2012-01-01

Decision Rule Induction for Service Sector Using Data Mining- A Rough Set Theory Approach

Zhonghua Hu
University of Texas at El Paso, cloud7@live.cn

Decision Rule Induction for Service Sector Using Data Mining - A Rough Set Theory Approach

Zhonghua Hu

Department of Industrial Manufacturing & System Engineering

APPROVED:

Tzu-Liang(Bill) Tseng, Ph.D., Chair

Eric D. Smith, Ph.D.

Yirong Lin, Ph.D.

Benjamin C. Flores, Ph.D.
Dean of the Graduate School

Decision Rule Induction for Service Sector Using Data Mining – A Rough Set Theory Approach

By

ZHONGHUA HU

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Industrial Manufacturing & System Engineering

THE UNIVERSITY OF TEXAS AT EL PASO

December 2012

Acknowledgements

With a deep sense of gratitude, I thank my family members and many people for their continuous support and encouragement throughout my life, and without their support this thesis would not have been a possibility. Firstly my sincere thanks to my thesis advisor, Dr. Bill Tseng for choosing me as his student and having belief in me to accomplish this research. His faith in my ability coupled with his positive outlook was the biggest source of strength at times of despair. I am also grateful to him for introducing me to new and interesting technologies in Industrial software development. Finally, I would like to sincerely thank him for all the financial support that he offered during the course of this program.

I would like to express my deepest gratitude to my parents, who inculcated the art of learning in any domain with pure, unselfish and honest passion - this fundamental quality made me grow and appreciate the world in the way I see.

I also want to thanks the Industrial Engineering faculty at The University of Texas at El Paso for guiding me through my master's program and giving me the right knowledge and experience to excel in my areas of interest.

I would also like to extend my gratitude to my thesis committee members, Dr. Eric D. Smith and Dr Yirong Lin for the dedication of their time and participation in the thesis.

Finally, I thank the God for showering all the blessings necessitated for the successful completion of this thesis work.

Abstract

Nowadays, data mining is more widely used than ever before; not only by the academic area, but also in the industry and business area. Apart from execution of business processes, the creation of knowledge base and its utilization for the benefit of the organization is becoming a strategy tool to compete. Despite of having ever growing data bases, the problem is that the finance company fails to fully capitalize the true benefits which can be gained from this great wealth of information. The data mining technology instead of classic statistical analysis is developed to help the people to discover the knowledge inside of the data. In this thesis, a rule-based Rough Set decision system is described to make analysis and prediction for the commercial opportunity in the banking sector.

Key words: Data mining, knowledge, data base, Rough Set, business area.

Table of Contents

List of Tables

List of Figures

Chapter:1 Introduction

1.1 Background

In recent years, due to the rapid improvement of computer performance, cost reduction and successful using of data management technology, the various departments obtain the increasingly high level of information technology, in the same time, producing and collecting huge amount of "ample data but barren knowledge". Decision-makers were difficult to extract valuable knowledge from the vast amounts of data. So this fact pushes people to develop the data analysis technology and related tools. Data analysis tools for information and knowledge can be widely used in business management, production control, market analysis, engineering design and scientific research.

Actually, data mining is a senior data analysis tool. Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of a future observation.

1.2 Motivation of this research

The motivation for this thesis work is from a data mining project on a bank market analysis table. The data of the analysis table is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often more than one contact to the same client in order to access if the product (bank term deposit) would be (or not) subscribed. The classification goal is to predict if the client will subscribe a term deposit. The goal of the thesis is to analyze a large volume of historical market data, extract some important variables, and develop predictive rules on the variables using Rough

Set approach of Data Mining. One of the fundamental tasks in data pre-processing stage is the data-splitting task, i.e., to divide the dataset into training and test sets. A training set is required to build a mining model, while a test set is required to detect overtraining of the discovered model. A validation set is required to test the validity of the model. A crucial aspect of data mining is that the discovered knowledge should be interesting, where the objective metrics for interestingness are surprisingness (unexpectedness), usefulness, and novelty.

The RSES (Rough Set Exploration System) software provides the means for data exploration, classification support, and knowledge discovery of various methods, particular those based on RST. RSES version 2.2 was applied to perform RST analysis. The training sets were used to generate the reducts and rules while the testing sets were used for validation.

The flowchart of this research thesis is illustrated in Chapter 1 introduces the basic information and background of Data Mining and motivation of this research. Chapter 2 focus on the literature reviews of KDD, Data Mining, Rough Set Theory. Chapter 3 explains the Rough set approach use RSES software. Chapter 4 discusses a case study of bank market analysis using Rough Set Theory and compares the result with other approaches. Chapter 5 concludes the result of the whole thesis. Chapter 6 lists some future research which needs to optimize or improve.

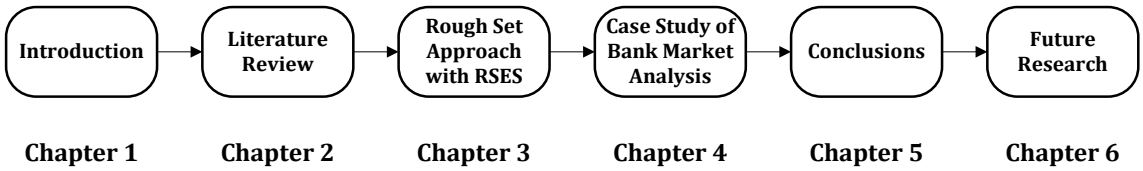


Figure 1.1: The organization of the thesis

Chapter:2 Literature Review

2.1 Knowledge Discovery in Database

2.1.1 What is KDD?

In the late of 1980s, the utilization of machine learning methods is beyond the field of computing and artificial intelligence, especially in the database market. In order to improve the market competitiveness, KDD (Knowledge Discovery in Databases) is proposed to describe all the methodology of discovering relation and decision rule from the known data. Gradually, KDD extended the description to the whole process inferred of the database information, from the determination of the initial business objectives to decision rule being made. Data mining is one integral part of the overall KDD process. The typical process of the KDD should be like this:

I, Data cleaning: remove the noise and inconsistent data.

II, Data aggregation: combine several types of data together

III, Data selection: select the relative data from database for analysis task

IV, Data transform: transform the data to the format which suitable for mining, such as categorical or classification.

V, Data mining: use different intelligent approach to extract the data model or pattern

VI: Model estimation: According one or some interesting estimation, identify the real interesting model of the knowledge representation.

VII, Knowledge representation: Using visualization and knowledge representation technique to provide the knowledge to the user.

The flowchart of the KDD process

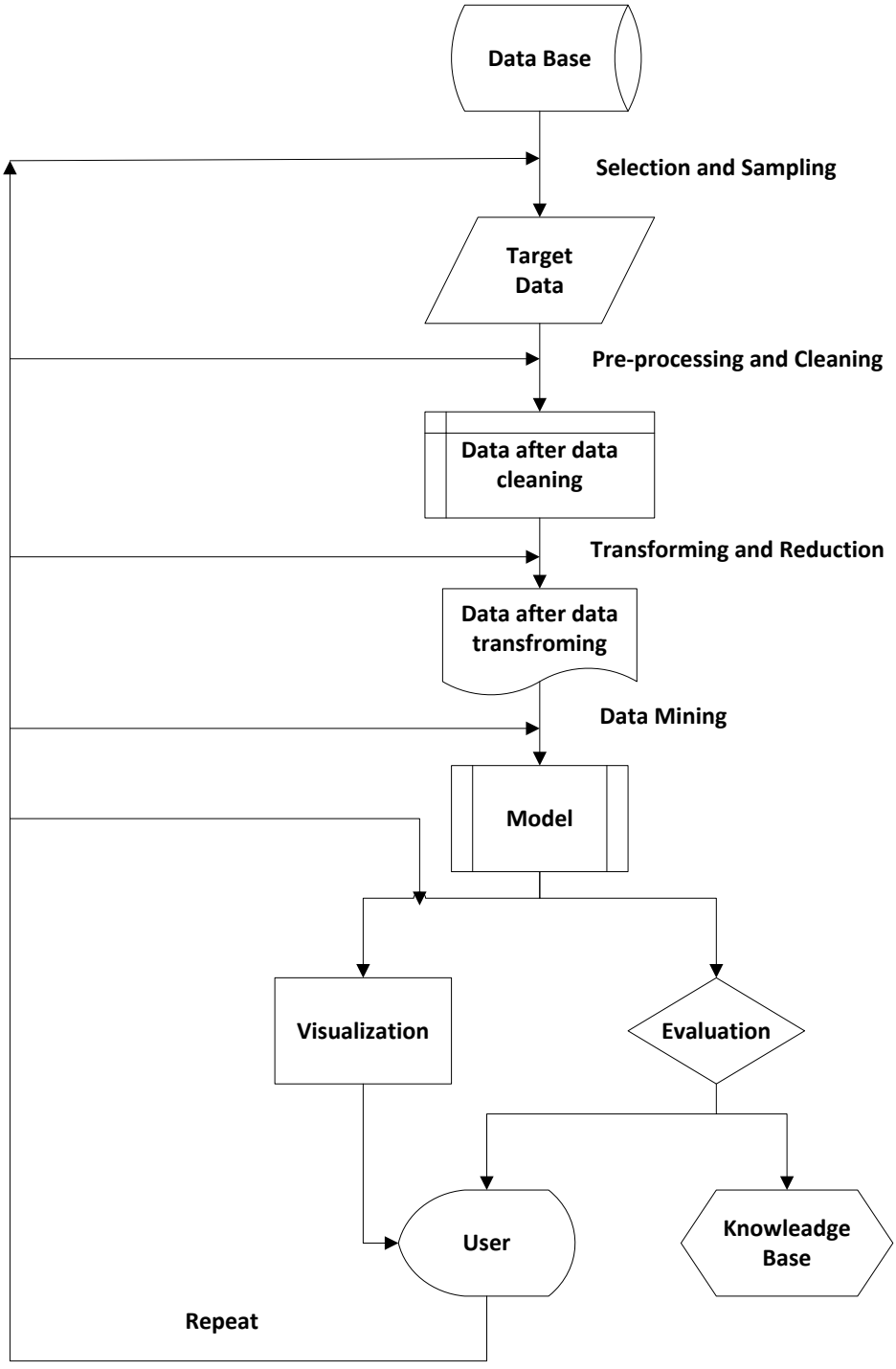


Figure 2.1: Flowchart of KDD

2.1.2 Data Selection

Database and Data Warehouse

A database system is always called as Data Base Management System (DBMS). It consists of several relative data and using software to manage and store these data. DBMS provide definition of structure of database, data query language (SQL), data store, run simultaneously, sharing and Distributed mechanism, authentication and authorization of access of data.

Relational Database consists of tables. Every table has a unique table name. Attribute (column or domain) sets constitute the structure of the table. The data stored in row in the table and each row called "a record". The key value used to be identified between different records. Some attributes in the table describe the relationship between the tables. This kind of semantic model is entity relationship (ER) model. Relational Database is the most popular database now, such as Oracle, SQL Server, Mysql, DB2 etc. Most of them are used to do Online Transaction Processing (OLTP).

