THE UNIVERSITY OF TURKISH AERONAUTICAL ASSOCIATION

INSTITUE OF SCIENCE AND TECHNOLOGY

APPLICATION OF MEAN GAIN RATIO (MGR) MODEL FOR THE CLUSTERING OF ELECTRICAL GENERATOR FAILURES

MASTER THESIS

Saddam Raheem Salih AL- Saadi

THE DEPARTMENT OF INFORMATION TECHNOLOGY

THE PROGRAM OF INFORMATION TECHNOLOGY

June 2017

THE UNIVERSITY OF TURKISH AERONAUTICAL ASSOCIATION

INSTITUE OF SCIENCE AND TECHNOLOGY

APPLICATION OF MEAN GAIN RATIO (MGR) MODEL FOR THE CLUSTERING OF ELECTRICAL GENERATOR FAILURES

MASTER THESIS

Saddam Raheem Salih AL- Saadi

1406050025

THE DEPARTMENT OF INFORMATION TECHNOLOGY

THE PROGRAM OF INFORMATION TECHNOLOGY

Supervisor : Assist. Prof. Dr. Tansel Dökeroglu

Co-Supervisor: Assist. Prof. Dr. Shadi AL SHEHABI

June 2017

Saddam Raheem, having student number 1406050025 and enrolled in the Master Program at the Institute of Science and Technology at the University of Turkish Aeronautical Association, after meeting all of the required conditions contained in the related regulations, has successfully accomplished, in front of the jury, the presentation of the thesis prepared with the title of: "Application of Mean Gain Ratio (MGR) model for the clustering of Electrical Generator Failures"

Supervisor : **Assist. Prof. Dr. Tansel Dökeroglu** ---------------

Türk Hava Kurumu Üniversitesi

Co-Supervisor: **Assist. Prof. Dr. Shadi AL SHEHABI** ---------------

Türk Hava Kurumu Üniversitesi

Jury Members: **Assist. Prof. Dr. Tansel Dökeroğlu** ---------------

Türk Hava Kurumu Üniversitesi

Prof. Dr. Ahmet Cosar ---------------

Orta Doğu Teknik Üniversitesi

Assist. Prof. Dr. Yuriy ALYEKSYEYENKOV ---------------

Türk Hava Kurumu Üniversitesi

Thesis Defense Date: 21.6.2017

THE UNIVERSITY OF TURKISH AERONAUTICAL ASSOCIATION

INSTITUTE OF SCIENCE AND TECHNOLOGY

I hereby declare that all the information in this study I presented as my Master's Thesis, called: Application of Mean Gain Ratio (MGR) model for the clustering of Electrical Generator Failures, has been presented in accordance with the academic rules and ethical conduct. I also declare and certify with my honor that I have fully cited and referenced all the sources I made use of in this present study.

21.06.2017

Saddam Raheem Salih AL-Saadi

ACKNOWLEDGEMENTS

Thanks to the most compassionate, gracious and merciful. May Allah's blessings and peace be upon our prophet Mohammed who protects us from the depths of darkness and leads us forth into light, and his household.

It is a pleasure to express my special thanks to my family and friends for their valuable support.

I would like to express my sincere gratitude to Asst. Prof. Dr Tansel Dökeroglu and Assist. Prof. Dr Shadi AL SHEHABI for supervision, special guidance, suggestions, and encouragement through the development of this thesis.

Also, I would like to present my thanks to the head of the department of IT and all the lecturers in IT who provided me with the scientific knowledge.

June 2017 **Saddam Raheem salih Al-saadi**

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Application of Mean Gain Ratio (MGR) model for the clustering of Electrical Generator Failures

AL-Saadi, Saddam

Master, Department of Information Technology

Supervisor : Assist. Prof. Dr. Tansel Dökeroglu

Co-Supervisor: Assist. Prof. Dr. Shadi AL SHEHABI

June- 2017, 105 pages

Categorical data clustering is getting more and more important part of data mining. In this study, we compared four data clustering methods which are VPRS, MTMDP, ITDR and MGR to cluster four real life databases. The VPRS, MTMDP and ITDR algorithms are based on the Rough Set Theory while the MGR algorithm is based on the Information Theory. Three of the databases used from UCI databases while the other database is collected for electrical generators failure from a mobile company in Iraq. Three performance measures are used to evaluate the performance of each method by calculating the purity and F-measure for the resulting clusters with respect to the database classes and the time consumed by each algorithm to process the databases. The comparison results show that the MGR has the superiority over the other algorithms. Thus, the MGR results are chosen to be proposed to the decision makers and it may potentially contribute to give a recommendation how to design intervention in order to improve the efficiency of the maintenance team performance and moreover to reduce electrical generators failure. In addition, we propose a new technique called Minimum Information Gain Roughness (MIGR) to select the clustering attribute based on information entropy in rough set theory. To evaluate the performance of this technique, three real life sample data sets (UCI) are chosen to be clustered using MIGR, the resulting clusters are compared to the clusters resulted from the Min-Min-Rough (MMR) and Information-Theoretic Dependency Roughness (ITDR) techniques which are compared with many other clustering techniques, such as k-modes, fuzzy centroids and fuzzy k-modes. Accuracy and F-measure are the measures chosen to compare the quality of the resulting clusters. The experimental results show that the MIGR algorithm outperforms the MMR and ITDR algorithms; therefore, it can be used for clustering categorical data.

Keywords: Categorical data (maintenance), Rough set theory, Clustering, Information system, Information theory, Decision markers.

CHAPTER ONE

INTRODUCTION

It is getting more and more important to cluster data sets into groups of objects in a way that the objects in each group are more similar to each other than the objects in the other groups. There are many theories proposed for data clustering. Information theory was introduced by Shannon [1]. Rough set theory was suggested by Pawlak [2]. It is a useful method for data analysis of vague information and it has been successfully employed in research areas involve knowledge discovery, decision analysis, data warehouse, pattern recognition, machine learning and data mining [3-6].The rough set theory has major potentiality in the fields of maintenance and industrial plants. Rough set theory is one of the mathematical techniques for extracting knowledge from huge data [7].

The approach of the rough set theory is based on the indiscernibility relation and clustering analysis. Clustering analysis leads to dividing a given database into sub-database with similar objects, and the technique is widely used in many applications [8]. The cluster analysis techniques often face with difficulty because of the fact that many of the data contained in modern databases are categorical in life. This requires the utilization of rough set algorithms that is one of the data mining tools for clustering categorical data [9]. Data mining functions are split into two categories: predictive data mining and descriptive data mining. The first category predicts the future trends of the variables. There are many data mining techniques to achieve that such as deviation analysis and prediction .The second category describes the properties of the objects. There are many data mining techniques to achieve that such as classification and clustering [10]. The main purpose of the algorithms is to handle uncertainty in the clustering process for categorical data clustering. The major purpose of the rough theory is clustering a database and to map it to the decision table [2] [11,12]. Moreover the divide-and-conquer method is used to find the clusters of objects. There is a need for strong clustering algorithms that can handle uncertainty in the process of clustering categorical data, thus, we propose clustering algorithms, based on rough set theory using a variable precision rough set (VPRS) clustering algorithm based on the maximum mean accuracy [13], Maximum Total Mean Distribution Precision (MTMDP) clustering algorithm based on distribution of approximation precision [14], Information-theoretic Dependency Roughness (ITDR) clustering algorithm based on the mean degree of rough entropy[15], and Mean Gain Ratio (MGR) clustering algorithm based on the maximum mean of gain ratio of each attribute [16]. In this study, we employ our proposed clustering techniques through three real-life datasets: Dermatology [17] , Breast Cancer which obtained from the UCI Machine Learning Repository [18], Soybean[46] and a real life Electrical Generator Failures dataset which is taken from a mobile phone company.

Figure 1.1: Electrical Generators failures sources identification.

We present a real dataset of Electrical Generators failures. This data were taken from a mobile phone company the study aims to analyze the influence of maintenance variables on electrical generators among mobile phone sites, which consists of 636sites (objects) and 38 causes of failures (attributes) grouped into three sources of failure that are mechanical, electrical and Sites management that are beyond the control of the maintenance team. How often each failure affects the availability of the generator was described by choosing one of five options (Never, Rare, Often, Frequent and Severe), these values are stored in a database as (1,2,3,4 and 5) Consecutively. This data was collected for the year 2015.

Experimental results on these three datasets show that MGR algorithm performs better than VPRS, MTMDP and ITDR in terms of performance and the process of selecting most effective attributes. So by using Mean Gain Ratio Model, we present how electrical generator failures can be grouped. Thus this study may potentially contribute to give a recommendation how to design intervention in order to improve the efficiency of the maintenance team performance and moreover to reduce electrical generators failure.

The framework of this work is organized as follows. Chapter two describes the related work. Chapter three presents a brief description on rough set-based algorithms for selecting a clustering attribute, following by the proposed MGR algorithm. Chapter four describes the experimental tests .Chapter five propose a new algorithm. Finally, the conclusion of this work is described in Chapter six.

1.1 Problem definition

The department of maintenance faces many challenges; some of them are technical problems like mechanical or electrical failures, and non-technical challenges like financial or management problems. These challenges increase with the intensive use of electrical generators. Furthermore, these challenges can be categorized as mechanical, electrical and Site management. A specific component failure may lead to a functional failure of the system/subsystem. The operational requirements should be considered carefully when processing maintenance tasks. Not all failures require an overall maintenance because of the probability of them occurring in remote sites or their effect is not important [19]. Site management, finance and telecom problems are some of these problems.

A simple database is used to store the electrical generators data. This technique has caused a lack in logging many maintenance activities that led to several problems such as:

1- Analyzing the data and procedural reports such as failure effect on each electrical generator repertoire and company performance reports. The reports are very important for decision-making process during the maintenance process.

2- Manage the documents and control of the inflow such as storage, retrieval, processing, routing, and distribution of in a secure and useful method, to ensure provide documents when required.

3- The number of sites increases in years, so it becomes difficult to manage the maintenance requests that are especially preventive and corrective maintenance. This makes the maintenance data huge, and it is not easy to analyze the influence of maintenance variables on electrical generators and extract the knowledge by the managers in charge. Therefore, it is necessary to develop a computer-aided approach to assist the decision-makers for extracting useful information (knowledge) from this data. One of the most important functions in data mining tools are the analyses of maintenance data that is based on clustering attributes to find the knowledge by using rough set algorithms. The data in rough sets theory is the orders in a table called decision table. Rows of the decision table correspond to objects and columns correspond to attributes. In the data set, a class label to indicate the class to which each row belongs. The class label is called as decision attribute, the rest of the attributes are the condition attributes and decision attributes [20].

Rough sets theory defines three regions based on the equivalent classes induced by the attribute values: lower approximation, upper approximation and boundary. Lower contains all the objects which are classified surely based on the data collected and upper approximation contains all the objects, which can be classified probably, while the boundary is the lower approximation [2]. The information system is used to selected clustering attribute based on the rough set and Information theory algorithms .It is represent sources of electrical generators failures. The table (1.1) : example dataset with five objects and three attributes .

Table 1.1: Dataset with five objects with three attributes

Object (sites)	Attribute1 (Electrical unit)	Attribute 2 (Mechanical unit)	Attribute 3 (Site management unit)
1	ATS	Radiator	NO Fuel
2	Over Voltage	Replacing lift pump	Commercial power
3	Cable short in generator	Repairing oil sensor	Over load
4	Replacing contactor	Damaged engine	False Alarm
5	Repairing fuse base	Dynamo	Fire alarm restarting

Table 1.1 is example, it shows the potential of the attributes for categorical data clustering in a real life categorical data set, the partitions defined by attributes differ as that in the above example; on the other hand, the objects in the same real clusters (classes) must have distinct value on some attributes from the objects in the other real clusters, consequently there exist some partitions defined by attributes which are close to the real clustering of objects; at least, there exist some equivalence classes (the set of objects which has the same value of the attribute) in these partitions which are close to the real clusters. Such partitions should share as much as possible information with the partitions defined by other attributes. The aim is to find such as equivalence classes and partitions to construct the clustering of the objects.

In this study, a notion information system based hierarchical divisive clustering algorithm for categorical data, called MGR is proposed. The object of Mean Gain Ratio (MGR) algorithm is to search some equivalence classes in the partitions defined by attributes as the clusters of the objects [16]. The initial step of Mean Gain Ratio (MGR) is to find clustering attribute. Clustering attribute is such an attribute that the partitions defined by it share the most information with the partitions defined by other attributes. In our algorithm, the information system-based notion of mean gain ratio (MGR) is used to determine the clustering attribute. The second step is to find objects groups by using divide and conquer method.

1.2 Data Mining

Data mining is the process of extracting interesting patterns and knowledge from a large amount of data. The data sources may be databases, data warehouses, Web documents, other information repositories, or the data streamed into the system dynamically [21]. Data mining also represents the intersection of many interdisciplinary such as machine learning, information retrieval, pattern recognition, data warehouse, statistics, database system and visualization [21]. In addition data mining contains several models and algorithms, i.e. association rule, clustering (Unsupervised Learning), classification (Supervised Learning), and etc. Supervised learning algorithms such as regression and classification “predictive” model. However, unsupervised data mining model is based on clustering "descriptive" rough set theory mathematical tools suggested by Pawlak [2].

1.3 Categorical Data Clustering Using Rough Set

The main aim of the Rough Set Theory (RTS) is to cluster dataset objects into groups depending on the clustering attribute chosen by the algorithm used. Many algorithms are used to decide the clustering attribute, VPRS (Variable Precision Rough Sets [13]), ITDR (Information-Theoretic Dependency Rough Set [15]), MTMDP (Total Mean Distribution Precision [14]) and MGR (Main Gain Ratio [16]) are some of these algorithms that are going to be discussed, executed and compared in this thesis.

RTS is widely used in many fields like using it for medical uses to analyze diabetic patients’ dataset [22], or for educational uses as to analyze students suffering study’s anxiety dataset [15] or for marketing uses as to analyze manufacturing and marketing applications [23]. In this thesis, we proposed the use of above algorithms in a new field (Electrical Generator Failures) to find the clustering attribute for the dataset.

In Data Mining, it is very important to cluster the objects of a dataset into homogeneous classes. This is a key operation in order to get the knowledge from a huge dataset, thus, Rough Set Theory and Information theory are used for this purpose.

1.4 Objectives of the thesis

The aim of this study is to propose an alternative approach based on data mining Rough Set Theory and Information Theory over the old techniques used based on excel datasheets and manual analysis to analyze maintenance variables in order to determine the clustering attribute and discover the most important variable that leads to electrical generator failures, leading to maintenance costs reduction. This also assists the decision-makers to figure out the most effective variable on the maintenance team performance. Lastly, make a comparison among the proposed data mining algorithms based on purity and F-measure of the resulting clusters and the time required by each algorithm to process the data. in addition, this study aim to propose a new algorithm for clustering attributes.

CHAPTER TWO

Related Work

2.1. Introduction

In this chapter, we review the contributions led to the development of the data clustering methods, the methods proposed to cluster datasets, the challenges that faces the data clustering methods, comparison results among different types of data clustering methods using sample and real life datasets and the factors used to compare these methods.

Despite the fact that there many algorithms that can be used to split objects with similar properties into groups, there are still some challenges that may be faced according to the algorithm capabilities to process uncertain data or to deal with categorical data [15].

A.M. Cruz [24] proposes the use of association rules and clustering methods to enhance the efficiency of medical equipment maintenance for the engineering facility in a hospital by finding the most causes of maintenance requests and the real causes of failures.

A. Maquee, A.A. Shojaie and D. Mosaddar [25] use k-means algorithm to cluster a bus maintenance data into homogenous clusters, then uses the Apriori algorithm to identify the causes for each record to lead the maintenance team to modify their maintenance schedules in a way that isolates severe conditions in separate groups.

Association rules method is used to analyze datasets that are related to maintenance, but it is a time consuming method according to the fact that it iterates through the dataset repeatedly until it concludes the results and it may not reach a convergence point, thus, results are not always guaranteed [15].

Pawlak [2] introduces the Rough Set Theory (RST) as a data clustering method that splits the objects of a huge dataset into groups depending on the attributes in order to find the knowledge of the dataset even without the existence of experience.

T. Herawan [26] shows that Rough Set Theory (RST) can be used to cluster two datasets that are related to cancer diseases by the dependency of the dataset on the attributes.

L Shenb, F.E.H. Taya[27] diagnoses valve error in a diesel engine that has multiple cylinders using Rough Set Theory (RST) by discretizing the attributes of the fault states in order to sort the faults or to analyze the dynamic attribute of the engine.

T. Herawan, R. Ghazali, I.T.R. Yanto and M.M. Deris [28] compares the use of two different Rough Set Theory (RST) algorithms to analyze two sample datasets and used the computational complicity and purity to measure the performance of each algorithm and compare them in order to decide the algorithm with better results.

L.J. Mazlack, A. He, Y. Zhu and S. Coppock [29] presents Total Roughness (TR) algorithm which is a Rough Set Theory (RST) algorithm that depends on calculating the total mean roughness for each sub-partition of values in every attribute in order to choose the clustering attribute used to cluster the dataset objects into groups.

D. Parmar, T. Wu and J. Blackhurst [30] proposes the use of Min-Min Roughness (MMR) algorithm based on Rough Set Theory (RST) for categorical data clustering with the ability to cluster uncertain dataset.

T. Herawan, J. H. Abawajy and M.M. Deris [31] introduces a new Rough Set Theory (RST) based algorithm which is Maximum Dependency Attribute (MDA) to split the dataset objects into groups in order to support decision making for complex fields with huge dataset and compare it with the Total Roughness (TR), Min-Min Roughness (MMR) algorithms regarding their accuracy and complexity.

Ziarko [32] introduced the Variable Precision Rough Set (VPRS) algorithm as a Rough Set Theory (RST) method that is capable of tolerating errors so it can overcome the uncertainty problem effectively.

Slezka and Ziarko [33] proposes that the VPRS algorithm is capable of data errors removal and noise resistance.

I.T.R. Yanto, P. Vitasari, T. Herawan and M.M. Deris [13] applied the VPRS algorithm over a real time dataset of students describing the factors causing anxiety for them in order to reduce these factors, enhancing the academic performance of the student.

T. Herawan and W.M.W. Mohd [22] shows that the Variable Precision Rough Set (VPRS) has the highest purity compared to Total Roughness (TR), Min-Min Roughness (MMR) and Maximum Dependency Attribute (MDA) when applied to a real life diabetics dataset.

T. Beaubuof, F.E. Petry and G. Arora [34] mentions that Shanon developed the information theory as a communication theory. This theory is widely used to characterize the datasets that has uncertain information through representing this information by rough entropy in all dataset kinds.

P. Kumar and B.K. Tripathy [9] modify the MMR algorithm to enhance the results creating a new Rough Set Theory (RST) based algorithm (MMeR).

B.K. Tripathy and A. Ghosh [35] propose a new Rough Set Theory (RST) algorithm called Standard-Deviation Roughness (SDR) that is able to handle non-homogenous data even if it contains uncertain information.

B.K. Tripathy and A. Ghosh [36] introduce the Standard-Deviation of Standard-Deviation of Roughness (SSDR) as a new Rough Set Theory Algorithm that handles categorical and numerical data.

I. Park and G. Choi [15] introduces the Information-Theoretic Dependency Roughness (ITDR) based on the measurement of the rough entropy and compares it to other Rough Set Theory (RST) algorithms that are Min-Min Roughness (MMR), (MMeR), Standard-Deviation Roughness (SDR), Standard-Deviation of Standard-Deviation of Roughness (SSDR) and Information-Theoretic Dependency Roughness (ITDR). Their comparison was based on the purity factor of the cluster groups using a UCI sample data and shows that the ITDR method has the higher purity among the algorithms under investigation.

K-means algorithm is a clustering method that can efficiently handle huge datasets. This algorithm is capable of processing only numerical datasets. In order to provide the ability to process real life datasets, Z. Huang [37] presents the k-modes algorithm which is an extension to the k-means algorithm that has the ability to handle categorical datasets.

M. Li, S. Deng, L Wang, S. Feng and J. Fan [14] proposes a new Rough Set Theory (RST) based algorithm called Maximum Total Distribution Precision (MTMDP) to cluster categorical dataset with the capability to handle uncertainty. Furthermore, the MTMDP is compared to Min-Min Roughness (MMR), which is a Rough Set Theory (RST) based algorithm, and few non-RST algorithms that are k-modes, fuzzy k-modes and fuzzy centroids by comparing the overall purity for each algorithm to the others'.

Z. He, X. Xu and S. Deng [38] presents the mutual information based algorithm k-ANMI which processes the data in a way that is very close to the way the k-mean algorithm does and makes use of each step's cluster by using the Average Normalized Mutual Information (ANMI) criterion which is based on mutual information to cluster categorical datasets.

Z. He, X. Xu and S. Deng [39] introduces another clustering algorithm that is also based on mutual information and uses the Average Normalized Mutual Information (ANMI) and is capable of clustering categorical datasets called (G-ANMI).

D. Barbara, J. Couto and Y. Li [40] presents the data clustering algorithm COOLCAT which is capable of clustering real-time data without the need to review the data clustered earlier. This algorithm relies on calculating the entropy to investigate the clustering attribute.

H. Qin, X. Ma, T. Herawan and J.M. Zain [16] proposes the new clustering algorithm Mean Gain Ratio (MGR) which is based on the information theory for categorical data clustering. Furthermore, the MGR algorithm is compared to the MMR, k-ANMI, G-ANMI and COOLCAT algorithms regarding the execution time and accuracy.

[41] defines the purity as a measure for the number of objects shared between classes and clusters. Higher purity means that the resulting structure of the cluster groups reflects the class structure more accurately. Furthermore, the precision is defined as the relation between the number of objects shared by clusters and classes with respect to the total number of object in that cluster; while the recall is the relation of the shared object between clusters and classes with respect to the total number of objects in class.

The F-measure is derived from the values of the precision and recall for the entire dataset.

[22] shows that VPRS algorithm has the highest purity compared to MMR, MDA and TR algorithms when applied to the real life data for diabetics.

[15] shows that applying the k-means, Fuzzy k-means, Fuzzy Centroids, SDR, SSDR, MMR and ITDR algorithms to the UCI machine learning dataset (Zoo dataset) [17] results that the ITDR algorithm has the highest purity in the comparison.

