35–58. I-Tech, Vienna, Austria.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65.

Sachdeva, S., & Kastore, B. (2014). Document Clustering : Similarity Measures. Tech. Rep. 11693, Indian Institute of Technology Kanpur.

Senan, N., Ibrahim, R., Nawi, N. M., Tri, I., Yanto, R., & Herawan, T. (2011). Rough Set Approach for Attributes Selection of Traditional Malay Musical Instruments Sounds Classification 1. *International Journal of database theory and Application*, *4*(3), 59–76.

Shelly, D. R., Hardebeck, J. L., Ellsworth, W. L., & Hill, D. P. (2016). A new strategy for earthquake focal mechanisms using waveform-correlation-derived relative polarities and cluster analysis: Application to the 2014 Long Valley Caldera earthquake swarm. *Journal of Geophysical Research: Solid Earth*, *121*(12), 8622–8641.

Shuanhu et al., W. (2004). Cluster Analysis of Gene Expression Data Based on Self-Splitting and Merging Competitive Learning. *IEEE Transactions on Information Technology in Biomedicine*, *8*(1), 5–15.

Singh, S., Mayfield, C., Prabhakar, S., Shah, R., & Hambrusch, S. (2007). Indexing uncertain categorical data. In *International Conference on Data Engineering*. pp. 616–625.

Sripada, S. C. (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.

Sun, B., Yao, H., Ji, R., Xu, P., Sun, X., & Yuan, K. (2010). Individual Home-Video Collecting Using a Co-clustering Method. *First International Conference on Pervasive Computing, Signal Processing and Applications*, pp. 1132–1135.

Suraj, Z. (2004). An Introduction to Rough Set Theory and Its Applications. In *ICENCO2004*.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Addison-Wesley.

Tripathy, B., & Ghosh, A. (2011a). SDR: An algorithm for clustering categorical data using rough set theory. *IEEE Recent Advances in Intelligent Computational Systems*, pp. 867–872.

Tripathy, B., Goyal, A., Chowdhury, R., & Sourav, P. A. (2017). MMeMeR: An algorithm for clustering heterogeneous data using rough set theory. *International Journal of Intelligent Systems and Applications*, *8*, 25–33.

Tripathy, B. K., & Ghosh, A. (2011b). SSDR : An Algorithm for Clustering Categorical Data Using Rough Set Theory. *Advances in Applied Science Research*, *2*(3), 314–326.

Tripathy, B. K., Goyal, A., & Sourav, P. A. (2016). A comparative analysis of rough intuitionistic fuzzy k-mode algorithm for clustering categorical data. *Research Journal of Pharmaceutical, Biological and Chemical Sciences*, *7*(5), 2787–2802.

Voges, K. E., & Pope, N. K. L. (2012). Rough Clustering Using an Evolutionary Algorithm. *Proceeding of 45th Hawaii International Conference on System Sciences*, pp. 1138–1145.

Voges, K. E., Pope, N. K. L., & Brown, M. R. (2002). Cluster Analysis of Marketing Data: A Comparison of K-Means, Rough Set, and Rough Genetic Approaches. In *Heuristics and Optimization for Knowledge Discovery*, pp. 208–216.

Wang, W., Gao, W., Wang, C., & Li, J. (2013). An Improved Algorithm for CART Based on the Rough Set Theory. In *Proceedings of Fourth Global Congress on Intelligent Systems*, January 2002. Ieee. pp. 11–15.

Wang, Y., Liu, P., Guo, H., Li, H., & Chen, X. (2010). Improved Hierarchical Clustering Algorithm for Software Architecture Recovery. In *International Conference on Intelligent Computing and Cognitive Informatics*. pp. 1–4.

Warren Liao, T. (2005). Clustering of time series data - A survey. *Pattern Recognition*, *38*(11), 1857–1874.

Wong, K.-P., Feng, D., Meikle, S. R., & Fulham, M. J. (2000). Segmentation of dynamic PET images using cluster analysis. *IEEE Symposium on Nuclear Science*, *3*, 126–130.

Wu, J., Hassan, A. E., & Holt, R. C. (2005). Comparison of clustering algorithms in the context of software evolution. In *IEEE International Conference on Software Maintenance, ICSM*, vol. 2005. pp. 525–535.

Wu, J., Xiong, H., & Chen, J. (2009). Adapting the right measures for K-means clustering. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, p. 877.

Wu, J. S., Lai, J. H., & Wang, C. D. (2011). A novel co-clustering method with intra-similarities. *Proceedings of IEEE International Conference on Data Mining, ICDM*, pp. 300–306.

Xu, H. Q., Besant, C. B., & Ristic, M. (2003). System for enhancing supply chain agility through exception handling. *International Journal of Production Research*, *41*(6), 1099–1114.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678.

Xue, G.-R., Lin, C., Yang, Q., Xi, W., Zeng, H.-J., Yu, Y., & Chen, Z. (2005). Scalable collaborative filtering using cluster-based smoothing. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05*, p. 114.

Yanto, I., Herawan, T., & Deris, M. (2011). Data clustering using variable precision rough set. *Intelligent Data Analysis*, *15*, 465–482.

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

Zaïane, O. R. (1999). (Chapter 1) Introduction to Data Mining. *Principles of Knowledge Discovery in Databases*, pp. 1–15.

Zhang, L., Li, Y., Sun, C., & Nadee, W. (2013). Rough Set Based Approach to Text Classification. In *Proceedings of International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. Ieee. pp. 245–252.

Zhao, Y. (2001). Criterion functions for document clustering: Experiments and analysis. Tech. rep., Department of Computer Science, University of Minnesota.

Zhao, Y., & KARYPIS, G. (2004). Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering . *Machine Learning*, *55*, 311–331.

Zhong, S., & Ghosh, J. (2005). Generative model-based document clustering: a comparative study. *Knowledge and Information Systems*, *8*(3), 374–384.

Zhou, Z., & Mu, L. (2016). Representative Virtual Machine Templates: An optimized virtual machine templates management mechanism for an Cloud system based on K-medoids Clustering. In *Proceedings of 35th Chinese Control Conference*

*(CCC)*. IEEE. pp. 5243–5248.

Zivkovic, Z. (2004). Improved adaptive Gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 2. IEEE. pp. 28–31.

Zsidisin, G. a., & Ellram, L. M. (2003). An agency theory investigation of supply risk management. *The Journal of Supply Chain Management*, *39*(August), 15–27.