A comprehensive definition of Data Warehouse is: it is a set of integrated, subject oriented, design for decision support function (DSF) databases. As the definition said, there's huge difference between Online Transaction Processing (OLTP) and Data Warehouse in structure and purpose of use. The main differences are:

I, Data in Data Warehouse is subject oriented and it based on one or multiple subject. OLTP database is process oriented. The purpose of manage data is to optimize the data updating and retrieving.

II, OLTP system is used in data processing, collecting and management. However, the data stored in the Data Warehouse usually is used to build statements, analyze and validate.

III, OLTP system deals with the data which is an industry or organization's daily operation. The records of the data are always accessed and updated of the transaction database. In contrast, In contrast, part of the data in the data warehouse is no longer used by OLTP environment. Most of the data in Data Warehouse is historical, timestamp and no longer changed (Read only).

The most efficient way of data mining in Data Warehouse is Multi-dimensional Data Analysis, also called Online Analytical Processing (OLAP).

The flowchart of Data Warehouse

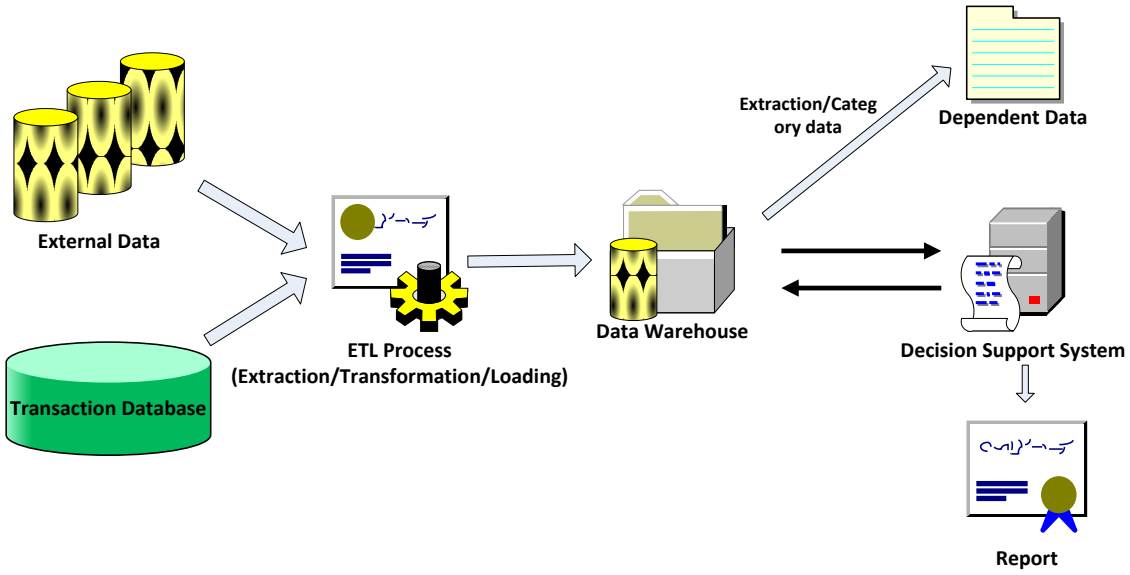


Figure 2.2: The flowchart of Data Warehouse

2.1.3 Data mining

Data mining is an interdisciplinary field, affected by a number of disciplines.

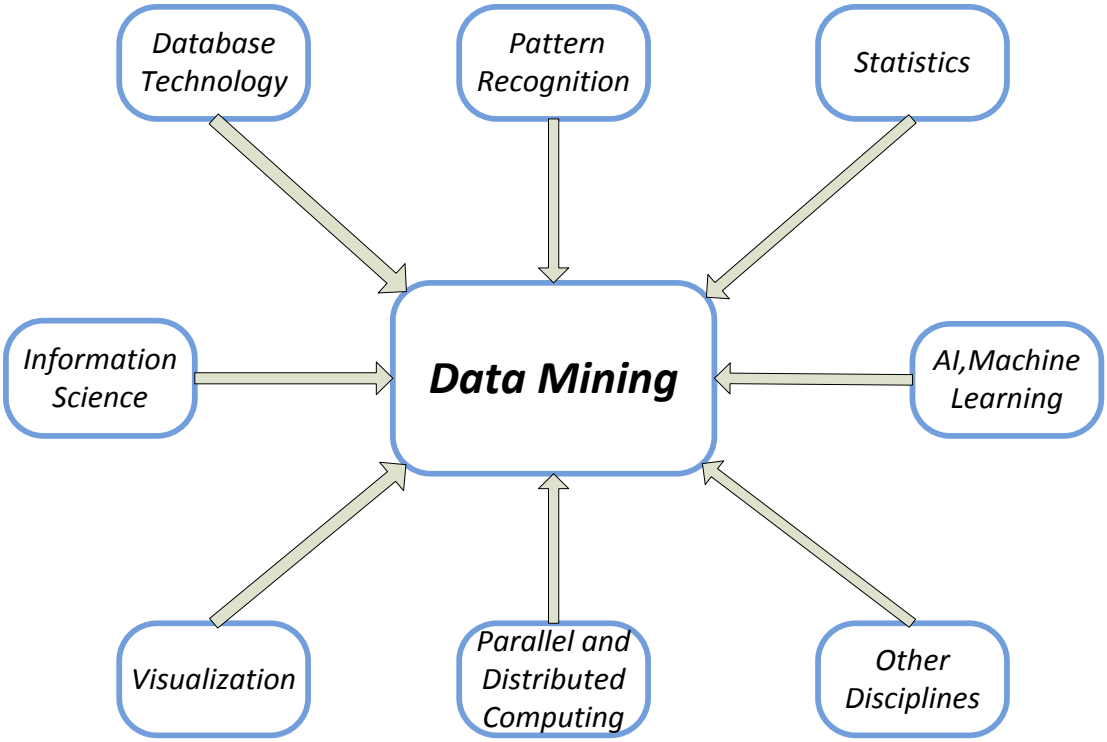


Figure 2.3: Data Mining relationship

In technology part

Not all information discovery tasks are data mining, such as query one record from the DBMS (Database Management System) or use internet search engine to find a specific webpage, they are called as information retrieval. Though these kinds of tasks are important and maybe related to complicated algorithm and data structure, they use the traditional technology of computer science and obvious characteristic to create the structure of index and then do retrieving information effectively.

The same point between information retrieval and data mining is extracting the data and information of interest. The difference is that the information retrieval process uses the pre-

defined information extraction rules. The success rate using pre-defined rules is much high than the other ways. Of course the pre-defined rules are based on the preliminary statistical analysis.

Data mining use to find the unknown relation between various types of phenomenon. There is a well-know example named “Beer and Diapers”, before we do the data mining, we may not imagine the relationship between these two products.

Also Data mining is not the same as OLAP. OLAP is used to show the relation of the 2D variables of the corresponding reports. However data mining can deal with all the variables with different ways of combination. The relationship between these several concepts is shown below.

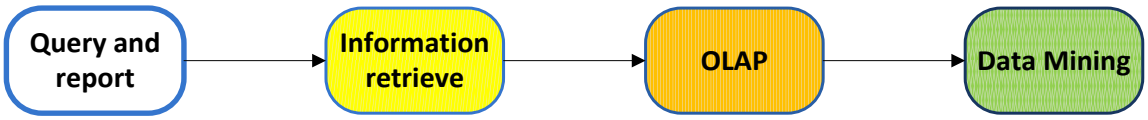


Figure 2.4: Relationship of the several concepts

In commercial part

Data mining is a new business information processing technology. Its main feature is to do extractions, transforming, analysis and modeling process of the data form a large commercial database and then discover the auxiliary business decisions critical data.

Data mining technology has been widely used in the banking and financial markets. Data mining methods, such as the characteristics of the selection and properties of the correlation calculation, helps to identify the important factors, excluding non-related factors. Data mining techniques can be used to find the evolution of the objects in the database features or objects trends. The bank

system can benefit a lot when using data mining approach to make some marketing prediction, such as customer obtain, cross selling and customer maintain.

The flowchart of Data Mining usually like this

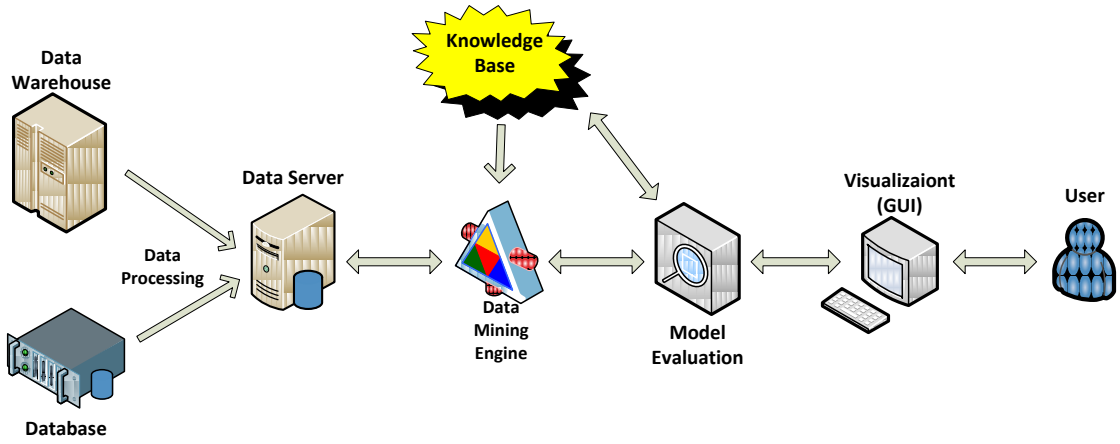


Figure 2.5: the flowchart of Data Mining process

2.1.4 Data Mining Methods

Table 2.1: Various application in data mining

Author	Method	Application
Chidanand Apte	Predictive modeling,Clustering	Frequent pattern extraction
Longbin Cao,…	Combined mining	Actionable patterns
Wu Xindong	NB Classifaction	Error-Aware Data

		Mining
Longbin Cao	Agent-driven distributed Mining	e-commerce,business intelligence, mining Web1.0,2.0,p2p DM
Ataee,P	Windowing, time series, clustering	ATM analysis
Bing Liu, Robert Grossman	Distance string matcing algorithm	Mining Web-page

From the review it was found that in general, Data Mining was used in all kinds of area which cannot be deal with by other discipline knowledge. Data Mining really helps the people not only in the research area but industrial area.

2.2 Rough Set Theory

2.2.1 Rough Set Theory basic idea

Rough set theory is proposed in the early 1980s by professor Pawlak from Warsaw University in Poland, it proposed a research that is not complete, and the expression of uncertainty knowledge and data, learning, inductive theory. It is the one describe incompleteness and uncertainty mathematical tools that can effectively analyze imprecise, inconsistent, incomplete, incomplete information, also do Data analysis and reasoning to discover hidden knowledge, reveal the potential rule. Therefore, the rough sets theory has been widely used in machine learning,

decision support, machine discovery, inductive reasoning, discover knowledge in databases, pattern recognition and other fields.