[14] shows the comparison between the MTMDP and MMR based on purity when applied to many UCI machine learning [18] datasets. The comparison shows that the MTMDP algorithm has higher purity than the MMR in all the datasets used.

[16] compares the MGR algorithm to MMR, k-ANMI, G-ANMI and COOLCAT algorithms. The comparison shows that the MGR has the highest purity among the compared algorithms.

As mentioned earlier, the VPRS, ITDR, MTMDP and MGR algorithms has the best results with respect to the algorithms they are compared to in each comparison. In this thesis, we compare these algorithms using sample data and real life data in order to measure the purity, execution time and F-measure for each algorithm to find which algorithm results the best clusters.

CHAPTER THREE

DATA CLUSTERING

3.1. Data Clustering.

Data clustering is a data mining process that divides the objects of a dataset into groups in a way that each object in a group is more related to the objects sharing the same group than the objects in the other groups. The clustering process is one of the most important data mining processes because of its ability to discover groups with interesting distributions in the datasets [42], thus, it is a key function of the Knowledge Discovery of Data (KDD) which results the useful knowledge from a huge dataset [43]

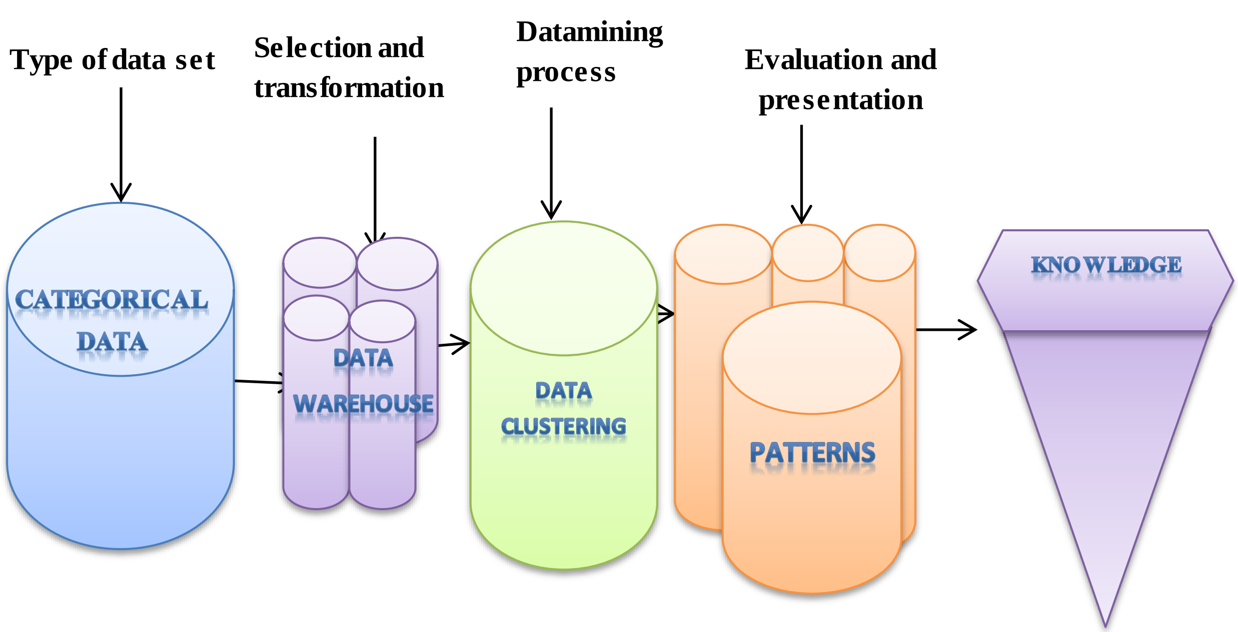


Figure 3.1: The KDD Process.

In order to find the cluster groups, it is important to examine the relations among the attributes so the clustering attribute can be chosen. There are many algorithms proposed to achieve that. Four important clustering attribute selection algorithms are discussed in this study; these algorithms were chosen according to their similarity and the fact that each algorithm is chosen as best results when compared to other algorithms as mentioned in chapter 2. Our contribution is to compare these algorithms on the basis of Purity, F-measure and execution time using sample datasets from UCI and real life dataset collected for electrical generators failures. These algorithms can be divided into two groups by the theory they belong to.

3.2.Rough Set Theory

The basic concepts of the Rough Set Theory can be defined by means of operation, closure and interior called approximations.

3.2.1 Information System[11]:

An information system is four - tuple (quadruple) $S = (u, B, \nu, f)$, where $u = \{s_1, s_2, s_3, \ldots, s_n\}$, $|u| = n$, u is the set of finite objects and $u \neq \emptyset$ called universe ,where $B \neq \emptyset$, B is a finite set of attributes, ν is a set of values set where ν $= \bigcup_{b \in B} \nu_b$, ν_b is represent the domain of attribute b. f is an information function denoted by $f: u \times B \rightarrow \nu$, $f(s, b)$ belong to ν_b, $\forall (s, b) \in u \times B$,such as database contains 636 sites, i.e. u=636, and 38attributes, i.e. $B = \{b_1, b_2, b_3, \ldots, b_{38}\}$ $i.e$ $|B| = 38$, This can be illustrated in terms of an information system table to choose clustering attribute supported the rough set algorithms as within the following table 3.1

Table 3.1: An information system

u=sites	$b1$	$b2$	………………………………………	$b_{\|B\|}$
s_1	$F(s_1,b_1)$	$F(s_1,b_2)$	………………………………………	$F(s_1,b_m)$
s_2	$F(s_2,b_1)$	$F(s_2,b_2)$	………………………………………	$F(s_2,b_m)$
s_3	$F(s_3,b_1)$	$F(s_3,b_2)$	………………………………………	$F(s_3,b_m)$
. .	. .	. .	……………………………………… ………………………………………	. .
636	$F(s_n,b_1)$	$F(s_n,b_2)$	………………………………………	$F(s_n,b_m)$

The initial point of rough set approximations is the indiscernibility relation, which is generated by information about objects of interest.

3.2.2 Example

The table 3.2 is an information system of six sites and three units valued attributes: electrical, mechanical and Site management

Table 3.2 Information system of six sites and three sources

site	Electric source	Mechanic source	Site management source
1	0	2	4
2	1	2	5
3	1	3	6
4	0	3	5
5	7	3	5
6	7	2	6

The universe u={1,2,3,4,5,6}

And attribute B={ Electric unit Mechanic unit ,Site management unit }

V_{Electric} ={0, 1,7}

V_{Mechanic} ={3, 2}

$V_{\text{site management}}$ ={4,6 ,5}

3.2.3.Indiscernibility Relation

$S = (u, B, v, f)$ represent an information system and $A \subseteq B$. let x, y be an element belong to universe u is called to be A-indiscernible (indiscernible by the set of attribute A subset of B in *information system)* iff $f(x,b) = f(y,b)$, $\forall b \in A$. clearly , each subset of B induces unique indiscernibility relation. Note that, an indiscernibility relation induced by the set of attribute A, denoted by *IND(A),* is an equivalence relation. It is well known that, an equivalence relation induces unique partition. The partition of *universe* induced by *IND(A)* in $S = (u, B, v, f)$ denoted by u/A and the equivalence class in the partition u/A containing $x \in u$, denoted by $[x]_A$. The concept of upper and lower approximations of a set can be defined as follows.

3.2.4 Set Approximations[44]:

$S = (u, B, v, f)$ be an information system, let $A \subseteq B$ and $X \subseteq u$. The A _lower approximation of X, denoted by $\underline{\boldsymbol{A}}(X)$ and A _upper approximations [19], denoted by $\overline{\boldsymbol{A}}(X)$of X, respectively, are defined by

$\underline{A}(X) = \{x \text{ belong to } u | [x]_A \subseteq X\}$,

$\overline{A}(X) = \{x \text{ blong to } u \,|[x]_A \cap X \neq \emptyset\}$.

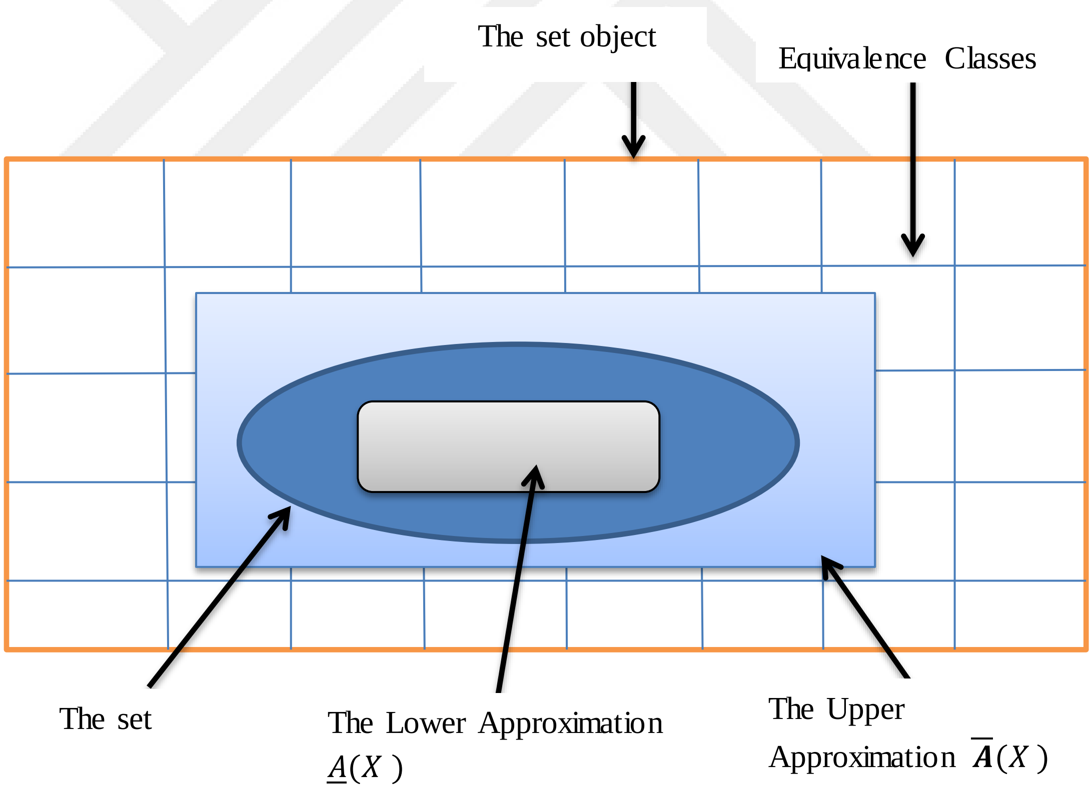


Figure 3.2: Set of approximation

3.2.5 Example

In above table 3.2, consider attribute electric unit

The set X(electric=0)={1,4}

The partition of *u* induce IND (electric)

u/electric={{1,4},{2,3},{6,5}}

3.3 Variable precision rough set algorithm(VPRS)

In this algorithm, variable precision of attributes is used to find the accuracy of approximation in rough set. Variable precision of attributes is used to find the accuracy of approximation in order to select the clustering attribute.

3.3.1. Error classification

Let a set u as a universe and x,*y subset u* , wherever *x,y are a non-empty* . The error classification rate of *x* relative to *y* is denoted by *Er(x ,y)*, is defined by

$$Er(x,y) = \begin{cases} 1 - \frac{|x \cap y|}{|x|}, & |x| > 0 \\ 0 \ , & |x| = 0 \end{cases} \quad (3.1)$$

3.3.2.Upper approximation and Lower approximation

Let u be a finite dataset and the real number δ and $\delta \in [0,0.5)$ and *Y* is a subset of u. The A_δ_lower approximation of Y, denoted by $\underline{A}_\delta$ (Y) and A_δ _upper approximation of *Y*, denoted by $\overline{A}_\delta$ (Y), respectively, are defined by

$$\underline{A}_\delta (Y) = \{y \in : Er([y]_A, Y) \leq \delta\} \quad (3.2)$$

and

$$\overline{A}_{\delta}\ (Y\ ),\ = \{x \in\ :\ Er([y]_A\ , Y\ ) < 1\text{-}\ \delta\ \} \tag{3.3}$$

The set $\underline{A}_{\delta}\ (Y\ )$ is called the positive region of Y which is the set of objects of u that can be classified into Y and error classification rate less than or equal to δ . This results in $\underline{A}_{\delta}\ (Y\ )\ \subseteq \overline{A}_{\delta}\ (Y\ )$, when $0 \le \delta < 0.5$, so the meaning of the upper and lower approximations is maintained.

3.3.3.Accuracy of approximation VPRS

The accuracy of approximation variable precision (accuracy of variable precision roughness) of any set Y subset of u w.r.t A subset of B is denoted by $\alpha_{A\delta}\ (y)$ is calculated as

$$\alpha_{A_{\delta}}(Y) = \frac{|\underline{A}_{\delta}\ (Y\ )|}{|\overline{A}_{\delta}\ (Y\ )|} \tag{3.4}$$

where $|Y|$ represents cardinality of Y. If $\delta = 0$, it is the traditional rough set algorithm of Pawlak

clearly , $\alpha_{A_{\delta}}(Y) \in [0,1]$, if $\alpha_{A_{\delta}}(Y)$=1 then Y is crisp with respect to A (Y is precise with respect to A), and otherwise, if $\alpha_{A_{\delta}}(Y)$=1 ,Y is rough with respect to A ,(Y is vague with respect A)

3.3.4 Proposition:

$S = (u, B, v, f)$ be an information system, $\alpha_A(Y)$be an roughness accuracy, $\alpha_{A_{\delta}}(Y)$is a variable precision roughness accuracy and given δ the variable precision error factor. If $(0 \le \delta < 0.5)$, then $\alpha_A(Y) \le \alpha_{A_{\delta}}(Y)$

3.3.5 Mean Accuracy of VPRS algorithm (MAC)

Suppose $b_i \in B$, $v(b_i)$ has r- different values, i.e. γ_r , r= 1,2,…,m and $Y(b_i = \gamma_r)$, r= 1,2,….,m is an objects subset having r- different values of attribute b_i . The accuracy of the set $Y(b_i = \gamma_r)$,r =1,2,….,m for given δ error factor, with respect to b_j , where $i \neq j$, denoted $\alpha_{\delta b_j}(Y \mid b_i = \gamma_r)$, is found by

$$\alpha_{\delta b_j}(Y \mid b_i = \gamma_r) = \frac{\left|\underline{A}_\delta Y_{b_j}(b_i = \gamma_r)\right|}{\left|\overline{A}_\delta Y_{b_j}(b_i = \gamma_r)\right|}, r = 1,2, \ldots., m \tag{3.5}$$

The mean accuracy of attribute $b_i \in B$ with respect to $b_j \in B$, where $i \neq j$, denoted by $MAC_{b_j}(b_i)$, is calculated as follows :

$$MAC_{b_j}(b_i) = \frac{\sum_{r=1}^{|v(b_i)|} \alpha_{\delta b_j}(Y \mid b_i = \gamma_r)}{|v(b_i)|} \tag{3.6}$$

Where $|v(b_i)|$ are the values set for the attribute $b_i \in B$.

3.3.6 Mean Average of VPRS algorithm (MA)

Given n attributes, mean accuracy of attribute $b_i \in B$ with respect to $b_j \in B$, where $i \neq j$, refers to the average of $MAC_{b_j}(b_i)$, denoted $MA(b_i)$, is evaluated by the formula

$$MA(b_i) = mean\left(MAC_{b_i}(b_i)\right), 1 \leq i , j \leq m . \tag{3.7}$$

3.3.7 The pseudo-code of VPRS algorithm

Algorithm: VPRS

Input: Data set

Output: Clustering attribute

Begin

Step 1.calculation the equivalence classes using the indiscernibility relation on each attribute.

Step 2. Calculate the Error classification *(Er)* of attribute b_i w.r.t all b_j, where *i isn't equal to j .*

Step 3. Calculate the $\underline{A}_{\delta}$ (Y) and $\underline{A}_{\delta}$ (Y) of attribute b_i w.r.t all b_j, where *i is not equal to j .*

Step 4. Calculate MAC of attribute b_i w.r.t all b_j, where i is not equal to j .

Step 5. choose a clustering attribute depended on the maximum MAC of attribute.

End

3.3.8 Example of VPRS algorithm

In above table (3.2) is information system of 6 site with 3 units valued attributes : electric , mechanic and Site management , there is no decision attribute defined a clustering then we will choose a clustering attribute among all candidates to get the value of the Variable precision rough set, the first step, we must get the equivalence classes.

Induced by indiscernibility relation of singletons attribute the three partitions of objects from table 3.2 are shown as follow:

1-X (Electric =0)={1,4}

2- X (Electric = 1)={2,3}

3-X (Electric =7)={5,6}

u / Electric = {{1,4},{2,3}{6,5}}

1- X (Mechanic = 2)={1,2,6}

2- X(Mechanic = 3)={3,4,5}

u / Mechanic = {{1,2,6},{3,4,5}}

1- X(Site management =4)={1}

2- X(Site management = 6)= {3,6}

3- X(Site management =5)={2,4,5}

u / Site management ={{1},{3,6},{2,4,5}}

By using the Formulae (1) attribute Mechanic w. r. t electric attribute is obtain as follows:

Er(0, 2)= $1 - \frac{|\{1\}|}{|\{1,4\}|} = 0.5$,

Er(1 , 2)= $1 - \frac{|\{2\}|}{|\{2,3\}|} = 0.5$

Er(7 , 2)= $1 - \frac{|\{5\}|}{|\{5,6\}|} = 0.5$

Er(0 , 3) = $1 - \frac{|\{4\}|}{|\{1,4\}|} = 0.5$,

Er(1,3)= $1 - \frac{|\{4\}|}{|\{2,3\}|} = 0.5$

Er(7 , 3)= $1 - \frac{|\{5\}|}{|\{5,6\}|} = 0.5$

by given δ=0.4 , the A_δ-lower and A_δ- upper approximation are:

$\left|\underline{A}_\delta(Mechanic = 3\ )\right| = |\{\emptyset\}| = 0$

$\left|\underline{A}_\delta(Mechanic = 2)\right| = |\{\emptyset\}| = 0$

$|\overline{A}_\delta\ (Mechanic = 3\ )| = |\{1,4,2,3,5,6\}| = 6$

$\left|\overline{A}_\delta(Mechanic = 2)\right| = |\{1,4,2,3,5,6\}| = 6$

The MAC of attribute Mechanic w.r.t electric are

$$MAC = \frac{\alpha_{0.4\ electric}\ (X|Mechanic = 2) + \alpha_{0.4\ electric}\ (X|Mechanic = 3)}{2} \quad = 0$$

by using the same steps, the MAC for each attribute w.r each to the Site management are computed, these calculation are summed up in table 3.3

Table 3.3: Maximal mean accuracy of VPRS algorithm

Attribute w.r.t	MAC		MA
Electric	Mechanic 0	Site management 0.33333	0.16667
Mechanic	Electric 0	Site management 0.4667	0.2333333
Site management	Mechanic 0	Electric 0.3333	0.16667

The VPRS algorithm from table 3.3, the attribute with the highest mean accuracy is attribute (Mechanic), so, the attribute Mechanic is select as clustering attribute .For object splitting, we use the divide -conquer method we can find cluster ,the objects depend on decision attribute selected ,note that equivalent classes of the attribute Mechanic is

u / Mechanic ={{1,2,6},{3,4,5}},

3.4. Maximum Total Mean Distribution Precision Algorithm (MTMDP)

Starting with the concept of distribution approximation precision which is derived from rough membership, MTMDP algorithm investigates the clustering attribute depending on the mean distribution precision (MDP) and the total mean distribution precision (TMDP)

3.4.1 Probabilistic Distribution Approximation[14]:

$S = (u, B, v, f)$ represent information system ,C ⊆B ,the rough membership value of an object $y \in Y$,Y subset of u and Y is non-empty, the probability of the object in Y given that the object is in $[y]_{C,}$ is the probabilistic interpretation of rough membership of an object$y \in Y$ *is denoted as* $\tau_y^C(y)$,

$$\tau_y^C(y) = p(Y|[y]_C) = \frac{|[y]_C \cap Y|}{|[y]_C|} \tag{3.8}$$

$\tau_y^C(y)$ represent Probabilistic Distribution Approximation set , and the Probabilistic Distribution Approximation set of Y based on attribute set C is :

$$\bar{C}^d(Y) = \left\{\frac{\tau_y^C(y)}{y}, y \in Y\right\} \tag{3.9}$$

Where "d" denoted the distribution approximation, $\bar{C}^d(Y)$is represented probabilistic rough set of Y and the member of each object y in using its rough membership value in Y given that the object is the equivalence class $[y]_C$

3.4.2The distribution approximation precision

$S = (u, B, v, f)$ represent information system ,and Y subset of universe u ,Y is a non –empty ,The distribution approximation precision of Y by the attribute set B is defined as follows

$$R_C^d(Y) = \frac{|\bar{C}^d(Y)|}{|Y|} = \frac{\sum_{y\in Y} \tau_y^C(y)}{|Y|} = \frac{1}{|Y|} \sum_{x \in u/B} \frac{|x \cap y|^2}{|x|} \tag{3.10}$$

clearly,$0 \le R_C^d(Y) \le 1$,if $R_C^d(Y) = 1$then Y is crisp

3.4.3. Mean Distribution Precision[14]

$S = (u, B, v, f)$ represent information system. $v(b_i)$ is the values of attribute b_i, then, the MDP of attribute b_i (b_i belong to B) w.r.t attribute b_j, where $b_i \neq b_j$, is defined as :

$$MDP_{b_j}(b_i) = \frac{\sum_{y \in u/\{b_i\}} R^d_{\{b_j\}}(Y)}{|v(b_i)|} \tag{3.11}$$

where $|v(a_i)|$ is the count of different values of attribute b_i, $MDP_{b_j}(b_i)$ consider the mean distribution precision of the equivalence classes induced by b_i w.r.t b_j. $MDP_{b_j}(b_i)$ ranges from 0 to 1.If $MDP_{b_j}(b_i) = 1$, then every equivalence class of $u/IND\{b_i\}$ is crisp w.r.t b_i.