Rough sets used in the field of data mining, to improve the analysis of incomplete data in large databases and the ability to learn, and has a wide application prospect and practical value. The rough set method using only the information provided by the data itself without any prior knowledge. Rough set is a powerful data analysis tools which can express and deal with incomplete information, retain key information on the minimum expression data simplification and seek knowledge, to identify and evaluate data dependence between relationship to explain the concept of a simple model and get easy proven rules knowledge from empirical data. Research of rough set is a collection of multi-valued attributes (characteristics, symptoms, attributes, etc.) described an object (observer cases) set. There is a value for each object and its attributes. Description of symbols, objects properties and descriptors are the three basic elements of expression of the decision-making problems.

2.2.2 Knowledge and Knowledge Base

In information processing systems, the first thing is the understanding and expression of the knowledge. Knowledge is the result of human understanding through the practice of the rule of motion of the objective world and human's practice and experience and refined. It is abstract with universal characteristics. From the point of view of cognitive science, the knowledge derived from human's classification of objective things. The concept is the description of the category of things or symbol and Knowledge of the relationship between the concepts. Every

species is described by some of the knowledge and classification of species using different attribute knowledge described in different classifications of species.

From the mathematical sense, the equivalence relation on the set and the set division is equal which means divide is classification.

Definition 1

Let $U \neq \emptyset$ be a finite set of objects we are interested in, call it as Universe. Any subset of $X \subset U$ which divided by the equivalent relation of the Universe can be called as a concept or category of the Universe. Any group of concept of the Universe is known as the knowledge. It represents the classification of the objects of the Universe.

Definition 2

K= (U, R), K represents the knowledge base, U (universe) is the set of all the objects. R is the equivalent relation of the Universe and equivalent relation equals to the classification. R is the set of single attribute or multi-attributes. We can use different R to classify the U into different classification.

Definition 3

K= (U, P) and M= (U, Q) are two knowledge base. If IND (P) = IND (Q), then say K and M (or Q and P) are equivalent, X. Hence, when K and M are the sets with the same basic category, the knowledge in both K and M can represent the exactly the same fact about the U precisely. This

concept means we can use different sets of attribute to describe the objects in order to show the fact of the U is absolutely the same.

2.2.3 Indiscernibility Relation

In rough set theory, knowledge is the ability of a classification. Assumes some knowledge uses attributes and attribute values to describe the objects of the Universe. If two objects have the same attributes and attribute values, then we call there's indiscernibility relation between them.

Definition 4

If U is a universe, R is the equivalent relation of the U, U/R represents all of the equivalent class derived from R. $[x]_R$ represents a equivalent class involve element x∈U. A knowledge base K= (U, P) is a relations system, U is the universe and P is a set of the equivalent class. If Q⊆P and Q≠ ∅ , then says Q is indiscernibility relation, marked as IND (Q).

2.2.4 Upper and Lower Approximation

Give a universe U, a set of equivalence relations R divided the U into disjoint basic equivalence classes U / R. Let XgU be an equivalence relation in R. When it can be expressed as some basic equivalence class union, we name it definable otherwise known as undefinable.

We may use two precise set rough set which are upper approximation set and the lower approximation set to define the rough set approximately. Rough set use upper approximation set and the lower approximation set to illustrate the concept of knowledge uncertainty and fuzziness.

Definition 5

If set X ⊆U, R is an equivalent relation, then $\underline{R}X = \{x | x \in U | \text{and } [x]_R \in X\}$ represents the lower approximation of R of the set X.

Also $\overline{R}X = \{x | x \in U \text{ and } [x]_R \cap X \neq \emptyset\}$ represents the upper approximation of R of the set X.

$BN_R(X) = \overline{R}X - \underline{R}X$ is the R boundary domain of set X

Define $Pos_R(X) = \underline{R}X$ as the positive domain of set X

And $NEG_R(X) = U - \overline{R}X$ as negative domain of set X

Example 2.1

If Universe $U = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\}$, R= $\{R_1, R_2\}$ is a group of the equivalent relation, R_1 and R_2 are two equivalent relations. Devide the U separately according to these equivalent relations.

$U/R_1 = \{\{e_1, e_2, e_3, e_4\}, \{e_5, e_6, e_7, e_8\}\}$ and $U/R_2 = \{\{e_1, e_2\}, \{e_3, e_4\}, \{e_5, e_6, e_7, e_8\}\}$ and $U/R = \{\{e_1, e_2\}, \{e_3, e_4\}, \{e_5\}, \{e_6, e_7, e_8\}\}$.

Set $X = \{e_2, e_3, e_6, e_7, e_8\}$ is a subset of U, then X cannot describe precisely use the basic equivalent relation U/R. So X is a rough set of Universe U.

And the lower approximation of X is $Pos(X) = \underline{R}(X) = \{e_6, e_7, e_8\}$

The upper approximation of X is $\overline{R}(X) = \{e_1, e_2, e_3, e_4, e_6, e_7, e_8\}$

The negative domain of set X is $\mathbf{Neg_R(X) = \{e_5\}}$.

2.2.5 Knowledge representation

Knowledge representation occupies a very important position in the intelligent data processing. In the intelligent systems, some of the object expressed in the way of nature language and others use data expression. Sometime the data are accurate, some show the default information or some are contradictory information. In order to process these data we need doing knowledge represent, which named knowledge representation system. Decision table is a special knowledge representation system.

The basic component of the knowledge representation system is the research of a collection of objects and the knowledge of these objects is described by the specified object characteristics (attributes) and their value.

Definition 6

A knowledge representation S may define as $S =< U, C, D, F >$. U is the Universe which contain all the objects $C \cup D = R$ is the set of attributes, subsets C and D are defined as conditional attributes and decision attributes. $V = \bigcup_{r \in R} V_r$ is the set of attributes' value, V_r represents the range of the value of the attribute $r \in R$. $F: U \times R \to V$ is a information function which define the value of object x of the U.

2.2.6 Decision table

The decision table contains large amounts of data in a field. It is the field of the sample database. It records the attribute values and decision-making value of a large number of samples. It is the carrier of domain knowledge. The purpose of knowledge acquisition is to generate useful and regularity knowledge through the analysis of this data set. For the decision table, the most important task is to generate decision rules.

Definition 7

Let $T = (U, P, C, D)$ is a decision table, U is universe, $P = C + D$ is the set of attributes. C is the set of conditional attributes and D is the set of decision attributes. If we remove a conditional attribute Pi, the decision table $T_1 = (U, P - \{P_i\}, C - \{P_i\}, D)$, compared with the original table T has $PosC(D) = PosC - \{P_i\}(D)$, we call attribute Pi is removable for the decision attribute D. If any conditional attribute for the decision attribute is not removable, says the set of conditional attribute C is independent from the set of decision attribute D.

Definition 8

If a decision table $T = (U, P, C, D)$ has a subset B of the set of conditional attribute C is independent from the decision attributes D and $PosB(D) = PosC(D)$, we call B is one D reduct of C.

2.2.7 Discernible Matrix

Discernible Matrix was proposed by Polish Mathematicians Sknowron.

Definition 9

If Decision table $T = \langle U, C, D, F \rangle$, $C = \{c_i | i = 1,2, \ldots, m\}$ is the conditional attribute and $D = \{d\}$ is the decision attribute, $U = \{x_1, x_2, \ldots x_n\}$ is the Universe, $c_i(x_j)$ is the value of sample x_j on the conditional attribute c_i. $M_D(i,j)$ represents the object ith row and jth column in the discernible matrix. Then the discernible matrix can be define as

Here: $i, j = 1,2,3, \ldots, n$

$$M_{D(i,j)} = \begin{cases} \{c_k | c_k \in C \wedge c_k(x_i) \neq c_k(x_j)\}, & d(x_i) \neq d(x_j) \\ 0 \quad , & d(x_i) = d(x_j) \end{cases}$$

Example 2.2

Decision table is a medical record table. It records the body temperature, cough, headache, whole body pain and influenza status of some people.

Table 2.2: Patient data of Influenza

U	Body temperature	cough	headache	Whole body pain	Influenza
1	Normal	No	No	Yes	No
2	Normal	No	Yes	No	No
3	A little higher	No	Yes	No	Yes
4	High	Yes	Yes	No	Yes

5	High	Yes	No	No	Yes
6	A little higher	Yes	Yes	No	Yes

After the data processing, get the discrete type of decision table

Table 2.3: Decision table of Table 2.2

U	a	b	c	d	E
x_1	0	0	0	1	0
x_2	0	0	1	0	0
x_3	1	0	1	0	1
x_4	2	1	1	0	1
x_5	2	1	0	0	1
x_6	1	1	1	0	1

Here, the set of conditional attribute is $\{a, b, c, d\}$ and the set of decision attribute is $\{E\}$

So the discernible matrix is

$$\begin{pmatrix} \emptyset & cd & abcde & abcde & abde & abcde \\ cd & \emptyset & ae & abe & abce & abe \\ abcde & ae & \emptyset & ab & abc & b \\ abcde & abe & ab & \emptyset & c & a \\ abde & abce & abc & c & \emptyset & ac \\ abcde & abe & b & a & ac & \emptyset \end{pmatrix}$$

There's no need to do reduction in $\{E\}$since it is the set of decision attributes. Follow the algorithm we can get the result that the conditional attribute $\{d\}$ can be removed.

Then we do the value of attributes reduction.

Now we get the new decision table without the conditional attribute $\{d\}$

Table 2.4: Decision table after remove the conditional attribute d

U	a	b	c	E
x_1	0	0	0	0
x_2	0	0	1	0
x_3	1	0	1	1
x_4	2	1	1	1
x_5	2	1	0	1
x_6	1	1	1	1

From the Proposition 2 we can compute the core value of the conditional attribute and get the smallest decision rule.

For Decision 1: $[1]_a = \{1,2\}$, $[1]_b = \{1,2,3\}$, $[1]_c = \{1,5\}$, $[1]_E = \{1,2\}$

Then we know

$[1]_a \cap [1]_b = \{1,2\} \cap \{1,2,3\} = \{1,2\} \subseteq [1]_E$, so c_0(represents the conditional attribute c with the value of 0) can be removed

$[1]_a \cap [1]_c = \{1,2\} \cap \{1,5\} = \{1\} \subseteq [1]_E$, so b_0 can be removed.

$[1]_b \cap [1]_c = \{1,2,3\} \cap \{1,5\} = \{1\} \subseteq [1]_E$, so a_0 can be removed.

For Decision 2: $[2]_a = \{1,2\}$, $[2]_b = \{1,2,3\}$, $[2]_c = \{2,3,4,6\}$, $[2]_E = \{1,2\}$

Then we know

$[2]_a \cap [2]_b = \{1,2\} \cap \{1,2,3\} = \{1,2\} \subseteq [2]_E$, so c_1 can be removed.

$[2]_a \cap [2]_c = \{1,2\} \cap \{2,3,4,6\} = \{2\} \subseteq [2]_E$, so b_0 can be removed.

$[2]_b \cap [2]_c = \{1,2,3\} \cap \{2,3,4,6\} = \{2,3\} \not\subset [2]_E$, so a_0 cannot be removed.

The logical meaning can be described as: $a_0 b_0 \vee a_0 c_1 \rightarrow E_0$

The same method can derive that in decision 3: a_1 cannot be removed, b_0 can be removed, c_1 cannot be removed, the logical meaning is

$a_1 b_0 \vee a_1 c_1 \vee b_0 c_1 \rightarrow E_1$

The same:

For decision 4: a_1,b_1,c_1 can be removed.

For decision 5: a_2, b_1, c_0 can be removed

For decision 6: a_1, b_1, c_1 can be removed.

From the above calculation, we may find the smallest decision rules, they are $a_0 b_0 \vee a_0 c_1 \rightarrow E_0$ and $a_1 b_0 \vee a_1 c_1 \vee b_0 c_1 \rightarrow E_1$ which mean (normal body temperature and do not cough) OR

(normal body temperature and have headache) do not catch influenza; (high body temperature and do not cough) OR (high body temperature and have headache) OR (do not cough and have headache) catch influenza.

2.2.8 Rough Set Theory Methods

Table 2.5: Various Application using Rough Set Theory

Author	Method	Application
Sucharita Mitra, Madhuchhanda Mitra	Knowledge-base development and time-domain feature extraction	ECG Classification
Ikno Kim, Yu-yi Chu, Junzo Watada	DNA-Based Algorithm	Efficient discovery of subsets of the lower approximations
Robert Nowicki	Combining Neuro-Fuzzy Architectures with Rough Set Theory	Solve Classifiction Problems with Incomplete Data
Pradipta Maji, Sushmita Paul	Rough sets for selection of molecular(MRMS)	Predict Biological Activity of Molecules
Gwanggil Jeon, Donghyung Kim	Rough set theory	Expert system in interlaced video sequences
Chen-Fu Chien, Li-Fei Chen	Rough set theory	Recruit and retain high-potential talents for semiconductor

		manufacturing
Qiang Li,; Jian-Hua Li; Gong-Shen Liu; Sheng-Hong Li	Mi,DF,CHI,IG with Rough Set and evaluated by naive Bayes mode	Topic-specific text filtering
G Ilczuk, R Mlynarski, …	LEM2 rule indction algorithm based on rough set rule induction	Medical Diagnosis Systems
Wang Chang-long, Qi Yan-ming	Variable precision rough set weight calculation	Web text classification
Vidhya.K.A, G.Aghila	Hybrid test mining(Rough Set and naive bayes classify)	Document classification
Andrew Kusiak	Rough Set Theory	Semiconductor Manufacturing
F.Fernandez-Riverola, F.Diaz, and J.M.Corchado	Rough Set Reduciton, feature subset selection algorithm	Reducing the memory size of a fuzzy case-based reasoning system
Peng Chen, Shuang Liu	Rough Set-based SVM Classifier	Text Categorization

Chapter:3 Methodology: Rough Sets Based Rule Induction

3.1 Determination of the Reducts of Attributes

Attribute reduction

Attribute reduction means deleting irrelevant or unimportant attributes when maintaining the same knowledge base classification ability. A collection of attributes may have multiple reductions.

Proposition 1

Remove the set of attributes P one by one from the decision table from the information system, after every remove action, check if there's no any new inconsistent, then the attribute can be removed otherwise cannot be removed.

Removing an attribute from the decision table repeatedly, each time removing an attribute, check the decision table, if there's no new inconsistent, the attribute can be reduced otherwise the attribute cannot be reduced.

Value Reduction

Attribute reduction is only removed redundant attribute from the decision table in some extent. When judging an object belongs to which classification, the different value of the attribute will cause the different effect of the classification.

Proposition 2

$\Phi \to \Psi$ is a decision rule of the decision table. Both Φ and Ψ are the logical formula of the decision table. The value υ of conditional attribute can be removed if only when $(\phi \to \Psi) \to (\phi - \{\upsilon\} \to \psi)$.

3.2 Decision Rule

The decision table contains a large amount of data in a domain, it is the sample database of the domain. It recorded values of attributes and decisions in a large number of samples. The purpose of knowledge acquisition is to get useful, regularity knowledge in the domain through the analysis of database. The regularity knowledge used to be recorded as the type of "Decision Rule". In the case of decision tables we think about a collection of data, which can be treated by means of various algebraic or statistical methods. Whereas decision rules are logical expressions (implications), of the form 'if…then', which belong to an entirely different realm, and require, in contrast to decision tables, logical means to deal with.

3.3 Problem Definition and Data Processing

The initial and most critical steps to start with are defining the problem, understanding the constraints, and determining the objectives. The prior knowledge can be learned with the help of domain experts in order to ensure the quality and usefulness of the extracted rules from the data. A corrected data should be collected after knowing the information about types and characteristics of data. A well defined sample size is also of importance. The data need to be well prepared before being used for analysis. Data preparation is usually tiring but is a necessary process and involves the following steps: Checking the data distribution, dealing with empty or

missing values, enriching data, and reducing the dimension in order to help transform data into analyzable formats. For our data we used a humanized, scientific method to discrete the continuous number data. Please refer Appendix

3.4 Rough Set Theory approach Implementation

RST approach uses training data set to perform the analysis, generation of reducts and the rules. The predefined Coverage, which means the number of objects matching the rules, is used as the preliminary screening criteria. Therefore, the rules with certain significance level are derived. In the real world, rules may be non-deterministic; that is, in a given data set, the same conditions may imply more than one decision outcome. In such cases, the strength of a rule can be quantified to choose the most frequent class as the best candidate. A strength index is introduced in order to identify meaningful reducts. A reduct with a higher value of the strength index is preferred over a reduct with a lower value index. Note that the comparison of the reducts is restricted to the same decision attribute and the number of attributes selected in the reducts.

Furthermore, the domain experts can help evaluate if the rules are reasonable and useful. If the rule satisfies the significance and usefulness, then it will be selected as the candidate rule. Decomposition trees are used to split data set into fragments not larger than a predefined size. These fragments, after decomposition represented as leafs in decomposition tree, are supposed to be more uniform and easier to cope with decision-wise.

Here’s a flowchart describe how to implement Rough Set Theory in a data mining problem.

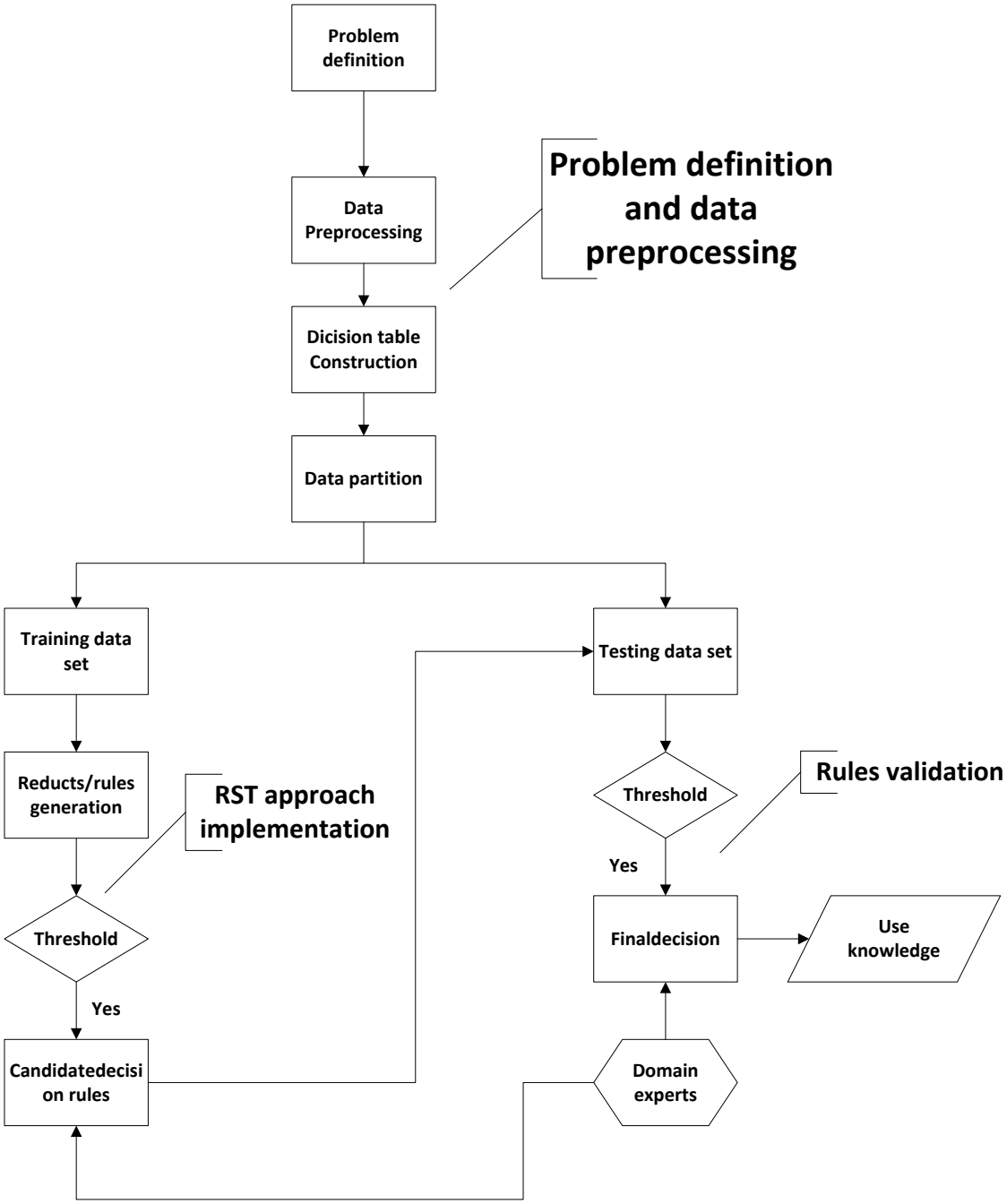


Figure 3.1: a flowchart describe how to implement Rough Set Theory in a data mining project

Example 3.1

There's a prediction of whether a person own a car when has a different background or condition.

Table 3.1: Data of a prediction mode if own a car

ID	Local?	Marital	Income	Own a car?
1	Yes	Single	125K	No
2	No	Married	100K	Yes
3	No	Single	80K	No
4	Yes	Married	120K	Yes
5	No	Divorced	95K	No
6	No	Married	60K	No
7	Yes	Divorced	220K	Yes
8	Yes	Single	65K	No
9	Yes	Married	75K	Yes
10	No	Divorced	90K	No
11	No	Single	110K	Yes
12	Yes	Single	85K	No

Flow the steps, do data processing, category the continuous number and change the words to discrete number and we got the decision table like this.