3.4.4 Maximum Total Mean Distribution Precision

$S = (u, B, v, f)$ represent information system, then the total mean distribution precision (TMDP) of attribute b_i (b_i $belong\ to\ B$) is:

$$TMDP(b_i) = \frac{\sum_{\substack{b_j \in B \\ b_i \neq b_j}} MDP_{b_j}(b_i)}{|B| - 1} \tag{3.12}$$

$$MTMDP(b_i) = Max_{bi \in B}\big(TMDP(b_i)\big), i = 1,2, \dots, n \tag{3.13}$$

$TMDP(b_i)$ represent the total mean distribution precision of the equivalence classes by attribute b_i.the range of $TMDP(b_i)$ from 0 to 1, clearly, $TMDP(b_i)$ include the total coupling between the equivalence classes by attribute b_i.

3.4.5.The pseudo-code of MTMDP algorithm

Algorithm: MDMTP

Input: Data Set

Output: Clustering attribute

Begin

1. Calculation the equivalence classes utilized the indiscernibility relation on each attribute.

2. Calculate Probabilistic Distribution Approximation($\tau_y^C(y)$)b_i w.r.t all b_j, where *i isn't equal to j* .

3.Calculate The distribution approximation precision$R_C^d(Y)$of attribute b_i w.r.t all b_j, where *i isn't equal to j*.

4. Calculate $MDP_{b_j}(b_i)$ of attribute b_i w.r.t all b_j, where *i isn't equal to j*.

5.Calculate $TMDP(b_i)$ of attribute b_i w.r.t all b_j, where *i isn't equal to j*.

6. Choose a clustering attribute based on the maximum$Max_{bi\in B}\big(TMDP(b_i)\big)$ of attribute.

End

3.4.6 Example of MTMDP

In above table (3.2)

Step 1:

1- X (electric =0)={1,4}
2- X (electric = 1)={2,3 }

3- X (electric =7)={5,6}

u / electric = {{1,4},{2,3}{6,5}}

3- X (Mechanic = 2)={1,2,6}

4- X(Mechanic = 3)={3,4,5}

u/ Mechanic = {{1,2,6},{3,4,5}}

2- X(Site management =4)={1}

3- X(Site management = 6)= {3,6}

4- X(Site management =5)={2,4,5}

u / Site management ={{1},{3,6},{2,4,5}}

Step2 . Calculate Probabilistic Distribution Approximation($\tau_x^C(x)$)

Table 3.4: Calculate Probabilistic Distribution Approximation

u	1	4	2	3	5	6
Electric source	0	0	1	1	7	7
Mechanic source	2	3	2	3	3	2
$\tau_{electric}^{Mechanic}(x)$	1/3	1/3	1/3	1/3	1/3	1/3

Step3.

$$R_C^d(Y) = \frac{|\bar{C}^d(Y)|}{|Y|} = \frac{\sum_{y\in Y}\tau_y^C(y)}{|Y|}$$

$$R_{Mechanic}^d(Y|Electric = 0) = \frac{\sum_{y\in Y}\tau_y^C(y)}{|Y|} = \frac{1/3 + 1/3}{|\{1,4\}|} = \frac{2/3}{2} = 0.333$$

$$R_{Mechanic}^d(Y|Electric = 1) = \frac{\sum_{y\in Y}\tau_y^C(y)}{|Y|} = \frac{1/3 + 1/3}{|\{2,3\}|} = \frac{2/3}{2} = 0.333$$

$$R_{Mechanic}^d(Y|Electric = 7) = \frac{\sum_{y\in Y}\tau_y^C(y)}{|Y|} = \frac{1/3 + 1/3}{|\{6,5\}|} = \frac{2/3}{2} = 0.333$$

Step 4 :

Calculation MDP of attribute *Electric* w.r.t all *Mechanic*,

$$MDP_{Mechanic}\ (Electric) = \frac{\sum_{y \in u/Mechanic} R^{d}_{Mechanic}(Y)}{|v(electric\ )|}$$

$$MDP_{Mechanic}\ (Electric) = \frac{0.333 + 0.333 + 0.333}{3} = 0.333$$

Table 3.5: Calculate the mean distribution precision

	$X_1(0)$	$X_2(1)$	$X_3(1)$	MDP
With respect to Mechanic	0.333	0.333	0.333	0.333
With respect to Site management	0.665	0.417	0.417	0.5

Step 5:

choose a clustering attribute depend on the maximum TMDP of attribute.

$$TMDP(electric) = \frac{\sum_{\substack{Mechanic \in B \\ b_i \neq b_j}} MDP_{Mechanic}(electric)}{|B| - 1} = (0.333 + 0.5)\backslash 2 = 0.417$$

All value MDP and TMDP

Table 3.6: Maximum Total Mean Distribution Precision

Attribute	mean distribution precision(MDP)			TMDP
	Electric	Mechanic	Site management	
Electric	0	0.333	0.5	0.417
Mechanic	0.5	0	0.611	0.55556
Site management	0.5	0.4047	0	0.453704

The attribute Mechanic has the maximum TMDP, then Mechanic attribute is selected as clustering attribute .

3.5 Information theoretic dependency roughness (ITDR)[15]:

Information theoretic dependency roughness (ITDR) can handle uncertainty in categorical data for clustering categorical data that deals with uncertainty as well. The rough set is applied to determine the clustering attribute based on the rough measure entropy [34] from all candidate attributes in dataset.

3.5.1 Definition :

$S = (u, B, v, f)$ is approximation space, and let $X, Y \subseteq B$,attribute Y totally based on attribute X denoted $X \Rightarrow Y$ if all values of attribute Y are determined uniquely by attributes values X, in other word attribute Y totally based on attribute X if a functional dependency between values Y and X exists, the following definition describes the generalized attribute dependency notion:

3.5.2 Information-theoretic dependency roughness (ITDR)

$S = (u, B, v, f)$ is approximation space, and $X, Y \subseteq B$,and X, Y are a non-empty. Information-theoretic dependency roughness (ITDR) of attribute Y on attributes X, denoted X⇒ Y, is defined by the following equation:

$$H(y_i|x_j) = \begin{cases} -\sum_{j=1}^{n} \frac{|x_j|}{|u|} log \frac{|x_j \cap y_i|}{|x_j|}, |x_j \cap y_i| > 0 \\ 1 \qquad\qquad\qquad\qquad |x_j \cap y_i| = 0 \end{cases} \tag{3.13}$$

Where $H(y_i|x_j)$ is a function from A, clearly $H \in [0,1]$, where H depicts the value of $H(y_i|x_j)$. Attribute y_i is said to depend totally (in a degree of H) on the attribute x_j if H=1, in the other word, y_i depends partially in x_j. Thus, attribute y_i depends totally (partially) on attribute x_j, if all (some) element of the universe u can be classified uniquely into equivalence classes of the partition u/y_i, employing x_j.

3.5.3 Min-roughness ITDR algorithm

Suppose b_i belongs to B, $v(b_i)$ has s-different values ,i.e. σ_s , $s = 1,2,..,m$. Let $y(\mathrm{bi} = \sigma_s)$be an object subset that is having s- different values of attribute b_i. Min roughness b_j of set $y(\mathrm{bi} = \sigma_s)$ $w.r.t$ bj, where i is not equal to j denoted $MRH(y_{i[\delta]}|x_j)$ is described by

$$MRH(y_{i[\delta]}|x_j) = \min(H(y|\mathrm{bi} = \delta)|x_j)) \tag{3.14}$$

3.5.4 Min- Mean- roughness ITDR algorithm

Min mean roughness of attribute of b_i w.r.t b_j, b_i and b_j belong to B, where i is not equal to j denoted by MMRH($b_i|b_j$) is calculated

$$\mathrm{MMRH}(y_i|x_j) = MRH(y_{i[\delta]}|x_j) + \cdots + MRH(y_{i[\mathrm{bi}_{|v(\mathrm{bi})|}]}|x_j)/|v(\mathrm{bi})| \tag{3.15}$$

$|v(\mathrm{bi})|$ *is represented values of attribute* bi belong to B

3.5.5 Min- Mean- Min- roughness ITDR algorithm

Given m attributes, min-mean –min –roughness of attribute b_i belongs to y, w.r.t b_j belongs to x, where i is not equal to j refers to min of MMRH($b_i|b_j$), denoted MMMRH($y_i|x_j$) is calculated using the equation

$$\text{MMMRH}(y_i|x_j) = \min\left(MMRH(y_1|x_1), \ldots \ldots, MMRH(y_m|x_m)\right) \tag{3.16}$$

The ITDR algorithm choose partition attribute based on the mean degree of rough entropy, more accuracy for partitioning attribute selection is implied by the rough entropy with the higher degree while clustering crispness is higher when the mean roughness is lower. ITDR determines the clustering attribute.

3.5.6 The pseudo-code of ITDR algorithm

Algorithm: ITDR

Input: Data set

Output: Clustering attribute

1: calculation the equivalence classes utilized the indiscernibility relation on each attribute.

2: calculate the entropy $H(y_i|x_j)$ of attribute b_i w.r.t all b_j, where i is not equal to j

3: calculate the Min roughness $MRH(y_{i[\delta]}|x_j)$ and Min mean roughness of attribute of b_i w.r.t b_j

4: choose a clustering attribute depended on the Max (Min entropy value on $MRH(y_{i[\delta]}|x_j)$) degree of dependency of attribute.

End

3.5.7 Example for ITDR algorithm

In table 3.2 there are six sites ($|u|$=6) with three value attribute ($|B|$=3). To get the ITDR of all attribute, the initial step of the algorithm is to get the equivalence classes in the same the example (VPRS-step 1).

Objects can be partitioned depending on the equivalence classes collected. Table 3.2 shows these partitions.

Formula (3.13) can be used to obtain the dependency degree of attribute. For attribute electric unit depends on attributes Site management unit and Mechanic unit. The mean roughness of electric attribute w.r.t Mechanic is calculate by using definition information theoretic dependency measure X(electric=0) w.r.t X(Mechanic=2),where X(electric=0)=y1={1,4}, X(Mechanic=2) =x1 ={1,2,6}

$$H(y_1|x_1) = -\frac{|\{1,2,6\}|}{|\{1,2,3,4,5,6\}|} log \frac{|\{1,4\} \cap \{1,2,6\}|}{|\{1,2,6\}|} = -\frac{3}{6} log\left(\frac{1}{3}\right) = 0.5493$$

The ITDR measure of X(electric=0) w.r.t X(Mechanic=3) ={3,4,5}is

$$H(y_1|x_2) = -\frac{|\{3,4,5\}|}{|\{1,2,3,4,5,6\}|} log \frac{|\{1,4\} \cap \{3,4,5\}|}{|\{3,4,5\}|} = -\frac{3}{6} log\left(\frac{1}{3}\right) = 0.5493$$

The ITDR measure of X(electric=0) with respect to X(Mechanic=2)and X(Mechanic=3) are 0.5493, 0.5493respectively, according to Formula(3.14)

The Min-roughness ITDR of X(electric=0) w.r.t X(Mechanic) is 0.5493,and X(electric=1) with respect to X(Mechanic) is 0.5493, according to Formula(3.15) the Mean – Min roughness on electric attribute w.r.t Mechanic is 0.5493, and repeat the same steps, the mean roughness on electric w.r.t (Site management attribute) these calculation are summarized in table 3.5 similar calculation are performed for all attributes

Table 3.7: Mean roughness calculation for attribute (Electric)

With respect to	X(Electric =0)	X(Electric =1)	Mean roughness
Mechanic	0.5493	0.5493	0.5493

Table 3.8: shows ITDR technique minimum degree of dependency of attribute Machine

Attribute	Mean roughness		Mean
Electric	Mechanic 0.5493	Site management 0.1540	0.3517
Mechanic	Electric 0.231	Site management 0.1014	0.166
Site management	Electric 0.231	Mechanic 0.4338	0.332

The min mean 0.166 occurs in attribute Mechanic, the Mechanic attribute is chose as clustering attribute. We use the divide-conquer method for objects splitting, the first split is Mechanic attribute which produces two cluster, the first cluster is {1,2,6}and second cluster is{6,5,3}.

3.6. Objects splitting:

Divide-conquer method is used to split the objects into clusters. For example, table (3.2) shows the clusters of the maintenance variables depend on the clustering attribute chosen by the algorithm i.e. Mechanic, in algorithms MTDP, VPRS and ITDR Notice that, partitioning the maintenance variables dataset using the Mechanic attribute as clustering attribute results {1,2,6},{4,5,3}.

we can split the maintenance variables by utilized the hierarchical tree as follows

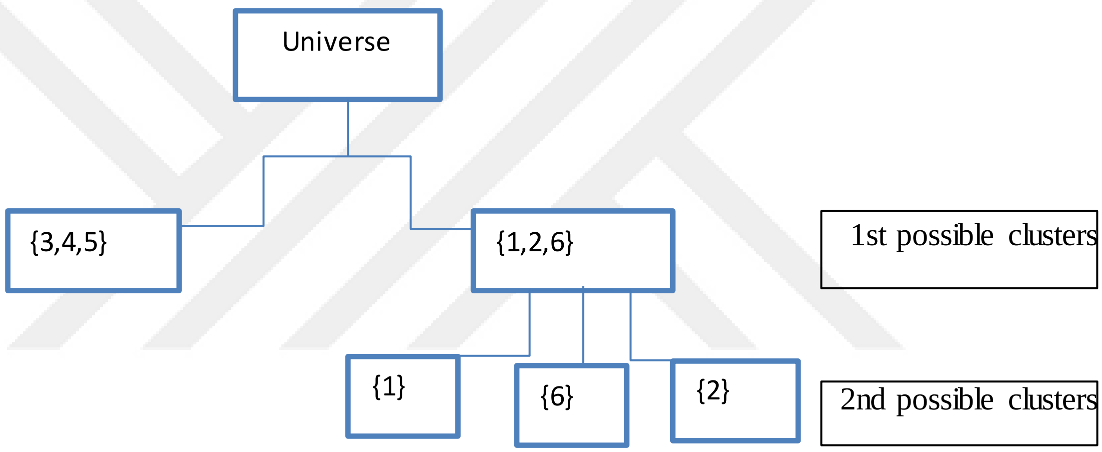


Figure 3.3: clustering results of the VPRS,MTMDP,ITDR algorithms

Furthermore, this technique is repeated by selecting the closest attribute, to the last clustering attribute selected, as a new clustering attribute in order to produce more clusters. The process terminates when a pre-defined number of clusters is reached or all the attributes are used for clustering.

3.7.Information Theory

Information theory is a useful mathematical tool that is used in many fields, such as statistics, mathematics and computer sciences. The information theory relies on the entropy, conditional entropy ,relative entropy and mutual information . These concepts are used and described in the following algorithm [34].

3.7.1.The MGR algorithms[16]

Mean Gain Ratio (MGR) is based on information theory and, it can handle uncertainty in categorical data for clustering categorical data. Mean gain ratio includes determine a clustering attribute and selecting an equivalence class based on the rough measure entropy [16] from all candidate attributes in the dataset. The calculate of MGR by using of some definition as follows:

3.7.1.1.Definition

Let b_i be attribute belong to B, assume $u \setminus b_i = \{x_1, x_2, x_3, \ldots, x_n\}$ the entropy of b_i about the partition is defined as

$$E(b_i) = -\sum_{a=1}^{n} \frac{|x_a|}{|u|} Log_2 \frac{|x_a|}{|u|} \tag{3.17}$$

Where n is domain size of b_i, x_a subset of u is an equivalence class ,a=1,2,….,n

3.7.1.2 Conditional Entropy(cE)

Let b_i, b_j be attributes that belong to B, assume $u \setminus b_i = \{x_1, x_2, x_3, \ldots, x_n\}$,

$u \setminus b_j = \{y_1, y_2, y_3, \ldots, y_m\}$, the conditional entropy(c E) of b_j w.r.t b_i is described as

$$cE_{b_i}(b_j) = -\sum_{h=1}^{m} \frac{|y_h|}{|u|} \sum_{a=1}^{n} \frac{|y_h \cap x_a|}{|y_h|} Log_2 \frac{|y_h \cap x_a|}{|y_h|} \tag{3.18}$$

Where y_h, x_a are subset u ,h=1,2,..,m, a=1,2,..,n

3.7.1.3 Information Gain

Let b_i, b_j be attributes that belong to B, the information gain (IG) of b_i,w.r.t b_j is described as

$$IG_{b_j}(b_i) = E(b_i) - cE_{b_i}(b_j)\,) \tag{3.19}$$

3.7.1.4 Gain Ration

Let b_i, b_j be attributes that belong to B ,the gain ration (GR) of b_i,with respect to b_j is described as

$$GR_{b_j}(b_i) = \frac{IG_{b_j}(b_i)}{E(b_i)} \tag{3.20}$$

3.7.1.5 Mean of Gain Ratio (MGR)

Let b_i be attributes that belong to B, the mean of gain ratio (MGR) of $b_i$is described as

$$MGR_{b_j}(b_i) = \frac{\sum_{\substack{j=1,\\ j\neq i}}^{|B|} GR_{b_j}(b_i)}{|B|-1} \tag{3.21}$$

3.7.2.The pseudo-code of MGR for selecting a clustering attribute

Algorithm: MGR

Input: Dataset

Output: Clustering attribute

Step 1: Calculate entropy of b_i

Step 2: Calculate conditional entropy(c E) of b_j with respect to b_i

Step3: Calculate IG of b_i,with respect to b_j.

Step4: Calculate (GR) of b_i,with respect to b_j

Step5: Calculate the mean of gain ratio (MGR) of b_i

Step 6: choose a clustering attribute depend on the maximum of MGR

End

3.7.3 Example for MGR algorithms

In table 3.2 there are six objects and three attributes, first, the mean of gain ratio for each attribute is calculated by determining the equivalence classes in portion of data set, let's take electric attribute defines a partition {{1,4},{2,3}{6,5}}, the entropy of electric attribute is

$$E(b_i) = -\sum_{a=1}^{n} \frac{|x_a|}{|u|} Log_2 \frac{|x_a|}{|u|}$$

$$\mathrm{E(electric)} = -\left(\frac{|\{1,4\}|}{|\{1,2,3,4,5,6|}Log_2\frac{|\{1,4\}|}{|\{1,2,3,4,5,6|} + \frac{|\{2,3\}|}{|\{1,2,3,4,5,6|}Log_2\frac{|\{2,3\}|}{|\{1,2,3,4,5,6|} + \frac{|\{6,5\}|}{|\{1,2,3,4,5,6|}Log_2\frac{|\{6,5\}|}{|\{1,2,3,4,5,6|}\right)$$

$$= -(\frac{2}{6}Log_2\frac{2}{6} + \frac{2}{6}Log_2\frac{2}{6} + \frac{2}{6}Log_2\frac{2}{6}) = 1.585$$

,and conditional entropy Mechanic w.r.t Electric is

$$cE_{\mathrm{Electric}}(Mechainc\ ) = -\sum_{h=1}^{m}\frac{|y_h|}{|u|}\sum_{a=1}^{n}\frac{|y_h \cap x_a|}{|y_h|}Log_2\frac{|y_h \cap x_a|}{|y_h|}$$

$$= -(\frac{|\{1,2,6\}|}{|\{1,2,3,4,5,6|}\left(\frac{|\{1,2,6\} \cap \{1,4\}|}{|\{1,2,6\}|}Log_2\ \frac{|\{1,2,6\} \cap \{1,4\}|}{|\{1,2,6\}|} + \frac{|\{1,2,6\} \cap \{2,3\}|}{|\{1,2,6\}|}Log_2\ \frac{|\{1,2,6\} \cap \{2,3\}|}{|\{1,2,6\}|} + \frac{|\{1,2,6\} \cap \{6,5\}|}{|\{1,2,6\}|}Log_2\ \frac{|\{1,2,6\} \cap \{6,5\}|}{|\{1,2,6\}|}\right)$$

$$+ \frac{|\{3,4,5\}|}{|\{1,2,3,4,5,6|}\left(\frac{|\{3,4,5\} \cap \{1,4\}|}{|\{3,4,5\}|}Log_2\ \frac{|\{3,4,5\} \cap \{1,4\}|}{|\{3,4,5\}|} + \frac{|\{3,4,5\} \cap \{2,3\}|}{|\{3,4,5\}|}Log_2\ \frac{|\{3,4,5\} \cap \{2,3\}|}{|\{3,4,5\}|} + \frac{|\{3,4,5\} \cap \{6,5\}|}{|\{3,4,5\}|}Log_2\ \frac{|\{3,4,5\} \cap \{6,5\}|}{|\{3,4,5\}|})\right)$$

$$= -\left(\frac{3}{6}\left(\frac{1}{3}\ Log_2\ \left(\frac{1}{3}\right)\frac{1}{3}\ Log_2\ \left(\frac{1}{3}\right) + \frac{1}{3}\ Log_2\ \left(\frac{1}{3}\right)\right) + \frac{3}{6}\left(\frac{1}{3}\ Log_2\ \left(\frac{1}{3}\right) + \frac{1}{3}\ Log_2\ \left(\frac{1}{3}\right) + \frac{0}{3}\ Log_2\ \left(\frac{1}{3}\right)\right)\right) = 1.585$$

by using Formula (19) the information gain of attribute Electric with respect to Mechanic is IG_{Mechanic}(Electric) =0,then the gain ratio of attribute Electric with respect to Mechanic is

$$GR_{Mechainc}(\text{Electric}) = \frac{IG_{b_j}(b_i)}{E(b_i)} = \frac{0}{1.585} = 0$$

the same procedure, the GR of Site management attributes are calculated ,we obtain the mean Gain ratio of attribute Electric is

$$MGR(electric) = \frac{\sum_{\substack{j=1,\\ j\neq i}}^{|B|} GR_{b_j}(b_i)}{|B|-1} = \frac{(0.2897+0)}{3-1} = 0.14485$$

, the same procedure, the MGR of Site management attributes and Mechanic attributes are calculated, as illustrated in the following table 3.6

Table 3.9:Calculation Gain ratio and MGR of all attribute

Attribute with respect to	Gain ratio		MGR
Electric	Mechanic 0.00	Site management 0.2897	0.144845
Mechanic	Electric 0.00	Site management 0.2075	0.103759
Site management	Electric 0.3147	Mechanic 0.1422	0.22844

The clustering attribute with highest MGR is chosen in table 3.6 showing that attribute (Site management) has the highest MGR, the Site management attribute is a clustering attribute.

3.7.4. The Object Splitting:

Divide-conquer method is used to split the objects into clusters.. For example, in Table 3.2 shows the clusters of the maintenance variables based on the clustering attribute chosen by the algorithm i.e. Mechanic, in algorithms MGR (see Appendix C)

Notice that, partitioning the maintenance variables dataset using the Site management attribute as clustering attribute results {1},{6,3},{2,4,5}.

we can split the maintenance variables utilized the hierarchical tree as follows

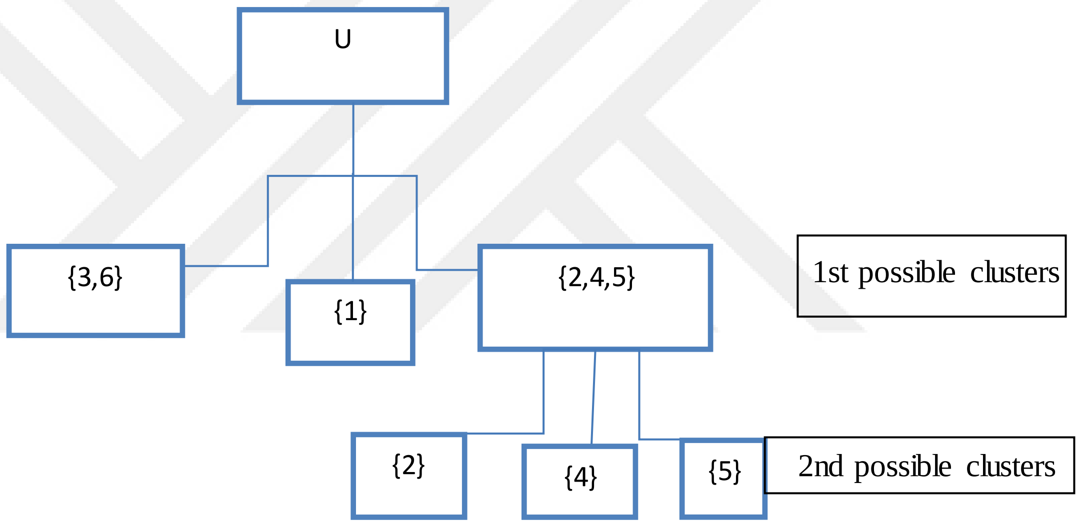


Figure 3.4: clustering results of the MGR algorithm

3.8.Comparison measures:

3.8.1.Overall Purity[41]

The purity of clusters is used as a measure to test the quality of clusters, the purity of a cluster is defined as

$$Purity\ (j) = \frac{bj}{mj}$$

$$Overall\ Purity\ = \frac{\sum_{i=1}^{m} b_j}{m}$$

Where bj is the count of objects in cluster j and its corresponding class, and cluster has the maximum value, furthermore, b_j is the count of objects belongs to a class label that dominates cluster j, where m_j is the count of objects in cluster j, m is the count of object in the dataset, thus, better clustering results are indicated by higher overall purity value, with perfect clustering, a value of 100% yields. High overall purity is easier to achieve when the number of clusters is larger, in particular, if every cluster includes only one object that mean the overall purity is one.

3.8.2.Precision measure[41]

The part of a cluster which content of objects of a specified class. The precision of cluster (β) w.r.t class (α) is

$$\mathrm{Precision}(\alpha, \beta\ ) = Pr_{\alpha\beta}\ = \frac{m_{\alpha\beta}}{m_{\beta}}$$

Where $m_{\alpha\beta}$ is the number of member of class (α) and cluster (β)

m_{β} reprsent the number of member of cluster β

3.8.3.Recall measure [41]

To measure that a cluster consists objects of a specified class. The recall of cluster (β) w.r.t class (α)

$$Recall(\alpha, \beta\ ) = Re_{\alpha\beta}\ = \frac{m_{\alpha\beta}}{m_{\alpha}}$$

where m_{α} represent the number of member of class (α)

3.8.4.F-measure [41]

A combine of precision and recall which measures the extent to that a cluster consist only objects of a particular class and all objects of that class. The F-measure of cluster β w.r.t class (α) is

$$F(\alpha\mathrm{i}, \beta\mathrm{j}) = \frac{2 \times Re_{\alpha\mathrm{i}\beta\mathrm{j}} \times Pr_{\alpha\mathrm{i}\beta\mathrm{j}}}{Re_{\alpha\mathrm{i}\beta\mathrm{j}} + Pr_{\alpha\mathrm{i}\beta\mathrm{j}}},$$

$$F(\alpha, \beta) = \sum_{\alpha i \in \alpha} \frac{|\alpha i|}{N} max_{\beta\mathrm{j} \in \beta}\{F(\alpha\mathrm{i}, \beta\mathrm{j})\} \qquad (3.22)$$

$where\ N\ is\ number\ of\ objects$

3.8.5. Execution Time

The time consumed by each algorithm to process every database is measured and compared. The execution time is an indication of how complex the algorithm is, thus, the less the execution time, the better the algorithm is.

CHAPTER FOUR

Experimental Results

In this chapter, all databases, performance evaluation factors and clustering results using the algorithms VPRS, ITDR, MTMDP and MGR are compared and discussed.

4.1.Benchmark Databases

Four databases are used to compare the results of the algorithms mentioned above. Three real life databases: Soybean, Wisconsin Breast Cancer and Dermatology which are obtained from the UCI Machine Learning Repository [45], and one real life database: Electrical Generator Failures which is collected for a mobile phone company in Iraq.

Soybean. This database is consisted of 47 objects on soybean diseases. These objects are classified in four classes, each class represents a disease, which are, Diaporthe Stem Canker, Charocal Rot, Rhizoctonia Root Rot and Phytophthora Rot. There are 35 categorical attributes describing the objects. The database is classified as 17 objects in the Phytophthora Rot disease and 10 objects in every other disease.

Wisconsin Breast Cancer. This database is consisted of 699 objects, 16 objects have missing values, these objects are neglected, thus, 683 objects are used. These objects are classified into two classes, each class represents a tumor type, that are Benign and Malignant. There are 9 categorical attributes describing the objects. The database is classified as 444 objects in the Benign class and 239 objects in the Malignant class.

Dermatology. This database is consisted of 366 objects, 8 objects have missing values, these objects are neglected, thus, 358 objects are used. These objects as classified into 6 classes, each class represents a skin disease that are Psoriasis, Seboric Dermatisis, Lichen Planus, Pityriasis Rosea, Cronic Dermatitis and Pityriasis Rubra Pilaris. The database is classified as 111 objects in the Psoriasis class, 60 objects in Seboric Dermatisis class, 71 objects in Lichen Planus class, 48 objects in Pityriasis Rosea class, 48 objects in Cronic Dermatitis class and 20 objects in Pityriasis Rubra Pilaris class.

Electrical Generator Failures. This database is consisted of 636 objects classified into 7 classes. Each class represents a failure source that are "Mechanical", "Electrical", "Sites Management", "Mechanical and Electrical", "Mechanical and Sites Management", "Electrical and Site management" and "Mechanical, Electrical and Site Management". The database is classified as 33 objects in "Mechanical" class, 40 objects in "Electrical" class, 22 objects in "Site Management", 150 objects in "Mechanical and Electrical" class, 36 objects in "Mechanical and Site Management" class, 88 objects in "Electrical and Site Management" class, 267 objects in "Mechanical, Electrical and Site Management" class. For detailed description of the database, see Appendix (A).

4.2.Experimental Analysis

In this section, each algorithm is tested against all the databases and the comparison factors are measured and compared for all algorithms. All algorithms are executed in a computer with an Intel Core i7-4500U CPU @ 2.40 GHz and 8.00 GB memory. The databases are managed with MYSQL server and the results are displayed using asp.net web application using C#.

4.2.1.Variable Precision Rough Set.

This algorithm is applied to all databases, the results are shown and discussed below.

Soybean database. The VPRS algorithm is used to divide the objects of the Soybean database into 4 clusters by choosing the attributes with the highest mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.1, table 4.2and table4.3.

Table 4.1: Soybean database clustering purity using VPRS algorithm

Clusters	Objects in cluster	Objects distribution in classes				Purity
		Class1	Class2	Class3	Class4	
Cluster1	28	9	0	8	11	0.39
Cluster2	9	1	0	2	6	0.67
Cluster3	4	0	4	0	0	1
Cluster4	6	0	6	0	0	1
Overall purity						0.76

Table 4.2: Soybean database clustering precision and recall using VPRS algorithm

Class	Cluster1		Cluster2		Cluster3		Cluster4	
	R	P	R	P	R	P	R	P
Class1	0.90	0.32	0	0	0.80	0.29	0.65	0.39
Class2	0.10	0.11	0	0	0.20	0.22	0.35	0.67
Class3	0	0	0.40	1	0	0	0	0
Class4	0	0	0.60	1	0	0	0	0

Table 4.3: Soybean database clustering F-measure using VPRS algorithm.

Class	F-measure distribution for clusters				F
	Cluster1	Cluster2	Cluster3	Cluster4	
Class1	0.47	0.11	0	0	0.47
Class2	0	0	0.57	0.75	0.75
Class3	0.42	0.21	0	0	0.42
Class4	0.49	0.46	0	0	0.49
F-measure					0.53

Dermatology database. The VPRS algorithm is used to divide the objects of the dermatology database into 6 clusters by choosing the attributes with the highest mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.4, table 4.5 and table 4.6.

Table 4.4: Dermatology database clustering purity using VPRS algorithm.

Clusters	Objects in cluster	Objects distribution in classes						Purity
		Class1	Class2	Class3	Class4	Class5	Class6	
Cluster1	213	77	36	1	45	44	10	0.36
Cluster2	2	0	0	2	0	0	0	1
Cluster3	35	0	0	35	0	0	0	1
Cluster4	23	0	0	23	0	0	0	1
Cluster5	24	2	15	0	3	4	0	0.63
Cluster6	61	32	9	10	0	0	10	0.52
Overall purity								0.75

Table 4.5:Dermatology database clustering precision and recall using VPRS algorithm.

Class	Cluster1		Cluster2		Cluster3		Cluster4		Cluster5		Cluster6	
	R	P	R	P	R	P	R	P	R	P	R	P
Class1	0.69	0.36	0.60	0.17	0.01	0	0.94	0.21	0.92	0.21	0.50	0.05
Class2	0	0	0	0	0.03	1	0	0	0	0	0	0
Class3	0	0	0	0	0.49	1	0	0	0	0	0	0
Class4	0	0	0	0	0.32	1	0	0	0	0	0	0
Class5	0.02	0.08	0.25	0.63	0	0	0.06	0.13	0.08	0.17	0	0
Class6	0.29	0.52	0.15	0.15	0.14	0.16	0	0	0	0	0.50	0.16

Table 4.6: Dermatology database clustering F-measure using VPRS algorithm.

Class	F-measure distribution for clusters						F
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	
Class1	0.48	0	0	0	0.03	0.09	0.48
Class2	0.26	0	0	0	0.36	0	0.36
Class3	0.01	0.05	0.66	0.49	0	0	0.66
Class4	0.34	0	0	0	0.08	0	0.34
Class5	0.34	0	0	0	0.11	0	0.34
Class6	0.09	0	0	0	0	0.25	0.25
F-measure							0.44

Breast cancer database. The VPRS algorithm is used to divide the objects of the breast cancer database into 2 clusters by choosing the attributes with the highest mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.7, table 4.8 and table 4.9.

Table 4.7: Breast cancer database clustering purity using VPRS algorithm.

Clusters	Objects in cluster	Objects distribution in classes		Purity
		Class1	Class2	
Cluster1	373	369	4	0.99
Cluster2	310	75	235	0.76
Overall purity				0.87

Table 4.8: Breast cancer database clustering precision and recall using VPRS algorithm.

Class	Cluster1		Cluster2	
	R	P	R	P
Class1	0.83	0.99	0.02	0.01
Class2	0.17	0.24	0.98	0.76

Table 4.9: Breast cancer database clustering F-measure using VPRS algorithm.

Class	F-measure distribution for clusters		F
	Cluster1	Cluster2	
Class1	0.90	0.01	0.90
Class2	0.20	0.86	0.86
F-measure			0.89

Electrical Generators Failure Database. The VPRS algorithm is used to divide the objects of the electrical generators failure database into 7 clusters by choosing the attributes with the highest mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.10, tables 4.11 and tables 4.12.

Table 4.10:Electrical generators failure database clustering purity using VPRS algorithm.

Clusters	Objects in cluster	Objects distribution in classes							Purity
		Class1	Class2	Class3	Class4	Class5	Class6	Class7	
Cluster1	566	30	40	22	122	31	87	234	0.41
Cluster2	44	3	0	0	16	2	0	23	0.52
Cluster3	10	0	0	0	4	3	0	3	0.40
Cluster4	1	0	0	0	1	0	0	0	1.00
Cluster5	11	0	0	0	5	0	1	5	0.45
Cluster6	2	0	0	0	1	0	0	1	0.50
Cluster7	2	0	0	0	1	0	0	1	0.50
Overall purity									0.54

Table 4.11: Electrical generators failure database clustering precision and recall using VPRS algorithm

Class	Cluster1		Cluster2		Cluster3		Cluster4		Cluster5		Cluster6		Cluster7	
	R	P	R	P	R	P	R	P	R	P	R	P	R	P
Class1	0.91	0.05	1	0.07	1	0.04	0.81	0.22	0.86	0.05	0.99	0.15	0.88	0.41
Class2	0.09	0.07	0	0	0	0	0.11	0.36	0.06	0.05	0	0	0.09	0.52
Class3	0	0	0	0	0	0	0.03	0.40	0.08	0.30	0	0	0.01	0.30
Class4	0	0	0	0	0	0	0.01	1	0	0	0	0	0	0
Class5	0	0	0	0	0	0	0.03	0.45	0	0	0.01	0.09	0.02	0.45
Class6	0	0	0	0	0	0	0.01	0.50	0	0	0	0	0	0.50
Class7	0	0	0	0	0	0	0.01	0.50	0	0	0	0	0	0.50

Table 4.12:Electrical generators failure database clustering F-measure using VPRS algorithm.

Class	F-measure distribution for clusters							F
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7	
Class1	0.10	0.08	0	0	0	0	0	0.10
Class2	0.13	0	0	0	0	0	0	0.13
Class3	0.07	0	0	0	0	0	0	0.07
Class4	0.34	0.	0.05	0.01	0.06	0.01	0.01	0.34
Class5	0.10	0.05	0.13	0	0	0	0	0.13
Class6	0.27	0	0	0	0.02	0	0	0.27
Class7	0.56	0.15	0.02	0.01	0.04	0.01	0.01	0.56
F-measure								0.38

4.3.Maximum Total Mean Distribution Precision.

The MTMDP algorithm is applied to all databases, the results are shown and discussed below.

Soybean database. The MTMDP algorithm is used to divide the objects of the Soybean database into 4 clusters by choosing the attributes with the maximum mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.13, table 4.14 and tables 4.15.

Table 4.13: Soybean database clustering purity using MTMDP algorithm.

Clusters	Objects in cluster	Objects distribution in classes				Purity
		Class1	Class2	Class3	Class4	
Cluster1	10	10	0	0	0	1
Cluster2	2	0	0	2	0	1
Cluster3	25	0	0	8	17	0.68
Cluster4	10	0	10	0	0	1
Overall purity						0.92

Table 4.14:Soybean database clustering precision and recall using MTMDP algorithm.

Class	Cluster1		Cluster2		Cluster3		Cluster4	
	R	P	R	P	R	P	R	P
Class1	1	1	0	0	0	0	0	0
Class2	0	0	0	0	0.20	1	0	0
Class3	0	0	0	0	0.80	0.32	1	0.68
Class4	0	0	1	1	0	0	0	0

Table 4.15: Soybean database clustering F-measure using MTMDP algorithm.

Class	F-measure distribution for clusters				F
	Cluster1	Cluster2	Cluster3	Cluster4	
Class1	1	0	0	0	1
Class2	0	0	0	1	1
Class3	0	0.33	0.46	0	0.46
Class4	0	0	0.81	0	0.81
F-measure					0.82

Dermatology database. The MTMDP algorithm is used to divide the objects of the dermatology database into 6 clusters by choosing the attributes with the highest mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.16, table 4.17 and table 4.18.

Table 4.16: Dermatology database clustering purity using MTMDP algorithm.

Clusters	Objects in cluster	Objects distribution in classes						Purity
		Class1	Class2	Class3	Class4	Class5	Class6	
Cluster1	213	77	36	1	45	44	10	0.36
Cluster2	2	0	0	2	0	0	0	1.00
Cluster3	36	0	0	35	0	0	0	1.00
Cluster4	23	0	0	23	0	0	0	1.00
Cluster5	24	2	15	0	3	4	0	0.63
Cluster6	61	32	9	10	0	0	10	0.52
Overall purity								0.75

Table 4.17: Dermatology database clustering precision and recall using MTMDP algorithm.

Class	Cluster1		Cluster2		Cluster3		Cluster4		Cluster5		Cluster6	
	R	P	R	P	R	P	R	P	R	P	R	P
Class1	0.69	0.36	0.60	0.17	0.01	0	0.94	0.21	0.92	0.21	0.50	0.05
Class2	0	0	0	0	0.03	1	0	0	0	0	0	0
Class3	0	0	0	0	0.49	1	0	0	0	0	0	0
Class4	0	0	0	0	0.32	1	0	0	0	0	0	0
Class5	0.02	0.08	0.25	0.63	0	0	0.06	0.13	0.08	0.17	0	0
Class6	0.29	0.52	0.15	0.15	0.14	0.16	0	0	0	0	0.50	0.16

Table 4.18: Dermatology database clustering F-measure using MTMDP algorithm.

	F-measure distribution for clusters						
Class	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	F
Class1	0.48	0	0	0	0.03	0.37	0.48
Class2	0.26	0	0	0	0.36	0.15	0.36
Class3	0.01	0.05	0.66	0.49	0	0.15	0.66
Class4	0.34	0	0	0	0.08	0	0.34
Class5	0.34	0	0	0	0.11	0	0.11
Class6	0.09	0	0	0	0	0.25	0.25
F-measure							0.44

Breast cancer database. The MTMDP algorithm is used to divide the objects of the breast cancer database into 2 clusters by choosing the attributes with the highest mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.19, table 4.20 and table 4.21.

Table 4.19: Breast cancer database clustering purity using MTMDP algorithm.

Clusters	Objects in cluster	Objects distribution in classes		Purity
		Class1	Class2	
Cluster1	373	369	4	
Cluster2	310	75	235	
Overall purity				0.87

Table 4.20: Breast cancer database clustering precision and recall using MTMDP algorithm.

Class	Cluster1		Cluster2	
	R	P	R	P
Class1	0.99	0.90	0.01	0.01
Class2	0.24	0.20	0.76	0.86

Table 4.21: Breast cancer database clustering F-measure using MTMDP algorithm.

Class	F-measure distribution for clusters		F
	Cluster1	Cluster2	
Class1	0.90	0.20	0.90
Class2	0.01	0.86	0.86
F-measure			0.89

Electrical Generators Failure Database. The MTMDP algorithm is used to divide the objects of the electrical generators failure database into 7 clusters by choosing the attributes with the highest mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.22, table 4.23 and table 4.24.

Table 4.22: Electrical generators failure database clustering purity using MTMDP algorithm.

Clusters	Objects in cluster	Objects distribution in classes							Purity
		Class1	Class2	Class3	Class4	Class5	Class6	Class7	
Cluster1	579	30	40	22	130	32	87	30	0.41
Cluster2	6	0	0	0	4	0	0	0	0.67
Cluster3	22	3	0	0	6	1	0	3	0.55
Cluster4	2	0	0	0	1	0	0	0	0.50
Cluster5	12	0	0	0	3	0	1	0	0.67
Cluster6	2	0	0	0	0	0	0	0	1.00
Cluster7	13	30	40	22	130	32	87	30	0.46
Overall purity									0.61

Table 4.23: Electrical generators failure database clustering precision and recall using MTMDP algorithm.

Class	Cluster1		Cluster2		Cluster3		Cluster4		Cluster5		Cluster6		Cluster7	
	R	P	R	P	R	P	R	P	R	P	R	P	R	P
Class1	0.91	0.05	1	0.07	1	0.04	0.87	0.22	0.89	0.06	0.99	0.15	0.89	0.41
Class2	0	0	0	0	0	0	0.03	0.67	0	0	0	0	0.01	0.33
Class3	0.09	0.14	0	0	0	0	0.04	0.27	0.03	0.05	0	0	0.04	0.55
Class4	0	0	0	0	0	0	0.01	0.50	0	0	0	0	0	0.50
Class5	0	0	0	0	0	0	0.02	0.25	0	0	0.01	0.08	0.03	0.67
Class6	0	0	0	0	0	0	0	0	0	0	0	0	0.01	1
Class7	0	0	0	0	0	0	0.04	0.46	0.08	0.23	0	0	0.01	0.31

Table 4.24: Electrical generators failure database clustering F-measure using MTMDP algorithm.

Class	F-measure distribution for clusters							F
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7	
Class1	0.10	0	0.11	0	0	0	0	0.11
Class2	0.13	0	0	0	0	0	0	0.13
Class3	0.07	0	0	0	0	0	0	0.07
Class4	0.36	0.05	0.07	0.01	0.04	0	0.07	0.34
Class5	0.10	0	0.03	0	0	0	0.12	0.12
Class6	0.26	0	0	0	0.02	0	0	0.26
Class7	0.56	0.01	0.08	0	0.06	0.01	0.03	0.56
F-measure								0.38

4.4.Information Theoretic Dependency Roughness.

The ITDR algorithm is applied to all databases, the results are shown and discussed below.

Soybean database. The ITDR algorithm is used to divide the objects of the Soybean database into 4 clusters by choosing the attributes with the maximum mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.25, table 4.26 and table 4.27.

Table 4.25: Soybean database clustering purity using ITDR algorithm.

Clusters	Objects in cluster	Objects distribution in classes				Purity
		Class1	Class2	Class3	Class4	
Cluster1	10	5	5	0	0	0.50
Cluster2	10	5	5	0	0	0.50
Cluster3	2	0	0	2	0	1.00
Cluster4	25	0	0	8	17	0.68
Overall purity						0.67

Table 4.26: Soybean database clustering precision and recall using ITDR algorithm.