Table 3.2: Decision Table after dataprocessing of Table 3.1

ID	Local?	Marital	Income	Own a Car?
1	1	0	3	2
2	2	2	2	1
3	2	1	1	2

4	1	2	3	1
5	2	3	2	2
6	2	2	1	2
7	1	3	3	1
8	1	1	1	2
9	1	2	1	1
10	2	3	2	2
11	2	1	3	1
12	1	1	2	2

Based on the table above, build the .tab (see appendix) file according to the RSES table format and syntax and load it in the RSES software. Split the table into two set. 70% of the data is training set and the rest 30% is testing set. Then use the training set to generate the decomposition tree and reduct. From the reduct can generate the rules as shown in the figure. Then classify the testing table with 2 approaches which is Rough Set rules and decomposition tree rules. At last, get the confusion tables of the result for the two methodologies.

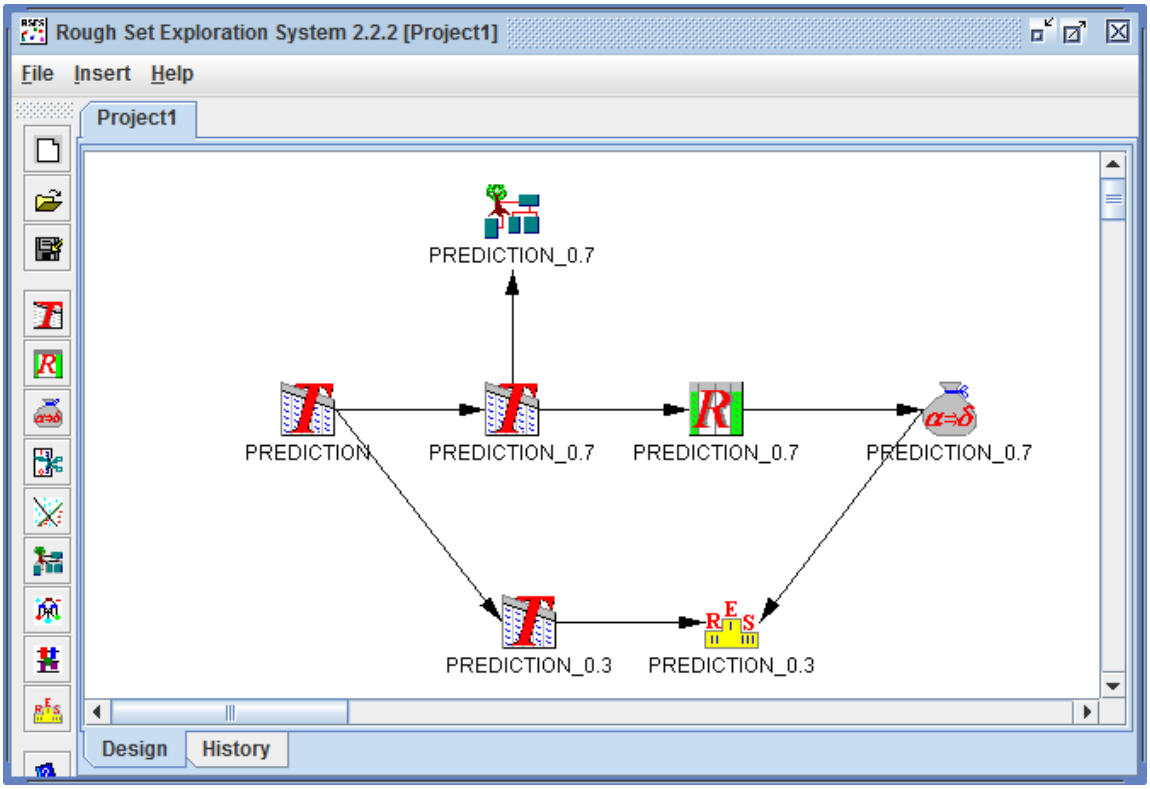


Figure 3.2: Rough Set Exploration System software showing the Prediction model

Here, we use decomposition tree approach to make a comparison.

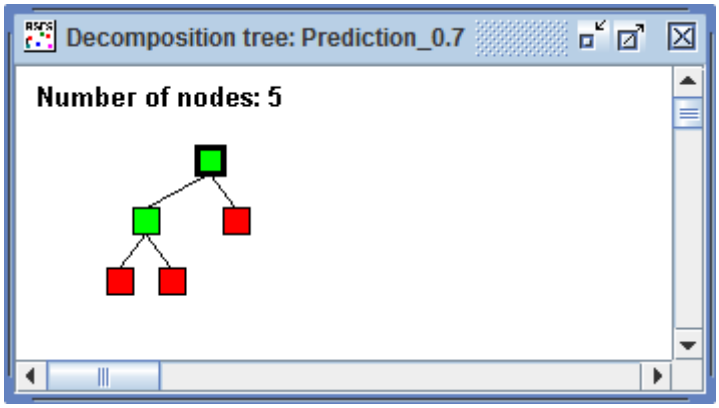


Figure 3.3:Decomposition Tree for Prediction model

Below is the detail information of the decomposition tree method's rules for the training set.

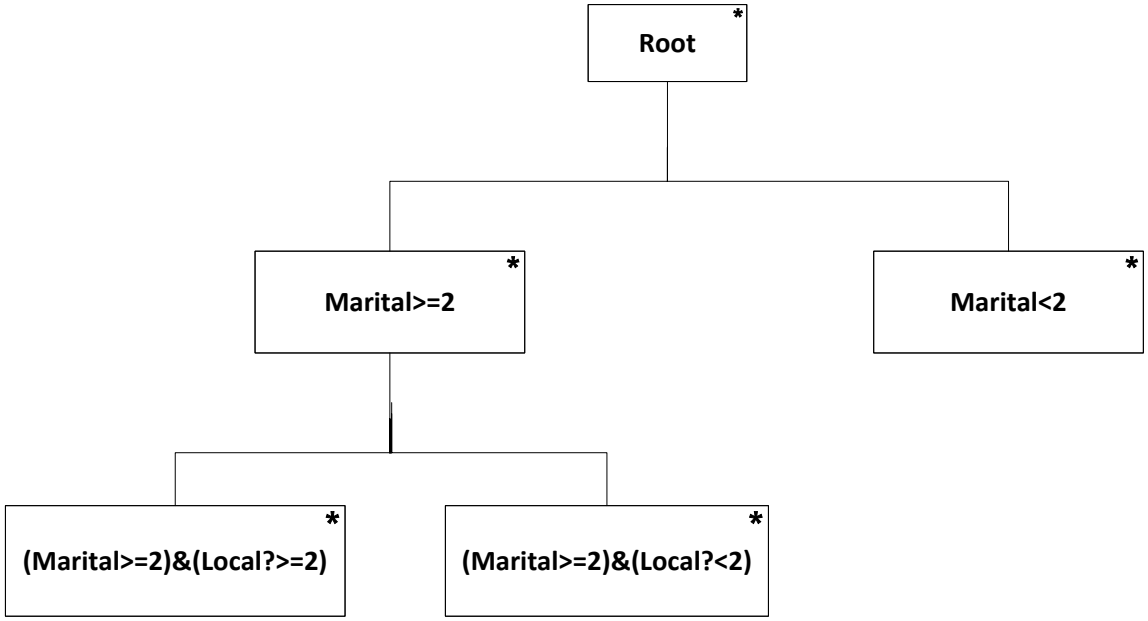


Figure 3.4:Decision tree for the Prediction model

Reduct set: Prediction_0.7

(1-2)	Size	Pos.Reg.	SC	Reducts
1	2	1	1	{ Local?, Marital }
2	2	1	1	{ Marital, Income }

Figure 3.5: Reduct of Prediction model generated by RSES

Below is the rule set of the training set by rough set approach.

Rule set: Prediction_0.7

(1-13)	Match	Decision rules
1	3	(Local?=1)&(Marital=1)=>(Car={2[3]})
2	1	(Local?=2)&(Marital=2)=>(Car={1[1]})
3	1	(Local?=2)&(Marital=1)=>(Car={2[1]})
4	1	(Local?=1)&(Marital=2)=>(Car={1[1]})
5	1	(Local?=1)&(Marital=3)=>(Car={1[1]})
6	1	(Local?=2)&(Marital=3)=>(Car={2[1]})
7	1	(Marital=1)&(Income=3)=>(Car={2[1]})
8	1	(Marital=2)&(Income=2)=>(Car={1[1]})
9	2	(Marital=1)&(Income=1)=>(Car={2[2]})
10	1	(Marital=2)&(Income=3)=>(Car={1[1]})
11	1	(Marital=3)&(Income=3)=>(Car={1[1]})
12	1	(Marital=3)&(Income=2)=>(Car={2[1]})
13	1	(Marital=1)&(Income=2)=>(Car={2[1]})

Figure 3.6:Rule set of Prediciton model generated by RSES

At last, classify the testing table using the decomposition tree and rough set rules and get the confusion table. From the table we can easily find the accuracy and coverage of the two approaches. Then we can determine which method is better and we may apply it in the future work.

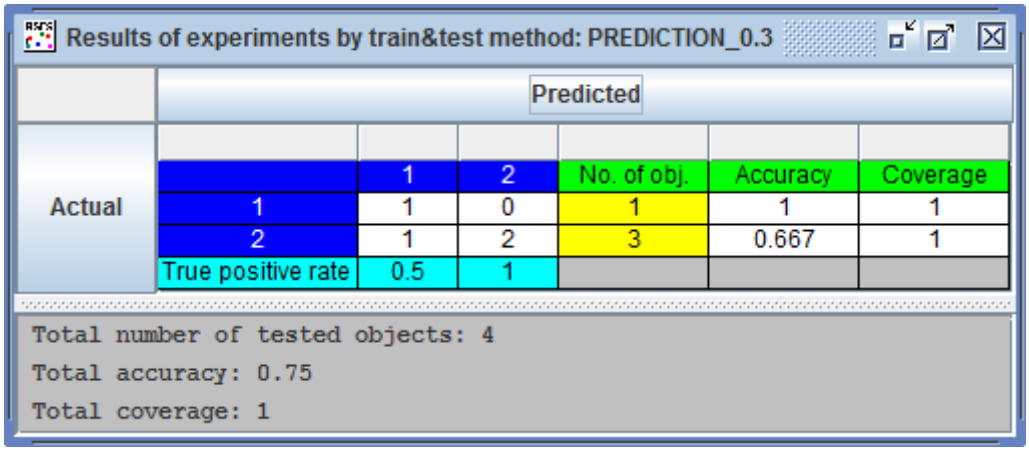


Figure 3.7: Confusion Table of accuracy and coverage using Rough Set Theory

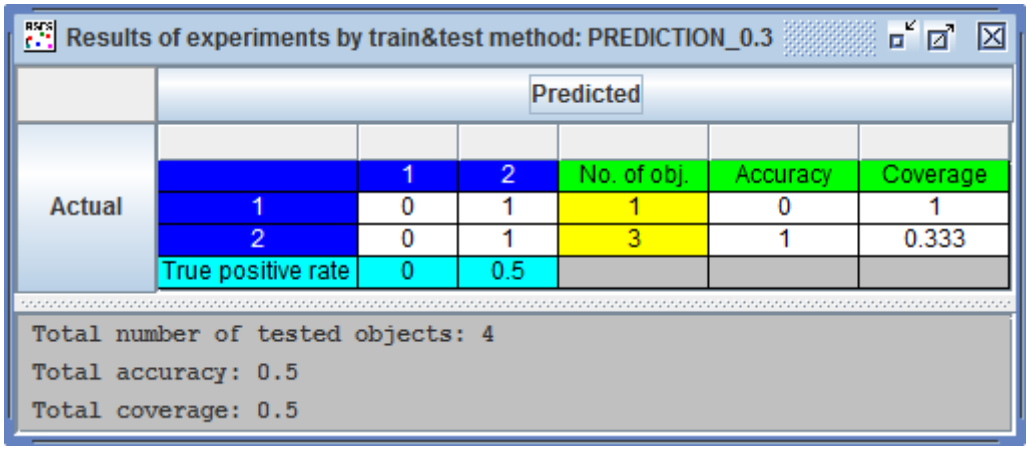


Figure 3.8: Confusion Table of accuracy and coverage using Decomposition Tree

Chapter:4 Case study

4.1 Problem Definition

In this study, a real world data set with around 45,000 records from financial area is used to identify the data mining approach's capability. As we all know, commercial bank system usually saves various type of information of the clients. This kind of information is collected by different use. Some group of category can generate some commercial opportunity with some business intelligent tools. Term deposit is the foundation of a commercial bank and it is the support of other business. Hence exploring the potential client who will subscribe a term deposit is really important and necessary.

This time, for our experiment, I use a small data (10% of the original data) set which is randomly selected from the original data set. So the final data I used is a 4521 instances data with 16 attributes.

This dataset is public available for research. The details are described in [Moro et al., 2011].

[Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimar ães, Portugal, October, 2011. EUROSIS.

The raw data set is related with direct marketing campaigns of a bank. Following table contains 15 input attributes of the bank clients.

Table 4.1: Table of conditional attributes in Bank Market Data

	Attribute Name	Description	Attribute Type
1	age	age of the client	numeric
2	job	type of job	Categorical: admin, unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services.
3	marital	Marital status	Categorical: married, divorced, single, note: divorced means divorced or widowed
4	education	Level of education	Categorical: unknown, secondary, primary, tertiary
5	Default	Has credit in default	Binary:yes,no
6	Balance	Average yearly balance, in euro	Numric
7	Housing	Has housing loan?	Binary: yes, no
8	Loan	Has personal loan?	Binary: yes, no
9	Contact	Contact communicate type	Categorical: unknown, telephone, cellular
10	Month	Last contact month of year	Categorical: jan,feb,mar…nov,dec

11	Duration	Last contact duration, in seconds	Numeric
12	campaign	Number of contacts performed during this campaign and for this client	Numeric include last contact
13	pdays	Number of days that passed by after the client was last contacted from a previous campaign	Numeric:-1 means client was not previously contacted
14	previous	Number of contacts preformed before this campaign and for this client	numeric
15	poutcome	Outcome of the previous marketing campaign	Categorical: unknown, other, failure, success

And the output variable (desired target)

Table 4.2: Table of decision attribute

	Attribute Name	Description	Attribute Type
1	y	Has the client subscribed a term deposit?	Binary: yes, no

In order to get more accurate data mining result, doing data transforming for the various data type such as continuous value is highly recommended. The detail data processing process is shown in the appendix.

4.2 Data Classification in different approaches

Data classification is the process which finds the common properties among a set of objects in a database or data set and classifies them into different classes, according to a classification model. To construct such a classification model, a sample database or data set is treated as the training set, in which each tuple consists of the same set of multiple attributes (feature) as the tuples in large database or data set, and additionally, each tuple has a known class identity (label) associated with it. The objective of the classification is to first analyze the training data and develop an accurate description or a model for each class using the features available in the data. Such class descriptions are then used to classify future test data in the database or data set or to develop a better description (called classification rules) for each class in the database or data set.

In this case, the data set will be split into two partitions as a ratio of 7:3 what means the 70% of the data set will be used to generate decision rules and the other 30% of the data set will be used to test the rules' accuracy. The number of training and testing data set will be shown as below.

Table 4.3: Count of Training and Testing Data

	Training Data	Testing Data	Total
Number	3164	1357	4521
Percentage	70%	30%	100%

4.3 Rough Set Approach and compare with other approaches

Rough set theory addresses the problem where the objects cannot always be assigned to a class crisply. Sometimes, the classes overlap, or it is unclear to which class an object should belong. In such a situation, classifying objects is not a 0-1 problem. This is true for data in the real world, especially in the some specific domain.

Rough set theory is equipped to handle such inconsistent or seemingly conflicting or vague examples of the data. Inconsistencies may occur due to, e.g., transcription errors, subjective determination of attribute values or outcomes, lack of information, etc. The theory of rough sets can handle any finite number of outcome categories and not just dichotomous outcomes.

This time, except Rough Set Theory, we also use decomposition tree to derive the rules in RSES, and test the C5.0 decision tree, Neural Network, Bayes Net and SVM these popular Data Mining approaches with the same data set in SPSS Modeler. The raw data (files) with specific format will be shown in the appendix.

4.4 Implementation

Build the model in RSES like the below figure.

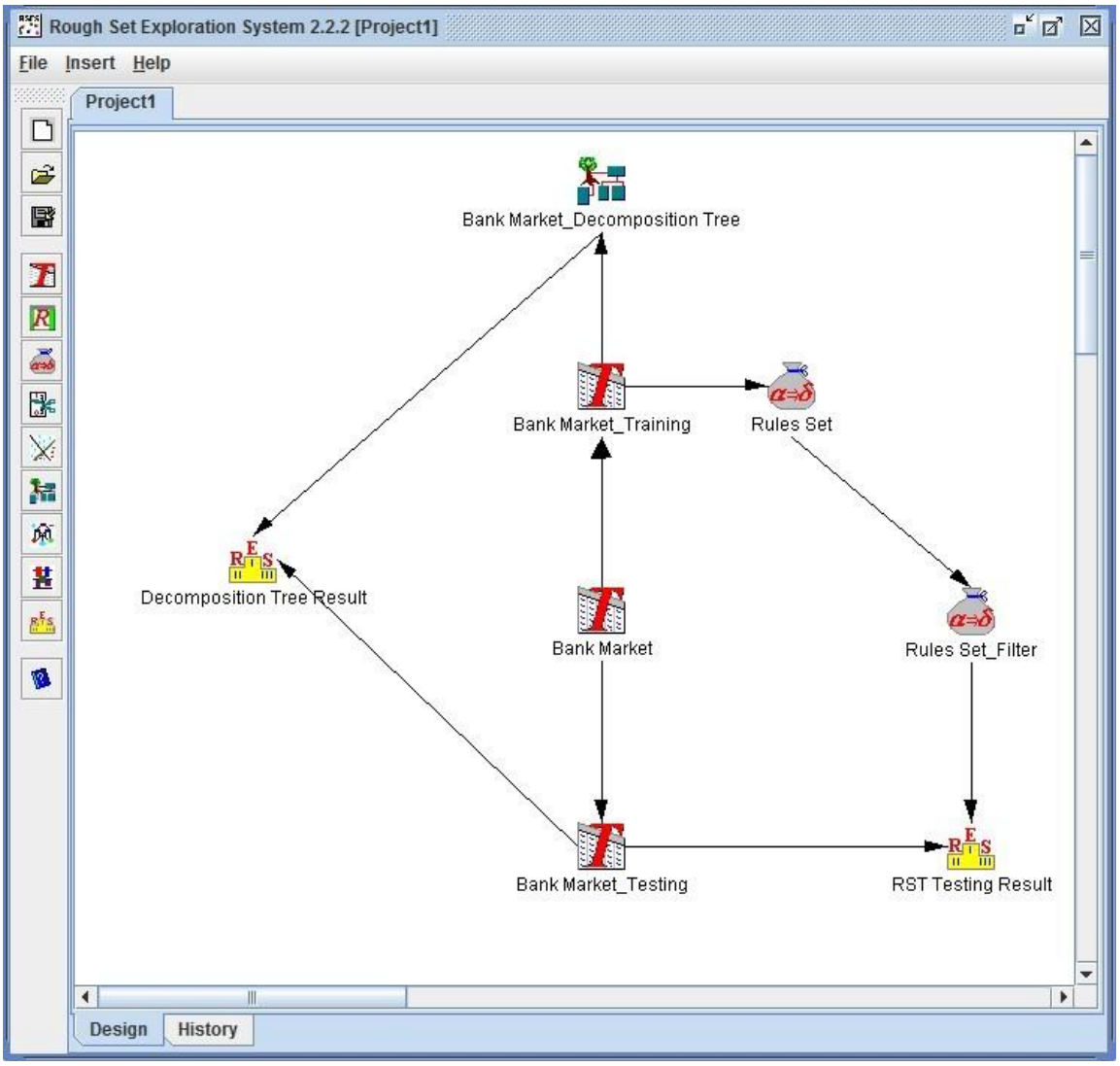


Figure 4.1: Bank Market design model in RSES

Explanation:

Step 1: Create a new project and insert a table.

Step2: Load the data with 4521 instance into the table.

Step3: Split the table into two parts with a ratio as 7:3 which represent training and testing set.

Step4: Generate the rules with Exhaustive Algorithm and also make decomposition for the training set

Step5: Classify both the decomposition tree and Rough set rules use the testing set.

Step6: Get the confusion table and make a comparison.

Next figure shows the whole data set table loading in the RSES.