Class	Cluster1		Cluster2		Cluster3		Cluster4	
	R	P	R	P	R	P	R	P
Class1	0.50	0.50	0.50	0.50	0	0	0	0
Class2	0.50	0.50	0.50	0.50	0	0	0	0
Class3	0	0	0	0	0.20	1	0	0
Class4	0	0	0	0	0.80	0.32	1	0.68

Table 4.27: Soybean database clustering F-measure using ITDR algorithm.

Class	F-measure distribution for clusters				F
	Cluster1	Cluster2	Cluster3	Cluster4	
Class1	0.50	0.50	0	0	0.50
Class2	0.50	0.50	0	0	0.50
Class3	0	0	0.33	0.46	0.46
Class4	0	0	0	0.81	0.81
F-measure					0.60

Dermatology database. The ITDR algorithm is used to divide the objects of the dermatology database into 6 clusters by choosing the attributes with the highest mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.28, table 4.29 and table 4.30.

Table 4.28: Dermatology database clustering purity using ITDR algorithm.

Clusters	Objects in cluster	Objects distribution in classes						Purity
		Class1	Class2	Class3	Class4	Class5	Class6	
Cluster1	57	50	0	0	0	7	0	0.88
Cluster2	13	3	2	0	6	0	2	0.46
Cluster3	34	1	6	1	18	1	7	0.53
Cluster4	12	0	1	1	8	0	2	0.67
Cluster5	20	17	0	0	1	1	1	0.85
Cluster6	222	40	51	69	15	39	8	0.31
Overall purity								0.62

Table 4.29:Dermatology database clustering precision and recall using ITDR algorithm.

Class	Cluster1		Cluster2		Cluster3		Cluster4		Cluster5		Cluster6	
	R	P	R	P	R	P	R	P	R	P	R	P
Class1	0.45	0.88	0	0	0	0	0	0	0.15	0.12	0	0
Class2	0.03	0.23	0.03	0.15	0	0	0.13	0.46	0	0	0.10	0.15
Class3	0.01	0.03	0.10	0.18	0.01	0.03	0.38	0.53	0.02	0.03	0.35	0.21
Class4	0	0	0.02	0.08	0.01	0.08	0.17	0.67	0	0	0.10	0.17
Class5	0.15	0.85	0	0	0	0.	0.02	0.05	0.02	0.05	0.05	0.05
Class6	0.36	0.18	0.85	0.23	0.97	0.31	0.31	0.07	0.81	0.18	0.40	0.04

Table 4.30: Dermatology database clustering F-measure using ITDR algorithm.

Class	F-measure distribution for clusters						F
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	
Class1	0.60	0.05	0.01	0.00	0.26	0.24	0.60
Class2	0.00	0.05	0.13	0.03	0.00	0.36	0.36
Class3	0.00	0.00	0.02	0.02	0.00	0.47	0.47
Class4	0.00	0.20	0.44	0.27	0.03	0.11	0.44
Class5	0.13	0.00	0.02	0.00	0.03	0.29	0.29
Class6	0.00	0.12	0.26	0.13	0.05	0.07	0.26
F-measure							0.45

Breast cancer database. The ITDR algorithm is used to divide the objects of the breast cancer database into 2 clusters by choosing the attributes with the highest mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.31, table 4.32and table 4.33.

Table 4.31: Breast cancer database clustering purity using ITDR algorithm.

Clusters	Objects in cluster	Objects distribution in classes		Purity
		Class1	Class2	
Cluster1	432	391	41	0.91
Cluster2	251	53	198	0.79
Overall purity				0.85

Table 4.32:Breast cancer database clustering precision and recall using ITDR algorithm.

Class	Cluster1		Cluster2	
	R	P	R	P
Class1	0.88	0.91	0.17	0.09
Class2	0.12	0.21	0.83	0.79

Table 4.33: Breast cancer database clustering F-measure using ITDR algorithm.

Class	F-measure distribution for clusters		F
	Cluster1	Cluster2	
Class1	0.89	0.15	0.89
Class2	0.12	0.81	0.81
F-measure			0.86

Electrical Generators Failure Database. The ITDR algorithm is used to divide the objects of the electrical generators failure database into 7 clusters by choosing the attributes with the highest mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.34, table 4.35 and table 4.36.

Table 4.34: Electrical generators failure database clustering purity using ITDR algorithm.

Clusters	Objects in cluster	Objects distribution in classes							Purity
		Class1	Class2	Class3	Class4	Class5	Class6	Class7	
Cluster1	536	26	40	22	114	31	88	215	0.40
Cluster2	21	2	0	0	5	1	0	13	0.62
Cluster3	29	1	0	0	12	2	0	14	0.48
Cluster4	3	1	0	0	1	0	0	1	0.33
Cluster5	40	2	0	0	15	2	0	21	0.53
Cluster6	1	0	0	0	1	0	0	0	1.00
Cluster7	6	1	0	0	2	0	0	3	0.50
Overall purity									0.55

Table 4.35:Electrical generators failure database clustering precision and recall using ITDR algorithm.

Class	Cluster1		Cluster2		Cluster3		Cluster4		Cluster5		Cluster6		Cluster7	
	R	P	R	P	R	P	R	P	R	P	R	P	R	P
Class1	0.79	0.05	1.00	0.07	1.00	0.04	0.76	0.21	0.86	0.06	1.00	0.16	0.81	0.40
Class2	0.06	0.10	0.00	0.00	0.00	0.00	0.03	0.24	0.03	0.05	0.00	0.00	0.05	0.62
Class3	0.03	0.03	0.00	0.00	0.00	0.00	0.08	0.41	0.06	0.07	0.00	0.00	0.05	0.48
Class4	0.03	0.33	0.00	0.00	0.00	0.00	0.01	0.33	0.00	0.00	0.00	0.00	0.00	0.33
Class5	0.06	0.05	0.00	0.00	0.00	0.00	0.10	0.38	0.06	0.05	0.00	0.00	0.08	0.53
Class6	0.00	0.00	0.00	0.00	0.00	0.00	0.01	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Class7	0.03	0.17	0.00	0.00	0.00	0.00	0.01	0.33	0.00	0.00	0.00	0.00	0.01	0.50

Table 4.36: Electrical generators failure database clustering F-measure using ITDR algorithm.

Class	F-measure distribution for clusters							F
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7	
Class1	0.09	0.07	0.03	0.06	0.05	0.00	0.05	0.09
Class2	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.14
Class3	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.08
Class4	0.33	0.06	0.13	0.01	0.16	0.01	0.03	0.33
Class5	0.11	0.04	0.06	0.00	0.05	0.00	0.00	0.11
Class6	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.28
Class7	0.54	0.09	0.09	0.01	0.14	0.00	0.02	0.54
F-measure								0.36

4.5.Mean Gain Ratio.

The MGR algorithm is applied to all databases, the results are shown and discussed below.

Soybean database. The MGR algorithm is used to divide the objects of the Soybean database into 4 clusters by choosing the attributes with the maximum mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.37, table 4.38 and table 4.39.

Table 4.37: Soybean database clustering purity using MGR algorithm.

Clusters	Objects in cluster	Objects distribution in classes				Purity
		Class1	Class2	Class3	Class4	
Cluster1	10	10	0	0	0	1.00
Cluster2	2	0	0	2	0	1.00
Cluster3	25	0	0	8	17	0.68
Cluster4	10	0	10	0	0	1.00
Overall purity						0.92

Table 4.38:Soybean database clustering precision and recall using MGR algorithm.

Class	Cluster1		Cluster2		Cluster3		Cluster4	
	R	P	R	P	R	P	R	P
Class1	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Class2	0.00	0.00	0.00	0.00	0.20	1.00	0.00	0.00
Class3	0.00	0.00	0.00	0.00	0.80	0.32	1.00	0.68
Class4	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00

Table 4.39: Soybean database clustering F-measure using MGR algorithm

Class	F-measure distribution for clusters				F
	Cluster1	Cluster2	Cluster3	Cluster4	
Class1	1.00	0.00	0.00	0.00	1.00
Class2	0.00	0.00	0.00	1.00	1.00
Class3	0.00	0.33	0.46	0.00	0.46
Class4	0.00	0.00	0.81	0.00	0.81
F-measure					0.82

Dermatology database. The MGR algorithm is used to divide the objects of the dermatology database into 6 clusters by choosing the attributes with the highest mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.40, table 4.41 and table 4.42.

Table 4.40: Dermatology database clustering purity using MGR algorithm.

Clusters	Objects in cluster	Objects distribution in classes						Purity
		Class1	Class2	Class3	Class4	Class5	Class6	
Cluster1	287	110	60	1	48	48	20	0.38
Cluster2	2	1	0	1	0	0	0	0.50
Cluster3	1	0	0	1	0	0	0	1.00
Cluster4	4	0	0	4	0	0	0	1.00
Cluster5	32	0	0	32	0	0	0	1.00
Cluster6	32	0	0	32	0	0	0	1.00
Overall purity								0.81

Table 4.41:Dermatology database clustering precision and recall using MGR algorithm.

Class	Cluster1		Cluster2		Cluster3		Cluster4		Cluster5		Cluster6	
	R	P	R	P	R	P	R	P	R	P	R	P
Class1	0.99	0.38	1.00	0.21	0.01	0.00	1.00	0.17	1.00	0.17	1.00	0.07
Class2	0.01	0.50	0.00	0.00	0.01	0.50	0.00	0.00	0.00	0.00	0.00	0.00
Class3	0.00	0.00	0.00	0.00	0.01	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Class4	0.00	0.00	0.00	0.00	0.06	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Class5	0.00	0.00	0.00	0.00	0.45	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Class6	0.00	0.00	0.00	0.00	0.45	1.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.42: Dermatology database clustering F-measure using MGR algorithm.

Class	F-measure distribution for clusters						F
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	
Class1	0.55	0.02	0.00	0.00	0.00	0.00	0.55
Class2	0.35	0.00	0.00	0.00	0.00	0.00	0.35
Class3	0.01	0.03	0.03	0.11	0.62	0.62	0.62
Class4	0.29	0.00	0.00	0.00	0.00	0.00	0.29
Class5	0.29	0.00	0.00	0.00	0.00	0.00	0.29
Class6	0.13	0.00	0.00	0.00	0.00	0.00	0.13
F-measure							0.44

Breast cancer database. The MGR algorithm is used to divide the objects of the breast cancer database into 2 clusters by choosing the attributes with the highest mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.43, table4.44 and table 4.45.

Table 4.43: Breast cancer database clustering purity using MGR algorithm.

Clusters	Objects in cluster	Objects distribution in classes		Purity
		Class1	Class2	
Cluster1	373	369	4	0.99
Cluster2	310	75	235	0.76
Overall purity				0.87

Table 4.44:Breast cancer database clustering precision and recall using MGR algorithm.

Class	Cluster1		Cluster2	
	R	P	R	P
Class1	0.83	0.99	0.02	0.01
Class2	0.17	0.24	0.98	0.76

Table 4.45: Breast cancer database clustering F-measure using MGR algorithm.

Class	F-measure distribution for clusters		F
	Cluster1	Cluster2	
Class1	0.90	0.20	0.90
Class2	0.01	0.86	0.86
F-measure			0.89

Electrical Generators Failure Database. The MGR algorithm is used to divide the objects of the electrical generators failure database into 7 clusters by choosing the attributes with the highest mean as clustering attributes. The overall purity and F-measure calculations are shown in tables 4.46, table 4.47and table4.48.

Table 4.46 :Electrical generators failure database clustering purity using MGR algorithm.

Clusters	Objects in cluster	Objects distribution in classes							Purity
		Class1	Class2	Class3	Class4	Class5	Class6	Class7	
Cluster1	618	33	40	22	147	35	86	255	0.41
Cluster2	2	0	0	0	0	0	1	1	0.50
Cluster3	5	0	0	0	2	0	0	3	0.60
Cluster4	7	0	0	0	0	1	1	5	0.71
Cluster5	1	0	0	0	0	0	0	1	1.00
Cluster6	2	0	0	0	0	0	0	2	1.00
Cluster7	1	0	0	0	1	0	0	0	1.00
Overall purity									0.75

Table 4.47: Electrical generators failure database clustering precision and recall using MGR algorithm.

Class	Cluster1		Cluster2		Cluster3		Cluster4		Cluster5		Cluster6		Cluster7	
	R	P	R	P	R	P	R	P	R	P	R	P	R	P
Class1	1.00	0.05	1.00	0.06	1.00	0.04	0.98	0.24	0.97	0.06	0.98	0.14	0.96	0.41
Class2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.50	0.00	0.50
Class3	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.40	0.00	0.00	0.00	0.00	0.01	0.60
Class4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.14	0.01	0.14	0.02	0.71
Class5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Class6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	1.00
Class7	0.00	0.00	0.00	0.00	0.00	0.00	0.01	1.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.48: **Electrical generators failure database clustering F-measure using MGR algorithm.**

Class	F-measure distribution for clusters							F
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7	
Class1	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.10
Class2	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.12
Class3	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.07
Class4	0.38	0.00	0.03	0.00	0.00	0.00	0.01	0.38
Class5	0.11	0.00	0.00	0.05	0.00	0.00	0.00	0.11
Class6	0.24	0.02	0.00	0.02	0.00	0.00	0.00	0.24
Class7	0.58	0.01	0.02	0.04	0.01	0.01	0.00	0.58
F-measure								0.39

4.6.Performance measures summary.

In order to illustrate the performance of each algorithm, the performance measures are summarized and the average for the four databases is calculated for each algorithm.

4.6.1.Overall purity averages.

The overall purities for the four algorithms on four databases are represented in Figure 4.1.

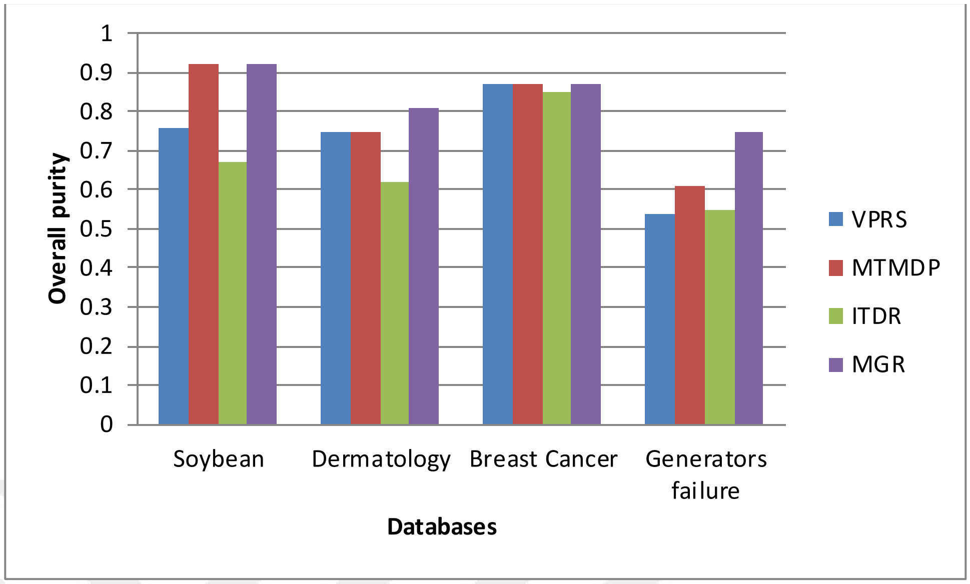


Figure 4.1: **Clustering overall purity of four algorithms on four databases.**

The overall purity average for each algorithm is calculated in table 4.49.

Table 4.49:Overall purity of four algorithms on four databases

Algorithms	Overall purity				Average
	Soybean	Dermatology	Breast Cancer	Generators Failure	
VPRS	0.76	0.75	0.87	0.54	0.73
MTMDP	0.92	0.75	0.87	0.61	0.79
ITDR	0.67	0.62	0.85	0.55	0.67
MGR	0.92	0.81	0.87	0.75	0.84

4.6.2.F-measure Averages.

The F-measure for the four algorithms on the four databases are represented in figure 4.2.

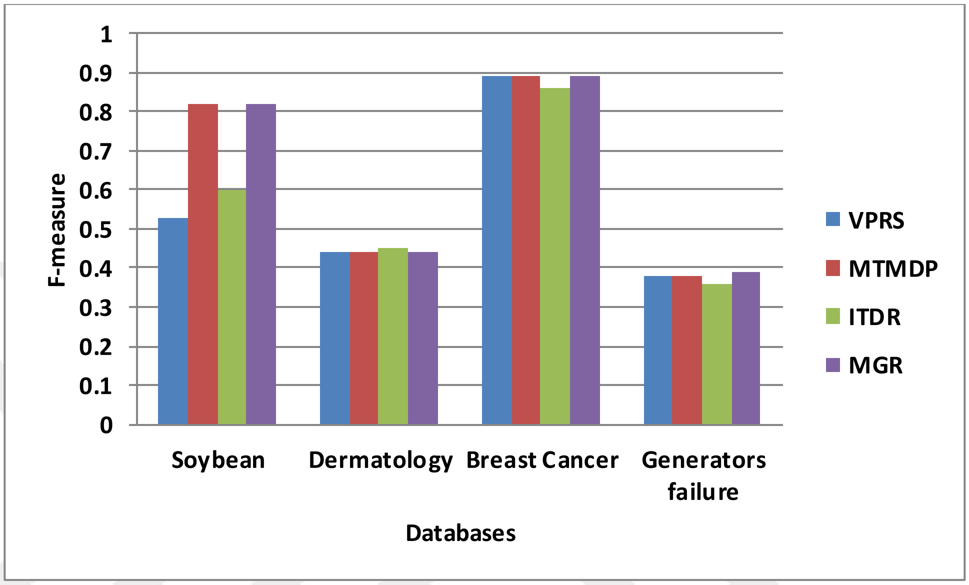


Figure 4.2: Clustering F-measure of four algorithms on four databases.

The F-measure average for each algorithm is calculated in table 4.50.

Table 4.50: F-measure of four algorithms on four databases.

Algorithms	F-measure				Average
	Soybean	Dermatology	Breast Cancer	Generators Failure	
VPRS	0.53	0.44	0.89	0.38	0.56
MTMDP	0.82	0.44	0.89	0.38	0.63
ITDR	0.6	0.45	0.86	0.36	0.57
MGR	0.82	0.44	0.89	0.39	0.64

4.6.3.Execution time.

The time consumed by each algorithm to calculate the results of each database is represented in figure 4.3.

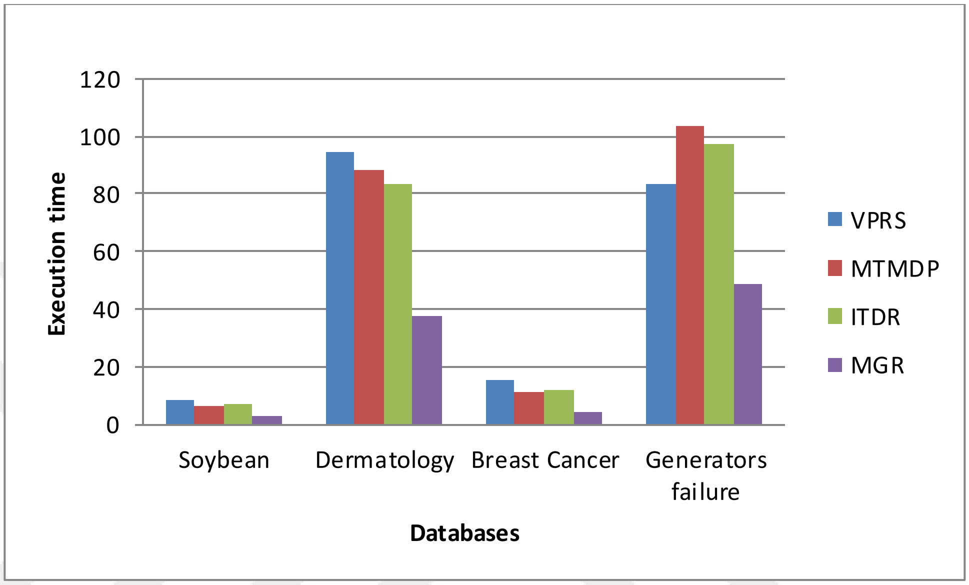


Figure 4.3: Execution time of four algorithms on four databases.

The average execution time for each algorithm on four databases are calculated in table 4.51.

Table 4.51: Execution time of four algorithms on four databases.

Algorithms	Execution time in seconds				Average
	Soybean	Dermatology	Breast Cancer	Generators Failure	
VPRS	8.16	94.56	15.39	83.52	50.41
MTMDP	6.37	88.47	11.15	103.88	52.47
ITDR	6.68	83.51	11.76	97.44	49.85
MGR	2.67	37.62	3.99	48.81	23.27

4.7.Performance Measures.

The performance measures of all the four algorithms applied to the four databases are discussed in this section.

4.7.1.Overall Purity.

The VPRS, MTMDP, ITRD and MGR algorithms are applied to the four databases Soybean, Dermatology, Breast cancer and Electrical generators failures. The overall purities for the resulting clusters are summarized in table 4.49. The MGR and MTMDP algorithms have the highest overall purity (0.92) when applied to the Soybean database, the MGR algorithm has the highest overall purity (0.81) when applied to the Dermatology database. VPRS, MTMDP and MGR have the highest overall purity (0.87) when applied to the Breast cancer database. For the Electrical generators failure database, the MGR has the highest overall purity (0.75). Overall purity average is calculated and presented on the same table 4.49, the MGR the best overall purity average (0.84).

4.7.2.F-Measure.

The F-measures for the Soybean, Dermatology, Breast cancer and Electrical generators failure databases when clustered using VPRS, MTMDP, ITDR and MGR algorithms are summarized in table 4.50. The highest F-measure for the Soybean database clusters (0.82) is achieved by the MTMTP and MGR algorithms. The highest F-measure for the Dermatology database clusters (0.45) is achieved by the ITDR algorithm. The VPRS, MTMDP and MGR achieved the highest F-measure (0.89) when clustering the Breast cancer database. The highest F-measure (0.39) is achieved by the MGR algorithm for the Electrical generators failure database clustering. F-Measure average is calculated and presented on the same table 4.50, the MGR the best F-Measure average (0.64).