Table: Bank Market

4521 / ...	job	marital	educa...	default	housing	loan	contact	month	poutc...	balance	durati...	camp...	pdays	previo...	age	Y
O:1	unemployed	married	primary	no	no	no	cellular	oct	unknown	LESS_THAN...	LESS_THAN...	1	NO_CONTR...	NO_CONTA...	<=35	no
O:2	services	married	secondary	no	yes	yes	cellular	may	failure	LESS_THAN...	LESS_THAN...	1	LESS_THAN...	4	<=35	no
O:3	management	single	tertiary	no	yes	no	cellular	apr	failure	LESS_THAN...	LESS_THAN...	1	LESS_THAN...	1	<=35	no
O:4	management	married	tertiary	no	yes	yes	unknown	jun	unknown	LESS_THAN...	LESS_THAN...	4	NO_CONTR...	NO_CONTA...	<=35	no
O:5	blue-collar	married	secondary	no	yes	no	unknown	may	unknown	LESS_THAN...	LESS_THAN...	1	NO_CONTR...	NO_CONTA...	<=65	no
O:6	management	single	tertiary	no	no	no	cellular	feb	failure	LESS_THAN...	LESS_THAN...	2	LESS_THAN...	3	<=35	no
O:7	self-employ...	married	tertiary	no	yes	no	cellular	may	other	LESS_THAN...	LESS_THAN...	1	LESS_THAN...	2	<=50	no
O:8	technician	married	secondary	no	yes	no	cellular	may	unknown	LESS_THAN...	LESS_THAN...	2	NO_CONTR...	NO_CONTA...	<=50	no
O:9	entrepreneur	married	tertiary	no	yes	no	unknown	may	unknown	LESS_THAN...	LESS_THAN...	2	NO_CONTR...	NO_CONTA...	<=50	no
O:10	services	married	primary	no	yes	yes	cellular	apr	failure	LESS_THAN...	LESS_THAN...	1	LESS_THAN...	2	<=50	no
O:11	services	married	secondary	no	yes	no	unknown	may	unknown	MORE_THA...	LESS_THAN...	1	NO_CONTR...	NO_CONTA...	<=50	no
O:12	admin.	married	secondary	no	yes	no	cellular	apr	unknown	LESS_THAN...	LESS_THAN...	2	NO_CONTR...	NO_CONTA...	<=50	no
O:13	technician	married	tertiary	no	no	no	cellular	aug	unknown	LESS_THAN...	LESS_THAN...	2	NO_CONTR...	NO_CONTA...	<=50	no
O:14	student	single	secondary	no	no	no	cellular	apr	unknown	LESS_THAN...	LESS_THAN...	1	NO_CONTR...	NO_CONTA...	<=21	yes
O:15	blue-collar	married	secondary	no	yes	yes	cellular	jan	failure	LESS_THAN...	LESS_THAN...	1	LESS_THAN...	1	<=35	no
O:16	management	married	tertiary	no	no	yes	cellular	aug	unknown	LESS_THAN...	LESS_THAN...	2	NO_CONTR...	NO_CONTA...	<=50	no
O:17	technician	married	secondary	no	no	no	cellular	aug	unknown	LESS_THAN...	LESS_THAN...	5	NO_CONTR...	NO_CONTA...	<=65	no
O:18	admin.	single	tertiary	no	yes	no	cellular	apr	failure	LESS_THAN...	LESS_THAN...	1	LESS_THAN...	2	<=50	no
O:19	blue-collar	single	primary	no	yes	no	unknown	may	unknown	LESS_THAN...	LESS_THAN...	1	NO_CONTR...	NO_CONTA...	<=25	no
O:20	services	married	secondary	no	no	no	cellular	jul	other	LESS_THAN...	LESS_THAN...	1	LESS_THAN...	1	<=35	no
O:21	management	divorced	unknown	no	yes	no	cellular	nov	unknown	LESS_THAN...	LESS_THAN...	2	NO_CONTR...	NO_CONTA...	<=50	no
O:22	management	divorced	tertiary	no	no	no	cellular	nov	unknown	LESS_THAN...	LESS_THAN...	3	NO_CONTR...	NO_CONTA...	<=50	no
O:23	services	single	secondary	no	no	no	unknown	jun	unknown	LESS_THAN...	LESS_THAN...	2	NO_CONTR...	NO_CONTA...	<=50	no
O:24	entrepreneur	married	secondary	no	no	no	cellular	jul	unknown	LESS_THAN...	LESS_THAN...	2	NO_CONTR...	NO_CONTA...	<=50	no
O:25	housemaid	married	tertiary	no	no	no	cellular	jan	unknown	LESS_THAN...	LESS_THAN...	3	NO_CONTR...	NO_CONTA...	<=35	no
O:26	management	married	tertiary	no	no	no	cellular	nov	unknown	MORE_THA...	LESS_THAN...	2	NO_CONTR...	NO_CONTA...	<=50	no
O:27	blue-collar	married	primary	no	yes	no	unknown	may	unknown	LESS_THAN...	LESS_THAN...	1	NO_CONTR...	NO_CONTA...	<=65	no
O:28	retired	married	unknown	no	no	no	telephone	aug	failure	LESS_THAN...	LESS_THAN...	1	LESS_THAN...	2	>65	no
O:29	self-employ...	married	secondary	no	no	yes	cellular	jul	unknown	LESS_THAN...	LESS_THAN...	2	NO_CONTR...	NO_CONTA...	<=65	no
O:30	admin.	married	secondary	no	no	yes	cellular	aug	unknown	LESS_THAN...	LESS_THAN...	2	NO_CONTR...	NO_CONTA...	<=65	no
O:31	retired	divorced	secondary	no	no	no	telephone	jul	unknown	LESS_THAN...	LESS_THAN...	2	NO_CONTR...	NO_CONTA...	>65	yes
O:32	technician	married	secondary	no	no	no	cellular	aug	unknown	LESS_THAN...	LESS_THAN...	3	NO_CONTR...	NO_CONTA...	<=35	no
O:33	management	married	secondary	no	no	no	cellular	nov	unknown	LESS_THAN...	LESS_THAN...	1	NO_CONTR...	NO_CONTA...	<=65	no
O:34	management	single	tertiary	no	yes	no	cellular	aug	unknown	LESS_THAN...	LESS_THAN...	6	NO_CONTR...	NO_CONTA...	<=35	yes
O:35	technician	married	tertiary	no	no	no	cellular	aug	unknown	LESS_THAN...	LESS_THAN...	3	NO_CONTR...	NO_CONTA...	<=50	yes
O:36	admin.	divorced	secondary	no	yes	no	unknown	may	unknown	LESS_THAN...	LESS_THAN...	1	NO_CONTR...	NO_CONTA...	<=50	no
O:37	retired	divorced	primary	no	no	no	telephone	oct	unknown	LESS_THAN...	LESS_THAN...	1	NO_CONTR...	NO_CONTA...	>65	yes
O:38	blue-collar	married	secondary	no	yes	no	cellular	nov	unknown	LESS_THAN...	LESS_THAN...	1	NO_CONTR...	NO_CONTA...	<=35	yes
O:39	management	married	secondary	no	yes	no	cellular	may	failure	LESS_THAN...	LESS_THAN...	1	LESS_THAN...	2	<=35	yes
O:40	services	single	tertiary	no	yes	no	unknown	may	unknown	LESS_THAN...	LESS_THAN...	MORE_THA...	NO_CONTR...	NO_CONTA...	<=25	no
O:41	management	single	tertiary	no	yes	no	unknown	nov	failure	MORE_THA...	LESS_THAN...	2	LESS_THAN...	3	<=50	no
O:42	management	single	tertiary	no	no	no	cellular	aug	unknown	LESS_THAN...	LESS_THAN...	2	NO_CONTR...	NO_CONTA...	<=50	no
O:43	blue-collar	married	secondary	no	yes	no	cellular	may	unknown	LESS_THAN...	LESS_THAN...	1	NO_CONTR...	NO_CONTA...	<=65	no
O:44	technician	married	tertiary	no	yes	no	cellular	may	unknown	LESS_THAN	LESS_THAN	3	NO_CONTR	NO_CONTA	<=35	no

Figure 4.2: Data after data processing loading in RSES table

This is the decomposition tree's diagram. In this case, I set the max leaf about 150.

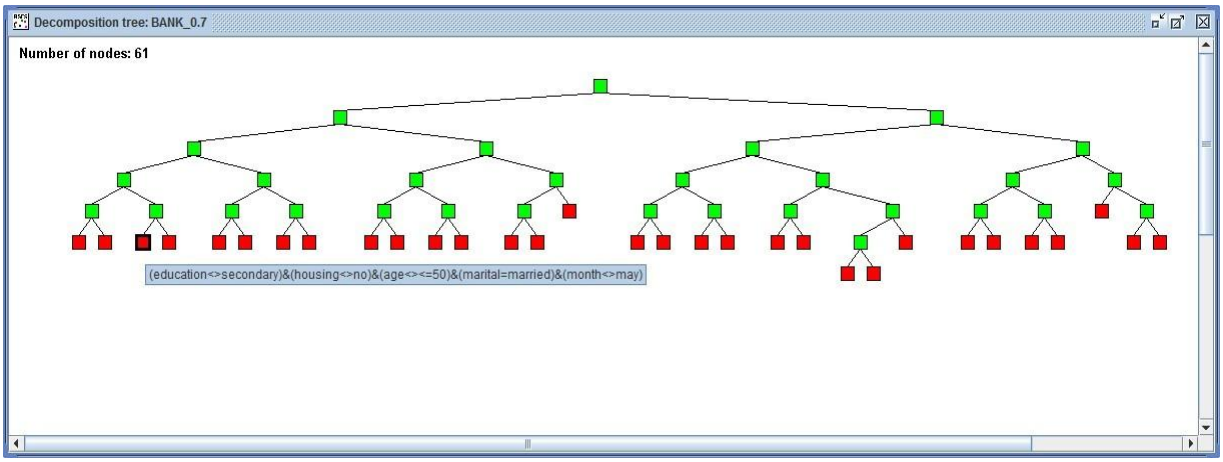


Figure 4.3: Decomposition Tree of the training data

Following 4 figures mean the Statistics for the rule set and the rule set after filter.