4.7.3. Execution Time.

The execution time for each algorithm is measured for the four databases as an indication of algorithm simplicity. The algorithm with the lowest time consumption is considered to be more simple, thus, more efficient. The MGR algorithm has the least time consumption when compared to the VPRS, MTMDP and ITDR algorithms for all the databases used. The average execution time for the MGR algorithm when applied to the Soybean, Dermatology, Breast cancer and Electrical generators failure is 23.27 sec.

4.8. Results Analysis.

As the MGR algorithm has the highest overall purity average, highest F-measure average and least execution time average compared to the VPRS, MTMDP and ITDR algorithms when used to cluster the four databases Soybean, Dermatology, Breast cancer and Electrical generators failure; and as the MGR has the best overall purity, F-measure and execution time compared to the VPRS, MTMDP and ITDR algorithms when applied to the electrical generators failure database; the MGR algorithm results are used to find the highest of mean attribute in the electrical generators failure database in order to help the decision makers to make appropriate modification in the maintenance team schedules and operations to improve the maintenance performance.

As presented in Appendix (B), the attribute with the highest mean (0.032) is the "Replacing air filter" (RAF) attribute while the second highest mean (0.029) is for the attribute "No fuel" (NF) attribute and the attribute "Owner problem" (OP) has the third highest mean (0.024), thus, these attributes are suggested to the decision makers as The highest of the mean score is the most potential attributes of Electrical generators failures in order to take proper decision to increase the efficiency of the maintenance team performance.

CHAPTER FIVE

PROPOSED ALGORITHM

5.1.MIGR algorithm

In this chapter, we propose a new algorithm for categorical data clustering called MIGR (minimum information gain roughness). We start with new concepts such as IG (Information Gain) and MIG (minimum Information Gain), followed by the pseudo-code description of MIGR algorithm in section(5.2) . This technique is applied to three real life sample datasets [47] and the selected clustering attributes are used to cluster the datasets using the Divide-Conquer method. The quality of the resulting clusters are measured based on Clustering accuracy and F-Measure compared to the quality of the clusters resulted using MMR and ITDR techniques[30],[15] The our contributions show the significance of clustering categorical data using a clustering attribute, Propose a novel Rough Set Theory based technique (MIGR) for selecting the clustering attributes and increasing rate accuracy in a selecting attribute . The calculate of MIGR by using of some definition as follows:

5.1.1.Definition [48,55] (the entropy of Shannon) let an information system $S = (u, B, v, f)$, Q subset of B, u /Q ={X1, X2, . . ., Xn}.

$$H(Q) = -\sum_{i=1}^{n} p(X_i)\log p(X_i) = -\sum_{i=1}^{n} \frac{|X_i|}{|u|} log \frac{|X_i|}{|u|} \qquad (5.1)$$

Is the definition of the entropy of Shannon H(Q) of Q, such that $p(x) = \frac{|X_i|}{|u|}$

5.1.2.Definition [48,55](the joint entropy) let an information system $S = (u, B, v, f)$, Q and P subset of B, U/Q ={X1, X2, . . ., Xn}and U/P ={Y1, Y2, . . ., Ym}.The definition of the joint entropy of Q and P is:

$$H(Q,P) = -\sum_{i=1}^{n} p(X_i, Y_j)\log p(X_i, Y_j) = -\sum_{i=1}^{n} \frac{|X_i \cap Y_j|}{|u|} log \frac{|X_i \cap Y_j|}{|u|} \tag{5.2}$$

Such that $p(X,Y)=\frac{|X_i \cap Y_i|}{|u|}$

5.1.3.Definition [55,60] , [36](information gain)let $S = (u, B, v, f)$be an information system, B is the attribute set ,Q and P are two subsets of B, such that U/Q ={X1, X2, . . ., Xn} and U/P ={Y1, Y2, . . ., Ym} ,the information gain of Q w.r.t P is defined by:

$$\mathrm{IG_Q}(P) = H(P) + H(Q) - H(p,Q)$$

$$= \begin{cases} -\sum_{i=1}^{n} \frac{|X_i|}{|U|} log \frac{|X_i|}{|U|} + \left(-\frac{|Y_j|}{|U|} log \frac{|Y_j|}{|U|}\right) - \left(-\frac{|X_i \cap Y_j|}{|U|} log \frac{|X_i \cap Y_j|}{|U|}\right) & , |X_i \cap Y_j| > 0 \\ 0 & , |X_i \cap Y_j| = 0 \end{cases} \tag{5.3}$$

Where the information gain= 0 if X and Y are independent .

5.1.4.Definition let Q_i belong to A, $v(Q_i)$ has s- various values, i.e. σ_s , $s = 1,2, \dots \dots, n$. Let $y(\mathrm{Qi} = x_s), s = 1,2, \dots, n$ be a subset of the objects having s- various values of attribute Q_i. the roughness of the set $y(\mathrm{Qi} = x_s), s = 1,2, \dots, n$ w.r.t P_j, *(i) is not equal to (j)* , denoted by $MIG_{Pj}(Qi = x_s)$ is :

$$MIG_{Pj}(Qi = x_s) = \min(IG_{Pj}(X, y(\mathrm{Qi} = x_s)) \tag{5.4}$$

the mean roughness of attribute of Q_i with respect to P_j, Q_i and P_j belong to B, such that *(i) is not equal to (j)* denoted by $MMIG(Q_i)$ gives :

$$\mathrm{MMIG(Qi)} = MIG_{Pj}(Qi) + \dots + MIG_{pj}(Qi_{|v(\mathrm{Qi})|})/|v(\mathrm{Qi})| \tag{5.5}$$

$|v(\mathrm{Qi})|$ *is represented values of attribute* Qi belongs to B

5.1.5.Definition **.**Suppose m attributes, min-mean –min –roughness of attribute Q_i belongs to y, w.r.t P_j belongs to x, such that *(i) is not equal to (j)* indicate to min of $MMRIG(Q_i, P_j)$, denoted $MMMIG(y_i, x_j)$ is calculated using the equation

$$\mathrm{MMMIG}(y_i, x_j) = \min(\mathrm{MMIG}_{Pi}(\mathrm{Qi}), \dots \dots, \mathrm{MMIG}_{Pi}(\mathrm{Qi})) \qquad (5.6)$$

5.2. The pseudo-code of MIGR algorithm

Algorithm:MIGR

Input: Data set

Output: Clustering attribute

1: calculation the equivalence classes utilized the indiscernibility relation on each attribute.

2: calculate the information gain of attribute P_i w.r.t all Q_j, where i is not equal to j

3: calculate the Min roughness $MIG(Q_i)$

4: calculate the Min mean roughness ($MMIG(Q_i)$) of attribute of P_i w.r.t all Q_j

5: choose a clustering attribute depended on the Min $MMIG(Q_i)$of attribute.

End

5.3.Example

The table 5.1 is an information system of six objects u=6 with six categorical valued attributes ,such that hair, teeth, eye, feet, milk and fly, attribute teeth has only three values, while the attributes hair, milk , eye , feet and fly have two values .

Table 5.1. An information system of Animal world dataset.

Rows	Hair	Teeth	Eye	Feet	Milk	Fly
1	Y	Blunt	Forward	Claw	Y	N
2	Y	N	Side	Claw	Y	N
3	Y	N	Side	Claw	Y	N
4	N	Pointed	Side	Claw	N	N
5	N	Pointed	Forward	Hoof	N	N
6	N	Blunt	Forward	Claw	N	Y

There doesn't exist any a pre-defined a clustering (decision) attribute. Thus , from all candidates we will chose a clustering attribute. To get the values of MIGR, first step, we must get the equivalence classes induced by indiscernibility relation of singleton attribute. The six partitions of object from Table1 are shown as follows:

1-X(Hair, Y)={1,2,3},X(Hair, N)={4,5,6},

u /Hair={{1,2,3},{4,5,6}}

2-X (Teeth, Blunt)={1,6}, X(Teeth, N)={2,3}, X(Teeth, Pointed)={4,5},

u/ Teeth= {{1,6} ,{2,3} ,{4,5}}

3-X (Eye, Forward)={1,5,6},X(Eye, Side)={2,4,3},

u / Eye={{1,5,6},{2,4,3}}

4-X (Feet, Hoof)={ 5},X(Feet, Claw)={1,2,4,6,3},

u/ Feet ={{1,2,4,6,3},{5}}

5- X(milk, Y)={1,2,3},X(milk, N)={4,5,6},

u/ milk={{1,2,3},{4,5,6}}

6- X(Fly, Y)={6},X(Fly, N)={1,2,3,4,5},

u /Fly ={{1,2,3,4,5},{6}}

The mean roughness on each attribute is calculated, the mean roughness on Hair w.r.t Eye is calculated ,there are two elementary sets for y_1 (Eye ,forward) = {1,5,6},y_2(Eye, side)={2,3,4},there are two elementary x_1(Hair,Y) ={1,2,3},x_2(Hair, N)={45,6},according to definition of entropy of $x_1$is

$$H(x_1) = \frac{-|X_1|}{|u|} log \frac{|X_1|}{|u|} = -\frac{|\{1,2,3\}|}{|6|} log \frac{|\{1,2,3\}|}{|6|} = 0.5$$

, and entropy of y_1 is

$$H(y_1) = \frac{-|y_1|}{|u|} log \frac{|y_1|}{|u|} = -\frac{|\{1,5,6\}|}{|6|} log \frac{|\{1,5,6\}|}{|6|} = 0.5$$

The joint entropy of x_1 and y_1 is

$$H(x_1, y_1) = -\frac{|X_i \cap Y_i|}{|u|} log \frac{|X_i \cap Y_i|}{|u|} = -\frac{|\{1,2,3\} \cap \{1,5,6\}|}{|6|} log \frac{|\{1,2,3\} \cap \{1,5,6\}|}{|6|}$$
$$= 0.43083$$

The information gain measures x_1 and y_1 is

$$\mathrm{IG}_1 = H(x_1) + H(y_1) - H(x_1, y_1) = 0.5 + 0.5 - 0.43083 = 0.56917$$

,according to definition of entropy of x_2

$$H(x_2) = \frac{-|X_2|}{|u|} log \frac{|X_2|}{|u|} = -\frac{|\{4,5,6\}\ |}{|6|} log \frac{|\{4,5,6\}|}{|6|} = 0.5$$

.The joint entropy of x_2 and y_1 is

$$H(x_2, y_1) = (x_2, y_1) = -\frac{|X_i \cap Y_i|}{|u|} log \frac{|X_i \cap Y_i|}{|u|}$$
$$= -\frac{|\{4,5,6\} \cap \{1,5,6\}|}{|6|} log \frac{|\{4,5,6\} \cap \{1,5,6\}|}{|6|} = 0.5283$$

The information gain measures x_1 and y_2 is

$$\mathrm{IG}_1 = H(x_2) + H(y_1) - H(x_2, y_1) = 0.5 + 0.5 - 0.5283 = 0.4717.$$

The min information gain on y_1 (Eye, Forward) ={1,5,6} with respect to Hair is IG = 0.4717. According to definition of entropy of x_2 is $H(x_2) = 0.5$ .And entropy of y_2 is

$$H(y_2) = \frac{-|y_2|}{|u|} log \frac{|y_2|}{|u|} = -\frac{|\{2,3,4\}|}{|6|} log \frac{|\{2,4,3\}|}{|6|} = 0.5$$

The joint entropy of x_2 and y_2 is

$$H(x_2, y_2) = -\frac{|X_i \cap Y_i|}{|u|} log \frac{|X_i \cap Y_i|}{|u|} = -\frac{|\{2,4,3\} \cap \{4,5,6\}|}{|6|} log \frac{|\{2,4,3\} \cap \{4,5,6\}|}{|6|} = 0.4308$$

The information gain measures x_2 and y_2 is

$$\mathrm{IG}_2 = H(x_2) + H(y_2) - H(x_2, y_2) = 0.5 + 0.5 - 0.4308 = 0.5691$$

.Where x_1(Hair,N) ={1,2,3}, y_2 (Eye,side) ={2,4,3},according to definition of entropy of x_1 is

$$H(x_1) = \frac{-|X_1|}{|u|} log \frac{|X_1|}{|u|} = -\frac{|\{1,2,3\}\ |}{|6|} log \frac{|\{1,2,3\}\ |}{|6|} = 0.5$$

and the entropy of y_2

$$H(y_2) = \frac{-|y_2|}{|u|} log \frac{|y_2|}{|u|} = -\frac{|\{2,3,4\}|}{|6|} log \frac{|\{2,4,3\}|}{|6|} = 0.5$$

The joint entropy of x_1 and y_2 is

$$H(x_1, y_2) = -\frac{|X_i \cap Y_i|}{|u|} log \frac{|X_i \cap Y_i|}{|u|} = -\frac{|\{2,4,3\} \cap \{1,2,3\}|}{|6|} log \frac{|\{2,4,3\} \cap \{1,2,3\}|}{|6|} = 0.52832$$

The information gain measures x_1 and y_2 is

$$\mathrm{IG}_2 = H(x_1) + H(y_2) - H(x_1, y_2) = 0.5 + 0.5 - 0.5283 = 0.4717$$

Table 5.2 .Mean roughness calculation for attribute {hair}.

w.r.t/x	Hair =Y	Hair =N
Eye=Forward	0.56917	0.4717
Eye= side	0.4717	0.56917

the min information gain on y_2 (Eye, side) ={2,3,4} with respect to Hair is IG_2 = 0.4717,The mean Hair with respect to Eye is 0.4717.Following the same procedure, the mean on all attributes with respect each to the other are computed. These calculations are summarized in Table5. 3.With MIGR technique, From Table5.3, the lower of mean of attributes is attribute Fly. Thus, attribute Fly is selected as a clustering attribute.

Table 5.3.(MIGR calculation)

Attribute (w.r.t)	Hair	Teeth	Eye	Feet	Milk	Fly	Mean
Hair	-	0	0.4717	0.0954	0	0.0954	0.1325107
Teeth	0.1992	-	0.1992	0.1056	0.1992	0.1056	0.161724
Eye	0.4717	0	-	0.0954	0.4717	0.0954	0.2268466
Feet	0.0954	0.1096	0.0954	-	0.0954	0.0242	0.08402345
Milk	0	0	0.4717	0.0954	-	0.0954	0.13251707
Fly	0.0954	0.1096	0.0954	0.0242	0.0954	-	0.08402345

For objects splitting, we use a divide-conquer method. We can cluster (partition) the objects based on the decision attribute selected, i.e., Fly . Notice that, the partition of the set of objects induced by attribute Fly is u /Fly ={{1,2,3,4,5},{6}}.To this, we can split the objects into two cluster as the first cluster {1,2,3,4,5}and second cluster {6}.

5.4. Benchmark datasets

Three real-life datasets were chosen to be experimented: **Soybean, Zoo** and **Breast Cancer**, which are obtained from the UCI Machine Learning Repository [47]. A brief description for each dataset is provided next.

Soybean. This dataset contains data about diseases in soybeans; it is consisted of 47 objects described using 35categorical attributes. Objects are classified into four classes according to the diseases found in the plant, which are Diaporthe Stem Canker, Phytophthora Rot, Rhizoctonia Root Rot and Charocal Rot. Objects are distributes as 10 for all classes except for the Phytophthora Rot which contains 17 objects.

Wisconsin Breast Cancer. This database is consisted of 699 objects, 16 objects have missing values. These objects are classified into two classes, each class represents a tumor type, that are Benign and Malignant. There are 9 categorical attributes describing the objects. The database is classified as 458 objects in the Benign class and 241 objects in the Malignant class.

The Zoo dataset. This database is consisted of 101 objects. . These objects are classified into seven classes, every object represents information of an animal by 17 categorical attributes. Each animal data point is classified into seven classes . Hence, The splitting data for MIGR is set at seven clusters.

5.5. Performance measure

In order to identify the technique with the better results, a performance measure must be set to measure the quality of the resulting clusters, thus, Clustering accuracy is measured for each dataset when applied to each technique. A Clustering accuracy is calculated using the following equation:

$\mathrm{Accuracy(i)} = \sum_{i}^{k} \frac{a_i}{n_i}$, where (a_i) is the maximum number of objects shared between this cluster any of the classes and (n_i) is the number of objects in the data set.

where (i) is the resulting clusters count. According to the equations above, the higher the clustering accuracy the better the clustering results and when objects in each cluster fall in one class, this results a 100% clustering accuracy. In general, the higher the number of resulting clusters the easier to achieve higher accuracy.

5.6. Comparison with other two algorithms

5.6.1. Accuracy

The MMR, ITDR, MTMDP, VPRS, MGR and MIGR algorithms are applied to the three datasets Soybean, Breast cancer and Zoo . The clustering accuracies for the resulting clusters are summarized in table 5.13. The MIGR algorithm has the highest average clustering accuracy (0.86) when applied to the Soybean, Breast cancer and Zoo, While MMR algorithm has average clustering accuracy (0.84) and ITDR and VPRS algorithms have average clustering accuracy (0.78). While MTMDP and MGR algorithm have average clustering accuracy (0.84).The average clustering accuracy is calculated and presented on the same table 5.13, the MIGR the highest average clustering accuracy (0.86). In summary, the MIGR algorithm has 2% higher average accuracy when compared to the MMR algorithm and 3% when compared to the MTMDP and ITDR algorithms and 8% when compared to the ITDR and VPRS algorithms.

Table 5. 4.Clustering results for Breast cancer dataset using MIGR algorithm.

Clusters	Objects in cluster	Distribution in classes		Purity
		Class1	Class2	
Cluster1	384	380	4	0.99
Cluster2	315	78	237	0.75
Accuracy				0.88

Table 5. 5.Clustering results for Breast cancer dataset using MMR algorithm.

Clusters	Objects in cluster	Distribution in classes		Purity
		Class1	Class2	
Cluster1	579	445	143	0.77
Cluster2	120	13	107	0.89
Accuracy				0.79

Table 5.6.Clustering results for Breast cancer dataset using ITDR algorithm.

Clusters	Objects in cluster	Distribution in classes		Purity
		Class1	Class2	
Cluster1	443	402	41	0.91
Cluster2	256	56	200	0.78
Accuracy				0.86

Table 5.7. Clustering results for Zoo dataset using MMR algorithm.

Clusters	Objects in cluster	Distribution in classes							Purity
		Class1	Class2	Class3	Class4	Class5	Class6	Class7	
Cluster1	6	0	0	3	0	3	0	0	0.5
Cluster2	39	39	0	0	0	0	0	0	1.00
Cluster3	2	0	0	1	0	1	0	0	0.5
Cluster4	14	0	0	1	13	0	0	0	0.93
Cluster5	12	0	0	0	0	0	2	10	0.83
Cluster6	8	2	0	0	0	0	6	0	0.75
Cluster7	20	0	20	0	0	0	0	0	1.00
Accuracy									0.91

Table 5. 8. Clustering results for Zoo dataset using ITDR algorithm.

Clusters	Objects in cluster	Distribution in classes							Purity
		Class1	Class2	Class3	Class4	Class5	Class6	Class7	
Cluster1	14	0	0	1	13	0	0	0	0.92
Cluster2	9	0	0	3	0	6	0	0	0.66
Cluster3	15	0	0	1	0	0	4	10	0.66
Cluster4	20	0	20	0	0	0	0	0	1.00
Cluster5	10	6	0	0	0	0	4	0	0.6
Cluster6	17	17	0	0	0	0	0	0	1.00
Cluster7	16	16	0	0	0	0	0	0	1.00
Accuracy									0.87

Table 5.9. Clustering results for Zoo dataset using MIGR algorithm.

Clusters	Objects in cluster	Distribution in classes							Purity
		Class1	Class2	Class3	Class4	Class5	Class6	Class7	
Cluster1	14	0	0	0	0	0	6	8	0.57
Cluster2	43	37	0	3	0	3	0	0	0.86
Cluster3	4	0	0	0	0	0	2	2	0.50
Cluster4	3	0	0	2	0	1	0	0	0.67
Cluster5	16	4	0	0	12	0	0	0	0.75
Cluster6	1	0	0	0	1	0	0	0	1.00
Cluster7	20	0	20	0	0	0	0	0	1.00
Accuracy									0.81

Table 5.10.Clustering results for Soybean dataset using MMR algorithm.

Clusters	Objects in clusters	Distribution in classes				Purity
		Class1	Class2	Class3	Class4	
Cluster 1	10	0	10	0	0	1
Cluster 2	10	10	0	0	0	1
Cluster 3	25	0	0	8	17	0.68
Cluster 4	2	0	0	2	0	1
Accuracy						0.83

Table5.11. Clustering results for Soybean dataset using ITDR algorithm.

Clusters	Objects in cluster	Distribution in classes				Purity
		Class1	Class2	Class3	Class4	
Cluster1	10	5	5	0	0	0.50
Cluster2	10	5	5	0	0	0.50
Cluster3	2	0	0	2	0	1.00
Cluster4	25	0	0	8	17	0.68
Accuracy						0.62

Table 5.12. Clustering results for Soybean dataset using MIGR algorithm.

Clusters	Objects in cluster	Distribution in classes				Purity
		Class1	Class2	Class3	Class4	
Cluster1	22	0	0	5	17	0.77
Cluster2	10	10	0	0	0	1
Cluster3	5	0	0	5	0	1
Cluster4	10	0	10	0	0	1
Accuracy						0.89

Table 5.13. Results summary for average clustering accuracies of six algorithms on three data sets.

Algorithms	Accuracy			Average
	Soybean	ZOO	Breast Cancer	
MMR	0.83	0.91	0.79	0.84
ITDR	0.62	0.87	0.86	0.78
MIGR	0.89	0.81	0.88	0.86
MTMDP	0.83	0.79	0.88	0.83
VPRS	0.57	0.88	0.88	0.78
MGR	0.83	0.78	0.88	0.83

5.6.2. F-Measure.

The F-measures for the Soybean, Zoo and Breast cancer datasets when clustered using MMR, ITDR, MTMDP, VPRS, MGR and MIGR algorithms are summarized in table 5. 14. The highest F-measure for the Soybean dataset clusters (0.88) is achieved by the MIGR algorithm. The highest F-measure for the ZOO dataset clusters (0.92) is achieved by the MMR algorithm. The MGR, VPRS and MTMDP achieved the highest F-measure (0.89) when clustering the Breast cancer dataset. F-Measure average is calculated and presented on the same table 5.14, the MIGR the best F-Measure average (0.87). In summary, the average F-measure for the MIGR algorithm is 4% higher than the MMR, MGR and MTMDP algorithms and 14% higher when compared to the ITDR algorithm and 12% higher when compared to the VPRS algorithm. The F-measure average for each algorithm is calculated in table 5.14.

Table 5.14. F-measure of six algorithms on three databases.

	F-measure			
Algorithms	Soybean	Zoo	Breast Cancer	Average
MMR	0.82	0.92	0.76	0.83
ITDR	0.60	0.74	0.86	0.73
MIGR	0.88	0.81	0.88	0.87
MTMDP	0.82	0.77	0.89	0.83
MGR	0.82	0.78	0.89	0.83
VPRS	0.53	0.84	0.89	0.75

5.6.3 Execution Time.

The time consumed by each algorithm to calculate the results of each database is represented in figure 5.1

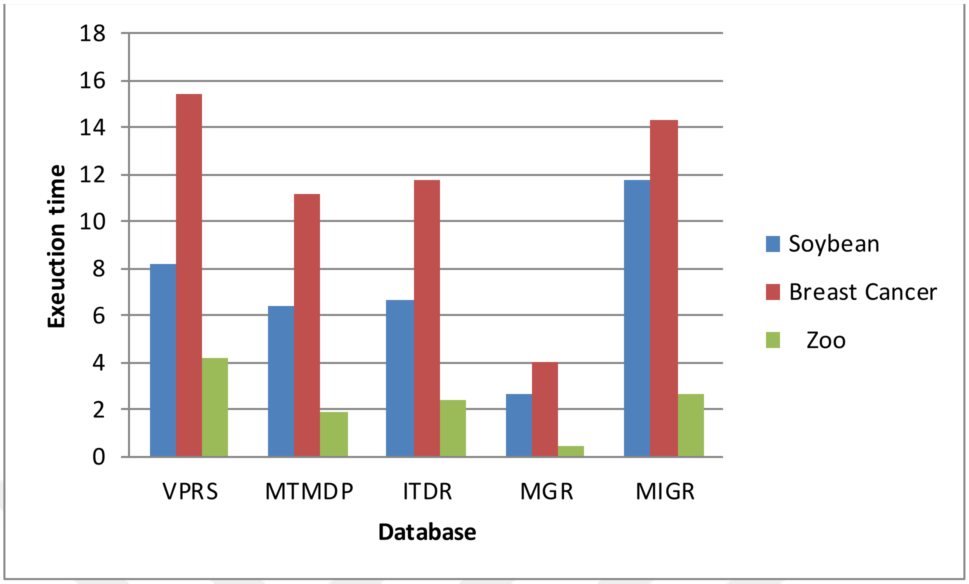


Figure 5.1: Execution time of five algorithms on three databases.

The execution time for each algorithm is measured for three databases as an indication of algorithm simplicity. The algorithm with the lowest time consumption is considered to be more simple, thus, more efficient. The MGR algorithm has the least time consumption when compared to the VPRS, MTMDP, MIGR and ITDR algorithms for all the databases used and we didn't program the MMR algorithm to compare it with VPRS, MTMDP, MGR, ITDR and MIGR. The average execution time for each algorithm on three databases is calculated in table 5.15.The MGR algorithm, when applied to the Soybean, Breast cancer and Zoo, is 2.38 sec

Table 5.15. Execution time of five algorithms on four databases.

Algorithms	Execution time in seconds			Average
	Soybean	Breast Cancer	Zoo	
VPRS	8.16	15.39	4.17	9.24
MTMDP	6.37	11.15	1.91	6.48
ITDR	6.68	11.76	2.4	6.95
MGR	2.67	3.99	0.469	2.38
MIGR	11.77	14.31	2.67	9.58

5.6.4. Clustering results

Clustering results for each dataset using MIGR, MMR, ITDR, VPRS, MTMDP and MGR algorithms are shown and discussed in this section. The clustering accuracy and **F-measure** are measured and compared for each clustering results. The resulting clusters of some other techniques are also compared, like fuzzy centroids, k-modes and fuzzy k-modes, which are unstable techniques when used to cluster categorical data. The modes initial values and dataset's objects order of processing affect these clustering techniques. Furthermore, a membership control parameter needs to be adjusted by the fuzzy k-modes to get better solutions. These unstable clustering techniques are compared directly with the literature results for the sake of the comparison objectivity.

Breast Cancer. As it contains two types of tumors, this dataset is clustered into two clusters using MIGR, MMR and ITDR algorithms. The clustering accuracies for the MIGR, MMR and ITDR are in tables 5.4, 5.5 and 5.6 respectively, In addition The clustering accuracies for the MGR, MTMDP and VPRS are in table 5.13 .These tables also show the accuracy of each algorithm when applied to the breast cancer dataset which illustrates that the MIGR, MGR,VPRS and MTMDP have the superiority over the MMR and ITDR algorithms with (0.88) accuracy, the result in M. Li, S. Deng et al.[14] show that the accuracy of MMR is (0.79). The ITDR has (0.86), While the MGR,MTMDP and VPRS have better performance than the MMR, MIGR and ITDR when compared using the F-measure as illustrated in table 5.14, the MGR,MTMDP and VPRS achieved 0.89 while MIGR, MMR and ITDR achieved 0.88, 0.76 and 0.86 respectively.

Furthermore, the results in F.Y. Cao et al.[62,63] show that the accuracy of k-modes, fuzzy k-modes and fuzzy k-modes for the breast cancer dataset are 0.83, 0.80 and 0.83, respectively. The MIGR algorithm outperforms k-modes and fuzzy k-modes for Breast Cancer dataset .

Zoo. With seven types of animals in this dataset, it is clustered into seven clusters using MMR, ITDR and MIGR algorithms. Objects distributions in the resulting clusters for MMR, ITDR and MIGR are shown in tables 5.7, 5.8 and 5.9 respectively.

The accuracy of each algorithm when applied to the zoo dataset is shown in each table. The best accuracy for the resulting clusters of the zoo dataset is achieved by the MMR with (0.91) while the results in IK .Park et al. [15] show that the accuracy of ITDR is (0.87), While the MIGR algorithm achieved (0.81). The MMR algorithm also has better performance when compared to the ITDR, MIGR, MGR, MTMDP and VPRS algorithms when applied to the zoo dataset and compared using the F-measure as illustrated in table 5.14 , the MMR has (0.92), while the ITDR, MIGR, MGR, MTMDP and VPRS have 0.74, 0.81, 0.78, 0.77 and 0.84 respectively.

Furthermore, results in Kim et al. [61] show that the accuracy of k-modes, fuzzy k-modes and fuzzy centroids on the Zoo dataset are 0.60, 0.64 and 0.75, respectively. Clearly, the MIGR algorithm and it outperforms k-modes, fuzzy k-modes and fuzzy centroids for Zoo dataset.

Soybean. This dataset is consisted of four diseases, therefore, it is clustered into four clusters using MMR, ITDR , MTMDP, MGR, VPRS and MIGR algorithms. The resulting objects distributions are shown in tables 5.10, 5.11 and 5.12 respectively. The resulting clusters accuracy shown in these tables show that the MIGR has the highest accuracy when compared to the clusters resulted from applying the MMR, ITDR, MTMDP, MGR and VPRS algorithms to the soybean dataset. The MIGR has (0.89) accuracy while the MMR, ITDR, MTMDP, MGR and VPRS algorithms have 0.83, 0.62,0.83, 0.83 and 0.57 respectively. When compared using the F-measure as illustrated in table 5.14, the MIGR has also the best clustering results when applies to the soybean dataset with (0.88) while the MMR, ITDR, MTMDP, MGR and VPRS achieved 0.83, 0.62, 0.82, 0.82 and 0.53 respectively.

Furthermore, results in Kim et al. [61] show that the accuracy of k-modes, fuzzy k-modes and fuzzy centroids on the Soybean dataset are 0.69, 0.77 and 0.97, respectively. Clearly, the MIGR algorithm and it outperforms k-modes and fuzzy k-modes in this case.

CHAPTER SIX

Conclusion

In this study, four data clustering methods are executed and compared using purity, F-measure and execution time as performance measures. These algorithms are chosen according to their similarity in the way the most effective attributes are concluded and that they have superiority over other algorithms when compared in earlier studies. For more precision, three UCI databases are used in addition to the Electrical Generator Failures database that was collected so the algorithm with the best performance measures is chosen to conclude the most effective attributes and suggested to the decision n makers in order to improve the maintenance team performance.

The average Purity and F-measure per algorithm is calculated for the four databases used. The MGR algorithm has the superiority over the VPRS, MTMDP and ITDR algorithms, thus, the results of this algorithm on electrical generators failure are proposed to the decision makers. These attributes severely affects the availability of the electrical generator when needed. This doesn't mean that these attributes happen frequently, but special attention must be taken to these attributes in order to maintain the availability of the electrical generator because the occurrence of one of these attributes will definitely disturb the site's operation.

Based on the MGR algorithm results, the attributes with the highest means are the most affective attributes, thus, the "Replacing Air Filter" (RAF), "No Fuel" (NF) and "Owner Problem" (OP) attributes are found as the most effective attributes on electrical generators failure and maintenance team performance. The source of the (RAF) failure is mechanical while the source of both the (NF) and (OP) is site management.

From the attributes concluded, two out of three do not require any maintenance, thus, these attributes affect the maintenance team performance efficiency alongside with the site performance which affects the stability of the service provided, eventually, affecting the company's reputation. These attributes need special attention from other departments than the maintenance department because of their effect on the site's stability as well as the maintenance team efficiency.

Our contribution is the use of the Rough Set Theory and Information Theory on the Electrical Generators Failures database collected to conclude the most effective attributes on the maintenance team performance and electrical generators' availability to suggest them to the decision makers in order to improve the performance of the maintenance team and the generators.

In future work, we recommend studying the factors affecting these attributes in order to improve the efficiency of the maintenance team performance.

Furthermore in chapter five we have proposed a new technique, MIGR (minimum information gain roughness), for selecting the clustering attribute to be used to cluster categorical data. In order to evaluate the performance of this algorithm, it is compared to a very similar categorical data clustering methods (MMR and ITDR) which are proven earlier to perform better than many other methods by (applying these algorithms to three real life UCI datasets and compare the resulting clusters using two performance measure, accuracy and F-measure). The comparison shows that MIGR results better clusters than the resulted from the MMR and ITDR when used for clustering categorical data.

Appendix A. Data Description:

Database Description

Electrical generators failures data was collected through a mobile phone company in Iraq and supplemental data associated with this thesis can be available in the online version at [45], the study aims to analyze the influence of maintenance variables on electrical generators among mobile phone sites, which consists of 636sites (objects) and 38 causes of failures (attributes) grouped into three sources of failure that are mechanical, electrical and Sites management that are beyond the control of the maintenance team. How often each failure affects the availability of the generator was described by choosing one of five options (Never, Rare, Often, Frequent and Severe), these values are stored in a database as (1,2,3,4 and 5) Consecutively. This data was collected for the year 2015.

Electrical Generators failures sources identification.

1 **- Mechanical sources.**

There are Nineteen attributes of Mechanical failure ; dynamo engine (DE) , radiator(RA),fan belt (Fb) , water pump (WP) ,high temperature(HT) , Oil sensor(OL), Repairing fuel injection pump (RFIP), Replacing nozzles(RN), oil consumption (OC), Repairing Relay of starter motor(RFIP),starter motor(SM), Replacing join(RJ), Pump setting(PS), Battery idle(BI), Replacing air filter(RAF), Replacing fuel filter(RFF), oil filter damaged(OFD) and Replacing Engine(RE).

2-Electrical sources.

There twelve attributes of Mechanical failure, Automatic Transfer Switch (ATS) contactor (AC),protection card (PC),Replacing fuses of dynamo Generator(RFDG),main circuit breaker(MCB),Phase failure (PF), Over current(OC), over voltage and Under Voltage (OVAUV),Repairing problem in wirings system(RPIWS), Replacing MDB contactor (RMC) , Auto-start(AS), programmable logic controller (PLC) setting(PS) and Replacing Automatic voltage regulator (RAVR).

3- Site management source include.

a- Financial. Like, running out of fuel.

b- Telecom department. Like, Radio base station problem.

There are seven attributes of Site management failure; False alarm(FA), over load(OL), commercial power problem (CPB), owner problem (OP), repairing power supply unit (RPSU), Radio base station problem (RBSP) and No Fuel (NF).

Electrical Generators Failures Classification:

With the failure source mentioned, each electrical generator failures for a year may fall in one of the following classes:

1. **Mechanical.** This class contains the sites that had only mechanical failures over the period data was collected through. There are 33 sites within this class.
2. **Electrical.** This class contains the sites that had only electrical failures over the period the data was collected though. There are 40 sites within this class.

3. **Sites management.** This class contains the sites that had only Site management failures over the period data was collected through. There are 22 sites within this class.
4. **Mechanical and Electrical.**This class contains the sites that had both mechanical and electrical failures during the period the data was collected through. There are 150 sites within this class.
5. **Mechanical and Site management .** This class contains the sites that had both Mechanical and Site management failures during the period the data was collected though. There are 36 sites within this class.

6. **Electrical and Site management .**This class contains the sites that had both Electrical and Site management failures during the period the data was collected though. There are 88 sites within this class.
7. **Electrical, Mechanical and Site management.**This class contains the sites that had the three kinds of failures (Electrical, Mechanical and Site management) during the period the data was collected through. There are 267 sites in this class.

Appendix B.

Table 1.MGR values for Electrical generators failures database

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13
	0.0423	0.0184	0.0019	0.0137	0.0025	0.0098	0.0009	0.0011	0.0096	0.0018	0.0012	0.0006
0.0464		0.0153	0.0003	0.0091	0.0082	0.0167	0.0003	0.0024	0.0034	0.0015	0.0006	0.0037
0.1321	0.1		0.002	0.0005	0	0.0289	0.0237	0.0026	0.0047	0.0026	0.0016	0.0039
0.025	0.0034	0.0037		0.0109	0.0136	0.0153	0.0014	0.0022	0.0263	0.0022	0.0014	0.0034
0.0432	0.0262	0.0002	0.0026		0.0114	0.0019	0.0021	0.0017	0.0015	0.0034	0.004	0.0004
0.0069	0.0202	0	0.0028	0.0098		0.009	0.0023	0.0008	0.0036	0.0008	0.0025	0.0055
0.0142	0.0221	0.0058	0.0017	0.0009	0.0048		0.0025	0.0076	0.0033	0.0049	0.0076	0.0002
0.0137	0.0045	0.0515	0.0016	0.0106	0.0132	0.0268		0.0022	0.0058	0.0022	0.0014	0.0033
0.0118	0.0234	0.0039	0.0018	0.0056	0.003	0.0557	0.0015		0.0063	0.0024	0.0462	0.0036
0.0472	0.0152	0.0032	0.0097	0.0023	0.0066	0.0113	0.0018	0.0029		0.0029	0.0115	0.0044
0.0189	0.0145	0.0039	0.0018	0.0115	0.003	0.036	0.0015	0.0024	0.0063		0.0015	0.0036
0.0183	0.0082	0.0035	0.0016	0.0195	0.0142	0.0819	0.0014	0.0679	0.0363	0.0022		0.0033
0.0047	0.0255	0.0042	0.0019	0.0009	0.0156	0.0013	0.0016	0.0026	0.0068	0.0026	0.0016	
0.0192	0.0082	0.0816	0.002	0.0128	0	0.0134	0.0016	0.0026	0.007	0.0026	0.0016	0.0039
0.0361	0.0069	0.0019	0.0023	0.0063	0.0024	0.0248	0.0019	0.0049	0.0081	0.0049	0.0019	0.0046
0.1207	0.0447	0.0028	0.0013	0.0083	0.0103	0.2396	0.0011	0.0017	0.0045	0.0017	0.0011	0.0026
0.0717	0.1749	0.0028	0.0013	0.0083	0.0103	0.0258	0.0011	0.0017	0.0045	0.0017	0.0011	0.0026
0.02	0.0277	0.0016	0.0037	0.0094	0.0082	0.0187	0.0002	0.0039	0.0051	0.0003	0.0059	0.0008
0.0909	0.094	0.0039	0.0018	0.0056	0.003	0.0051	0.0015	0.0024	0.0063	0.0024	0.0015	0.0036
0.0052	0.0097	0.0018	0.0008	0.0012	0.0031	0.0045	0.0028	0.0067	0.0031	0.0011	0.001	0.0014
0.0026	0.0234	0.0013	0.0002	0.0038	0.0037	0.0068	0.0015	0.0056	0.0003	0.0019	0.0015	0.0067
0.0139	0.0301	0.0017	0.0011	0.0102	0.0059	0.0257	0.0016	0.0038	0.0101	0.005	0.0024	0.0075
0.009	0.0048	0.0025	0.0015	0.0067	0.0051	0.0156	0.0003	0.0064	0.0014	0.0007	0.0021	0.003
0.0288	0.0727	0.0042	0.0732	0.0125	0.059	0.0176	0.0265	0.0026	0.0058	0.0026	0.0016	0.0039
0.0924	0.0428	0.0321	0.002	0.0079	0.0046	0.0337	0.0017	0.013	0.003	0.0027	0.0193	0.0041
0.0013	0.0093	0.004	0.0019	0.0121	0.0009	0.0019	0.0015	0.0025	0.0066	0.0025	0.0015	0.0037
0.0095	0.0037	0.004	0.0018	0.0118	0.0147	0.0411	0.0015	0.0024	0.05	0.0024	0.0394	0.0037
0.0145	0.007	0.0004	0.0029	0.001	0.0012	0.0049	0.0013	0.0038	0.005	0.0038	0.0024	0.0003
0.0034	0.0493	0.0031	0.0014	0.0091	0.0114	0.0284	0.0012	0.0019	0.005	0.0019	0.0012	0.0028
0.0026	0.0045	0.0027	0.001	0.0025	0.0043	0.0063	0.0021	0.006	0.0005	0.0025	0.007	0.0008
0.0102	0.027	0.0039	0.0018	0.0031	0.0573	0.032	0.0015	0.0024	0.0063	0.0024	0.0015	0.0159
0.0034	0.0493	0.0031	0.0014	0.0091	0.0114	0.0437	0.0012	0.0019	0.005	0.0019	0.0012	0.0028
0.0042	0.006	0.0004	0.0025	0.001	0.0205	0.0091	0.0033	0.0014	0.009	0.0034	0.0033	0.0051
0.0042	0.0105	0.0029	0.002	0.0052	0.0038	0.0073	0.0012	0.0023	0.0034	0.0014	0.003	0.0008
0.0464	0.0583	0.0259	0.0017	0.0108	0.0759	0.0056	0.0014	0.0022	0.0059	0.0022	0.0014	0.0033
0.0183	0.0572	0.0515	0.0016	0.0106	0.0721	0.0016	0.0014	0.0022	0.0058	0.0022	0.0014	0.0033
0.007	0.0133	0.0057	0.0026	0.0028	0.0029	0.011	0.0022	0.0035	0.0121	0.0009	0.0022	0.0029
0.1332	0.0493	0.0031	0.0014	0.1112	0.0114	0.0284	0.0012	0.0019	0.005	0.0019	0.0012	0.0028

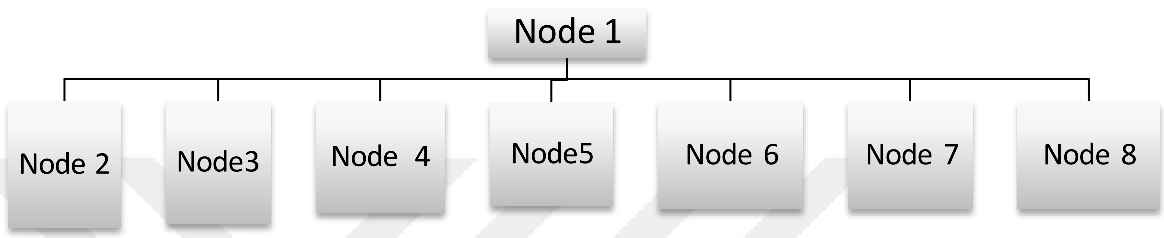
Table 1.MGR values for Electrical generators failures data (Continued)

A14	A15	A16	A17	A18	A19	A20	A21	A22	A23	A24	A25	A26
0.0027	0.0084	0.002	0.0012	0.0186	0.0086	0.0058	0.0019	0.006	0.0071	0.0038	0.0144	0.0001
0.0013	0.0018	0.0008	0.0031	0.0283	0.0097	0.0119	0.0194	0.0141	0.0042	0.0104	0.0073	0.0012
0.0816	0.0031	0.0003	0.0003	0.0104	0.0026	0.0143	0.0068	0.0053	0.0141	0.0039	0.0359	0.0033
0.0037	0.0071	0.0003	0.0003	0.0462	0.0022	0.0117	0.0018	0.0061	0.0155	0.1284	0.0042	0.0028
0.0056	0.0046	0.0004	0.0004	0.0277	0.0017	0.0042	0.009	0.0138	0.0168	0.0052	0.0039	0.0043
0	0.0015	0.0005	0.0005	0.0208	0.0008	0.0094	0.0076	0.0068	0.011	0.0209	0.002	0.0003
0.0027	0.0083	0.0057	0.0006	0.0252	0.0007	0.0072	0.0074	0.0159	0.0179	0.0033	0.0076	0.0003
0.0035	0.0069	0.0003	0.0003	0.0036	0.0022	0.0482	0.0174	0.0109	0.0035	0.0541	0.0041	0.0027
0.0039	0.0119	0.0003	0.0003	0.0381	0.0024	0.0788	0.0444	0.0173	0.0537	0.0036	0.0215	0.003
0.0048	0.0093	0.0004	0.0004	0.0234	0.0029	0.017	0.0013	0.0213	0.0055	0.0038	0.0023	0.0037
0.0039	0.0119	0.0003	0.0003	0.0025	0.0024	0.0124	0.0149	0.0225	0.0056	0.0036	0.0045	0.003
0.0035	0.0069	0.0003	0.0003	0.0855	0.0022	0.0175	0.0174	0.0158	0.0262	0.0033	0.0468	0.0027
0.0042	0.0082	0.0003	0.0003	0.0055	0.0026	0.0119	0.0382	0.0246	0.0179	0.0039	0.0049	0.0032
	0.0031	0.0003	0.0003	0.0226	0.0026	0.0274	0.0183	0.0202	0.0159	0.0039	0.0083	0.0033
0.0019		0.0004	0.0004	0.0354	0.0216	0.0344	0.0166	0.0073	0.0205	0.0046	0.0012	0.0382
0.0028	0.0054		0.0002	0.0526	0.0017	0.1372	0.0318	0.0124	0.2335	0.0026	0.0032	0.0021
0.0028	0.0054	0.0002		0.0526	0.0017	0.0667	0.0318	0.0124	0.0291	0.0026	0.0032	0.0021
0.0034	0.0088	0.0009	0.0009		0.0095	0.0084	0.0105	0.0107	0.0093	0.0044	0.0002	0.0018
0.0039	0.053	0.0003	0.0003	0.094		0.0124	0.0407	0.0225	0.0065	0.0036	0.0045	0.003
0.0034	0.0072	0.002	0.001	0.007	0.0011		0.01	0.0032	0.0068	0.0025	0.0013	0.0007
0.0034	0.0051	0.0007	0.0007	0.013	0.0051	0.0148		0.0016	0.0125	0.0022	0.0063	0.0001
0.0066	0.004	0.0005	0.0005	0.0233	0.005	0.0083	0.0027		0.0055	0.0001	0.0003	0.0001
0.0028	0.006	0.0048	0.0006	0.0109	0.0008	0.0095	0.0119	0.003		0.003	0.0011	0.0035
0.0042	0.0082	0.0003	0.0003	0.0315	0.0026	0.0209	0.0126	0.0003	0.0179		0.0049	0.0032
0.0074	0.0018	0.0003	0.0003	0.0014	0.0027	0.0091	0.0305	0.0009	0.0058	0.0041		0.0034
0.004	0.0783	0.0003	0.0003	0.015	0.0025	0.0066	0.001	0.0005	0.0248	0.0037	0.0047	
0.004	0.0077	0.0003	0.0003	0.064	0.0294	0.0067	0.0454	0.001	0.0727	0.0037	0.0046	0.0249
0.0014	0.0007	0.0005	0.0005	0.0064	0.0038	0.0119	0.01	0.0018	0.0211	0.0058	0.0041	0.0083
0.0031	0.0059	0.0002	0.0002	0.058	0.0019	0.0098	0.0319	0.0137	0.0421	0.0028	0.0035	0.0023
0.0101	0.0026	0.0014	0.0014	0.0093	0.0026	0.0103	0.0014	0.0019	0.0079	0.0008	0.0026	0.0006
0.0039	0.032	0.0003	0.0003	0.0494	0.0221	0.0322	0.0412	0.0124	0.0072	0.0036	0.0045	0.0186
0.1895	0.0059	0.0002	0.0002	0.058	0.0019	0.0736	0.0351	0.0858	0.0321	0.0028	0.0035	0.0023
0.0004	0.0013	0.0004	0.0004	0.0044	0.0034	0.0119	0.005	0.0018	0.0123	0.0051	0.0034	0.0042
0.0025	0.0029	0.0013	0.0011	0.0142	0.0023	0.0083	0.0018	0.0085	0.0154	0.0025	0.0014	0.001
0.0036	0.007	0.0003	0.0003	0.0135	0.0022	0.0183	0.1232	0.0161	0.1006	0.0033	0.0042	0.0028
0.0035	0.0069	0.0003	0.0003	0.0073	0.0022	0.0058	0.0407	0.0158	0.0035	0.0033	0.0041	0.0027
0.0109	0.0025	0.0004	0.0004	0.0055	0.0035	0.0194	0.01	0.0169	0.0302	0.0053	0.0009	0.0039
0.0031	0.0059	0.0002	0.0002	0.058	0.0019	0.0098	0.0351	0.0137	0.0321	0.0028	0.0035	0.0023

Table 1.MGR values for Electrical generators failures data(continued)

A27	A28	A29	A30	A31	A32	A33	A34	A35	A36	A37	A38	Mean
0.001	0.0069	0.0001	0.003	0.0013	0.0001	0.0015	0.0068	0.0047	0.0012	0.0033	0.004	0.0059
0.0004	0.0036	0.0016	0.0058	0.0038	0.0016	0.0024	0.0185	0.0065	0.004	0.007	0.0016	0.0075
0.003	0.0013	0.0007	0.0224	0.0036	0.0007	0.001	0.0336	0.0188	0.0237	0.0195	0.0007	0.0166
0.0025	0.0182	0.0006	0.0158	0.0031	0.0006	0.0121	0.0426	0.0022	0.0014	0.0167	0.0006	0.0123
0.0039	0.0015	0.0009	0.0092	0.0013	0.0009	0.0011	0.0265	0.0034	0.0021	0.0042	0.0105	0.0071
0.0042	0.0015	0.0009	0.0135	0.0201	0.0009	0.0199	0.0165	0.0208	0.0126	0.0038	0.0009	0.0071
0.0062	0.0034	0.0012	0.0105	0.006	0.0019	0.0047	0.0169	0.0008	0.0002	0.0076	0.0012	0.0065
0.0024	0.0099	0.0005	0.0385	0.003	0.0005	0.0186	0.0296	0.0022	0.0014	0.0161	0.0005	0.0113
0.0027	0.0192	0.0006	0.0737	0.0033	0.0006	0.0054	0.0393	0.0024	0.0015	0.0176	0.0006	0.0165
0.0256	0.0116	0.0007	0.0028	0.004	0.0007	0.0158	0.0271	0.0029	0.0018	0.0283	0.0007	0.0091
0.0027	0.0192	0.0006	0.0305	0.0033	0.0006	0.0128	0.0231	0.0024	0.0015	0.0044	0.0006	0.0079
0.0638	0.0176	0.0005	0.126	0.003	0.0005	0.0186	0.074	0.0022	0.0014	0.0161	0.0005	0.0219
0.0029	0.0011	0.0006	0.0068	0.0158	0.0006	0.0139	0.01	0.0026	0.0016	0.0106	0.0006	0.0071
0.003	0.0049	0.0007	0.0837	0.0036	0.0404	0.001	0.0282	0.0026	0.0016	0.037	0.0007	0.0132
0.0034	0.0014	0.0008	0.0129	0.0179	0.0008	0.002	0.0201	0.0031	0.0019	0.0051	0.0008	0.0097
0.0019	0.0138	0.0004	0.0969	0.0023	0.0004	0.0092	0.1228	0.0017	0.0011	0.0126	0.0004	0.0321
0.0019	0.0138	0.0004	0.0983	0.0023	0.0004	0.0092	0.1073	0.0017	0.0011	0.0126	0.0004	0.0207
0.0071	0.0033	0.0019	0.0116	0.0069	0.0019	0.0017	0.0245	0.0015	0.0005	0.0028	0.0019	0.0065
0.0324	0.0192	0.0006	0.0314	0.0303	0.0006	0.0128	0.0393	0.0024	0.0015	0.0176	0.0006	0.0177
0.0006	0.0051	0.0003	0.0107	0.0038	0.002	0.0038	0.0119	0.0017	0.0003	0.0083	0.0003	0.0037
0.0063	0.0063	0.0013	0.0022	0.0071	0.0014	0.0024	0.0039	0.0166	0.0035	0.0063	0.0014	0.005
0.0002	0.0019	0.0009	0.005	0.0038	0.006	0.0015	0.0318	0.0038	0.0024	0.0187	0.0009	0.0068
0.0095	0.0126	0.0016	0.0115	0.0012	0.0012	0.0056	0.0312	0.0129	0.0003	0.0181	0.0012	0.006
0.0029	0.0209	0.0006	0.0068	0.0036	0.0006	0.0139	0.0306	0.0026	0.0016	0.0191	0.0006	0.0141
0.003	0.0124	0.0007	0.0191	0.0037	0.0007	0.0077	0.0142	0.0027	0.0017	0.0026	0.0007	0.0107
0.0229	0.0347	0.0006	0.0066	0.0213	0.0006	0.0134	0.0139	0.0025	0.0015	0.0165	0.0006	0.0088
	0.0152	0.0006	0.0077	0.0033	0.0006	0.0131	0.0293	0.0024	0.0015	0.018	0.0006	0.0147
0.0033		0.001	0.0074	0.0053	0.0051	0.0023	0.0256	0.0038	0.0024	0.0081	0.001	0.0051
0.0021	0.0152		0.005	0.0026	0.0005	0.0101	0.0242	0.0019	0.0012	0.0139	0.0005	0.0101
0.0007	0.003	0.0001		0.002	0.0001	0.0038	0.0304	0.002	0.0005	0.013	0.0103	0.0044
0.0027	0.0193	0.0006	0.0182		0.0006	0.0198	0.0499	0.0024	0.0015	0.0395	0.0006	0.0148
0.0021	0.0814	0.0005	0.005	0.0026		0.0101	0.0242	0.0019	0.0012	0.0139	0.0005	0.0208
0.0038	0.003	0.0008	0.0121	0.0072	0.0008		0.0526	0.0191	0.0021	0.0221	0.0008	0.0067
0.0019	0.0076	0.0004	0.0219	0.004	0.0004	0.0117		0.0022	0.0023	0.0042	0.0004	0.0046
0.0025	0.018	0.0006	0.0225	0.0031	0.0006	0.0675	0.0355		0.1304	0.0697	0.0006	0.024
0.0024	0.0176	0.0005	0.0091	0.003	0.0005	0.0117	0.0562	0.2053		0.0161	0.0005	0.0175
0.0039	0.0081	0.0009	0.0317	0.0108	0.0009	0.0167	0.0142	0.0148	0.0022		0.0058	0.0078
0.0021	0.0152	0.0005	0.3996	0.0026	0.0005	0.0101	0.0242	0.0019	0.0012	0.0931		0.029

Appendix C



Figure(A) The electrical generator clusters obtained

The following sets are related to each node in Figure (A)

Node 1 is the set consist of all object

Node 2 is the set consist of

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100,101,102,103,104,105,106,107,108,109,110,111,112,113,114,115,116,117,118,119,120,121,122,123,124,125,126,127,128,129,130,131,132,133,134,135,136,137,138,139,140,141,142,144,145,146,147,148,149,150,151,152,153,154,155,156,157,158,159,160,161,162,163,164,165,166,167,168,169,170,171,172,173,174,175,176,177,178,179,180,181,182,183,184,185,186,187,188,189,190,191,192,193,194,195,196,197,198,199,200,201,202,203,205,206,207,208,209,211,212,213,214,215,216,217,218,219,220,221,222,223,224,225,226,227,228,229,230,231,232,233,234,235,236,237,238,239,240,241,242,243,244,245,246,247,248,249,250,251,252,253,254,255,256,257,258,259,260,261,262,263,264,265,266,267,268,269,270,271,272,273,274,275,276,277,278,279,280,281,282,283,284,285,286,287,288,289,290,291,292,293,294,295,296,297,298,299,300,301,302,303,304,305,306,307,308,309,310,311,312,313,314,315,316,317,318,319,320,321,322,323,324,325,326,327,328,329,330,331,332,333,334,335,337,338,339

,340,341,342,343,344,345,346,347,348,349,351,352,353,354,355,356,357,358,359,360,361,362,363,364,367,368,369,370,371,372,373,374,375,376,377,378,379,380,381,383,384,385,386,388,389,390,391,392,393,394,395,396,397,398,399,400,401,402,404,406,407,408,410,411,412,413,414,415,416,417,418,419,420,421,422,423,424,425,426,427,428,429,431,432,433,434,435,436,437,438,439,440,441,443,444,445,446,447,448,449,450,451,452,453,455,456,457,458,459,460,461,462,463,464,465,466,467,468,469,470,471,472,473,474,475,476,478,479,480,481,482,483,484,485,486,487,488,489,490,491,493,494,495,496,497,498,499,500,501,502,503,504,505,506,507,508,509,510,511,512,513,514,515,516,517,518,519,521,522,523,524,525,526,527,528,529,530,531,532,533,534,535,536,537,538,539,540,541,542,543,544,545,546,547,548,549,550,551,552,553,554,555,556,557,558,559,560,561,562,563,564,565,566,567,568,569,570,571,572,573,574,575,576,577,578,579,580,581,582,583,584,585,586,587,588,589,590,591,592,593,594,595,596,597,598,599,600,601,602,603,604,605,606,607,608,609,610,611,612,613,614,615,616,617,618,619,620,621,622,623,624,625,626,627,628,629,630,631,632,633,634,635,636

Node 3 is the set consist of:

204,366

Node 4 is the set consist of:

143,365,382,403,405

Node 5 is the set consist of:

210,336,409,430,454,477,492

Node 6 is the set consist of:

387

Node 7 is the set consist of:

350,442

Node 8 is the set consist of:

520

REFERENCES

[1] T. M. Cover, J. A. Thomas , Elements of Information Theory , A Wiley-Inter science Publication (1991) pp(3).

[2] Z. Pawlak, "Rough sets ", International Journal of computer and Information Science, 11 (1982) 341-356.

[3] Ananthanarayana .VS, Narasimha .MM, Subramanian. DK, "Tree structure for efficient data mining using rough sets", Pattern Recognit Lett 24(2003) 851–862.

[4] A.Skowron, Extracting Laws from decision tables: A rough set approach , Computational Intelligence,Vol (II), No(2)(1995) 371-388.

[5] Peters .JF, Skowron . A, A rough set approach to knowledge discovery, Int J, Intell ,Syst 17(2)(2002) 109–112.

[6] S. Tsumoto, Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model", information Sciences international journal , 162(2004) 65–80.

[7] M. C. Magro , P. Pinceti, A confirmation technique for predictive maintenance using the Rough Set Theory, Computers & Industrial Engineering, Vol 56 (2009) 1319–1327.

[8]A.K. Jain,Data clustering: 50 years beyond k-means, International Conference on Pattern Recognition (ICPR), Vol (31),issue(8) ,(2010) 651-666.

[9] P. Kumar, B.K.Tripathy, MMeR : an algorithm for clustering heterogeneous data using rough set theory, International Journal of Rapid Manufacturing, Vol(1),No(2), (2009)189-207.

[10] U. Fayyad, G. Piatetsky-Shapiro , P. Smyth, From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence,(1996) 1-34.

[11] Z.Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data ,Kluwer Academic Publishers(1991).

[12] Z.Pawlak ,A. Skowron, Rudiments of rough sets, Information Sciences, 177(2007) no(1),pp(3-27).

[13] I.T.R.Yanto ,P. Vitasari , T. Herawan, M. M.Deris, Applying variable precision rough set model for clustering student suffering study's anxiety, Expert Systems with Applications, 39(2012) 219-230

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

[15] IK .Park, G. S.Choi, Rough set approach for clustering categorical data using information-theoretic dependency measure, information system , 48(2015)289-295.

[16] H. Qin, X. Ma, T. Herawan, J. M. Zain, MGR: An information theory based hierarchical divisive clustering algorithm for categorical data, Knowledge-Based Systems,67(2014)401-411.

[17] https://archive.ics.uci.edu/ml/machine-learning-databases/dermatology/ .

[18]https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original).

[19] A. Chalifoux , J. Baird, Reliability centered maintenance guide for facilities and collateral equipment, NASA ,(2000)pp(3-13).

[20] X. Hu, Knowledge Discovery in Databases: An Attribute Oriented Rough Set Approach (Ph.D. dissertation), Computer Science Faculty of Graduate studies ,University of Regina(1995).

[21] P. Bastos, I. Lopes and L. Pires,A Maintenance Prediction System using Data Mining Techniques, Proceedings of the World Congress on Engineering(2012).

[22] T. Herawan1 , W.W. Mohd, Applying Variable Precision Rough Set for Clustering Diabetics Dataset" ,International Journal of Multimedia and Ubiquitous Engineering 9(2014) 219-230

[23] A. J. Vakharia , J. Mahajan, Clustering of objects and attributes for manufacturing and marketing applications”, European Journal of Operational Research , 123(2000), 640-651

[24] A.M.Cruz, Evaluating record history of medical device using association discovery and clustering techniques”, Expert systems with applications , 40(2013), 5292–5305 .

[25] A. Maquee , A. A.Shojaie ,D. Mosaddar, Clustering and association rules in analyzing the efficiency of maintenance system of an urban bus network”, Int J Syst Assur Eng Manag,3(3),(2012) 175-183.

[26] T. HERAWAN, ROUGH CLUSTERING FOR CANCER DATASETS, International Journal of Modern Physics, 9(2012) 240-258

[27] F. E.H. tay, L. Shen, Fault diagnosis based on Rough Set Theory, Engineering Applications of Artificial Intelligence 16(2003) 39–43.

[28] T. Herawan, R. Ghazali, I. T. R.Y anto, and M. M. Deris, Rough Set Approach for Categorical Data Clustering, International Journal of Database Theory and Application , 3(2010) 1.

[29] L. Mazlack, A. He, Y. Zhu, S. Coppock, A rough set approach in choosing partitioning attributes, in Proceedings of the ISCA 13thInternational Conference (CAINE-2000)1–6.

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

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

[32]W. Ziarko, Variable precision rough set model, Journal of computer and system sciences , 46(1993)39-59.

[33] D.Slezka , W. Ziarko, The investigation of the Bayesian rough set model, International Journal of Approximate Reasoning ,40(2005) 81–91

[34] T. Beaubouef, F. E .Petry ,G.Arora, Information- theoretic measures of uncertainty for rough sets and rough relational data bases , J .Inf .Sci. 109(1998)185–195.

[35]B.K. Tripathy, A.Ghosh, SDR: An Algorithm for Clustering Categorical Data Using Rough Set Theory, Communicated to the International IEEE conference to be held in Kerala(2011).

[36] B. K. Tripathy ,A. Ghosh, SSDR: An Algorithm for Clustering Categorical Data Using Rough Set Theory , Advances in Applied Science Research, 2 (3),(2011) 314-326

[37] Z. Huang, Extensions to the K-means algorithm for clustering large data sets with categorical values", Data Min .Know l.Discov.2(3) (1998) 283–304

[38] Z. He, X. Xu , S. Deng ,"k-ANMI: A mutual information based clustering algorithm for categorical data", Information Fusion ,vol(9), pp(223–233)

[39] S. Deng, Z. He, X. Xu, G-ANMI : A mutual information based genetic clustering algorithm for categorical data , Knowledge- Based Systems , 23(2010) 144–149

[40] D. Barbara, J. Couto and Y. Li, COOLCAT: An entropy-based algorithm for categorical Clustering", Eleventh International Conference on Information and Knowledge Management ,(2002) 582-589

[41] P.N.Tan,M.Steinbach,V.Kumar, introduction to data mining, Addison-Wesley Longman Inc, chapter 8,(2006) pp(549)

[42] S. Guha, R.Rastogi ,K. Shim, CURE: an efficient clustering algorithm for large databases, SIGMOD 'Proceedings of the 1998 ACM SIGMOD international conference on Management of data,(1998) 73-84.

[43] U. Fayyad, G. P. Shapiro and P.Smyth, From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence(1996).

[44] Ziarko, W, Set approximation quality measures in the variable precision rough set model, In Proceedings of the Second International Conference on Hybrid Intelligent Systems (HIS02), Soft Computing Systems , 87(2002) 442–452.

[45] http://www.diyalaedu.com/Dataset

[46] https://archive.ics.uci.edu/ml/datasets/Soybean+(Small)

[47] UCI Machine Repository<http://www.ics.uci.edu_/mlearn/MLRepository.html>, 2011.

[48] R. Jensen, Q. Shen, Fuzzy-rough sets assisted attribute selection, IEEE Transactions on Fuzzy Systems 15 (2007) 73–78.

[49] N. Iizuka, M. Oka, H. Yamada-Okabe, M. Nishida, Y. Maeda, N. Mori, T. Takao, T. Tamesa, A. Tangoku, H. Tabuchi, Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection, The Lancet 361 (2003) 923–929.

[50] X.H. Hu, N. Cercone, Learning in relational databases: a rough set approach,Computational Intelligence 11 (1995) 323–338.

[51] R. Jensen, Q. Shen, New approaches to fuzzy-rough feature selection, IEEE Transactions on fuzzy Systems 17 (2009) 824–838.

[52] P. Jia, J. Dai, Y. Pan, M. Zhu, Novel algorithm for attribute reduction based on mutual-information gain ratio, Journal of Zhejiang University (Engineering Edition) 40 (2005) 1041–1044.

[53] K. Kaneiwa, A rough set approach to multiple dataset analysis, Applied Soft Computing 11 (2011) 2538–2547.

[54] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westermann, F. erthold, M.Schwab, C. Antonescu, C. Peterson, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Medicine 7 (2001) 673–679.

[55] T.T. Lee, An information-theoretic analysis of relational databases-part I: data dependencies and information metric, IEEE Transactions on Software Engineering 13 (1987) 1049–1061.

[56] N. Mac Parthalain, Q. Shen, Exploring the boundary region of tolerance rough sets for feature selection, Pattern Recognition 42 (2009) 655–667.

[57] C. Martine De, C. Chris, E.K. Etienne, Fuzzy rough sets: the forgotten step, IEEE Transactions on Fuzzy Systems 15 (2007) 121–130.

[58] J. Mi, W. Wu, W. Zhang, Approaches to knowledge reduction based on variable precision rough set model, Information Sciences 159 (2004) 255-272.

[59] D. Miao, G. Hu, A heuristic algorithm for knowledge reduction, Computer Research and Development 36 (1999) 681–684.

[60] Dai. Jianhua, Xu. Qing, Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification, Applied Soft Computing, 13 (2013) 211–221.

[61] D. Kim, K. Lee, D. Lee, Fuzzy clustering of categorical data using fuzzy centroids, Pattern Recognition Letters. 25 (11) (2004) 1263–1271.

[62] F.Y. Cao, J.Y. Liang, D.Y. Li, L. Bai, A new initialization method for categorical data clustering, Expert Systems with Applications. 36 (2009) 10223–10228.

[63] F.Y. Cao, J.Y. Liang, D.Y. Li, X.W. Zhao, A weighting k-modes algorithm for subspace clustering of categorical data, Neurocomputing. 108 (2013) 23–30.