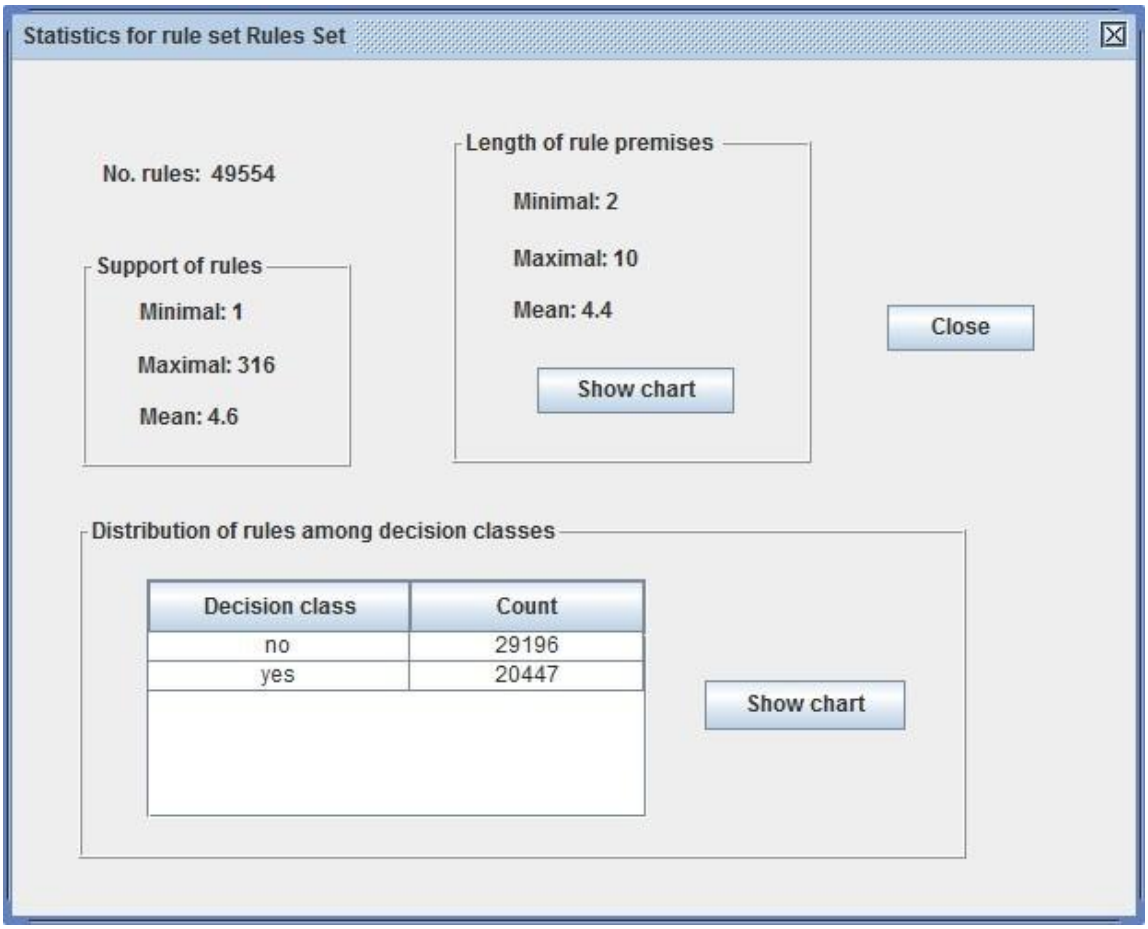


Figure 4.4: Statistics for rule set generated by training set

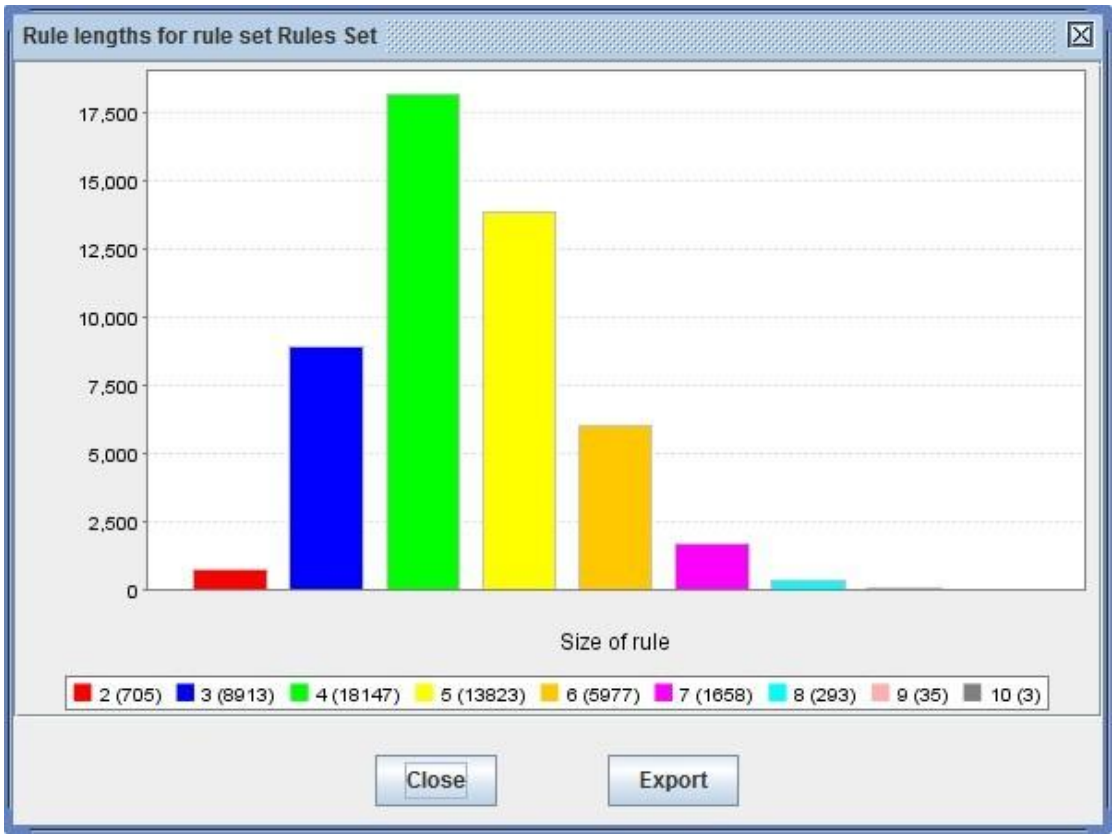


Figure 4.5: chart of rule lengths for rules set generated by training set

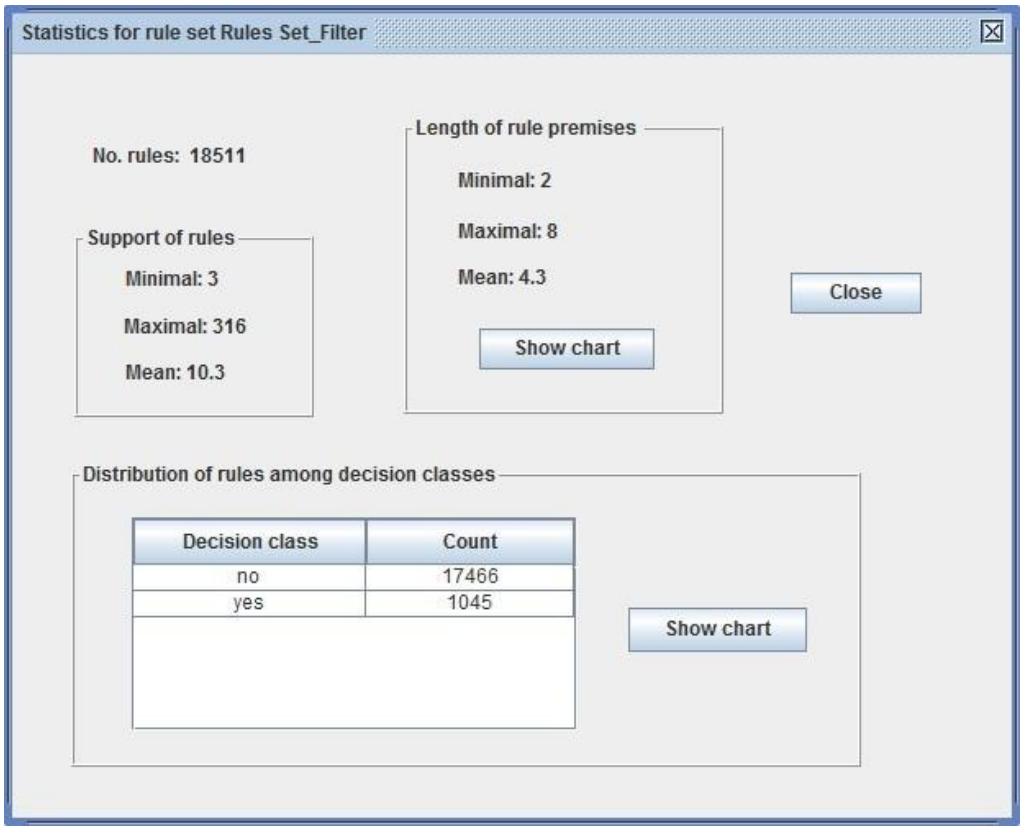


Figure 4.6: Statistics for rule set rules after filtration

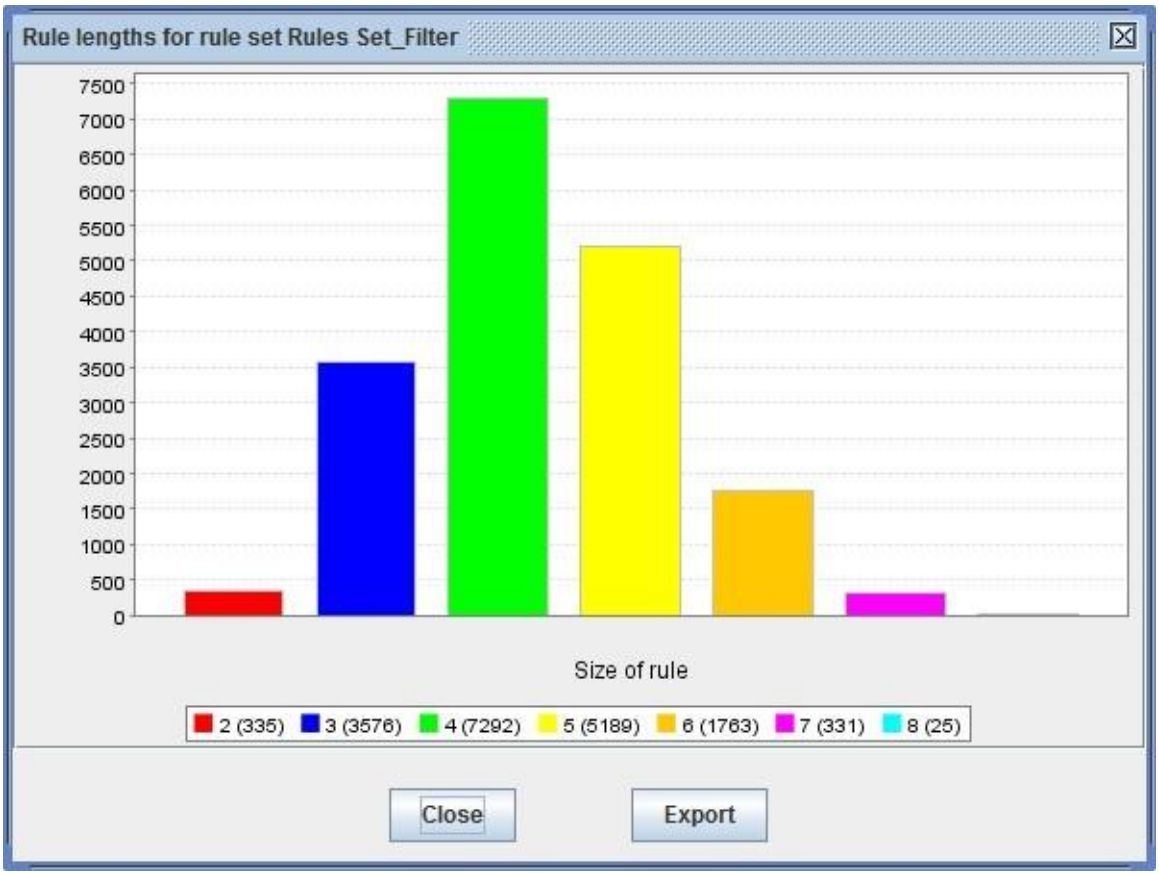
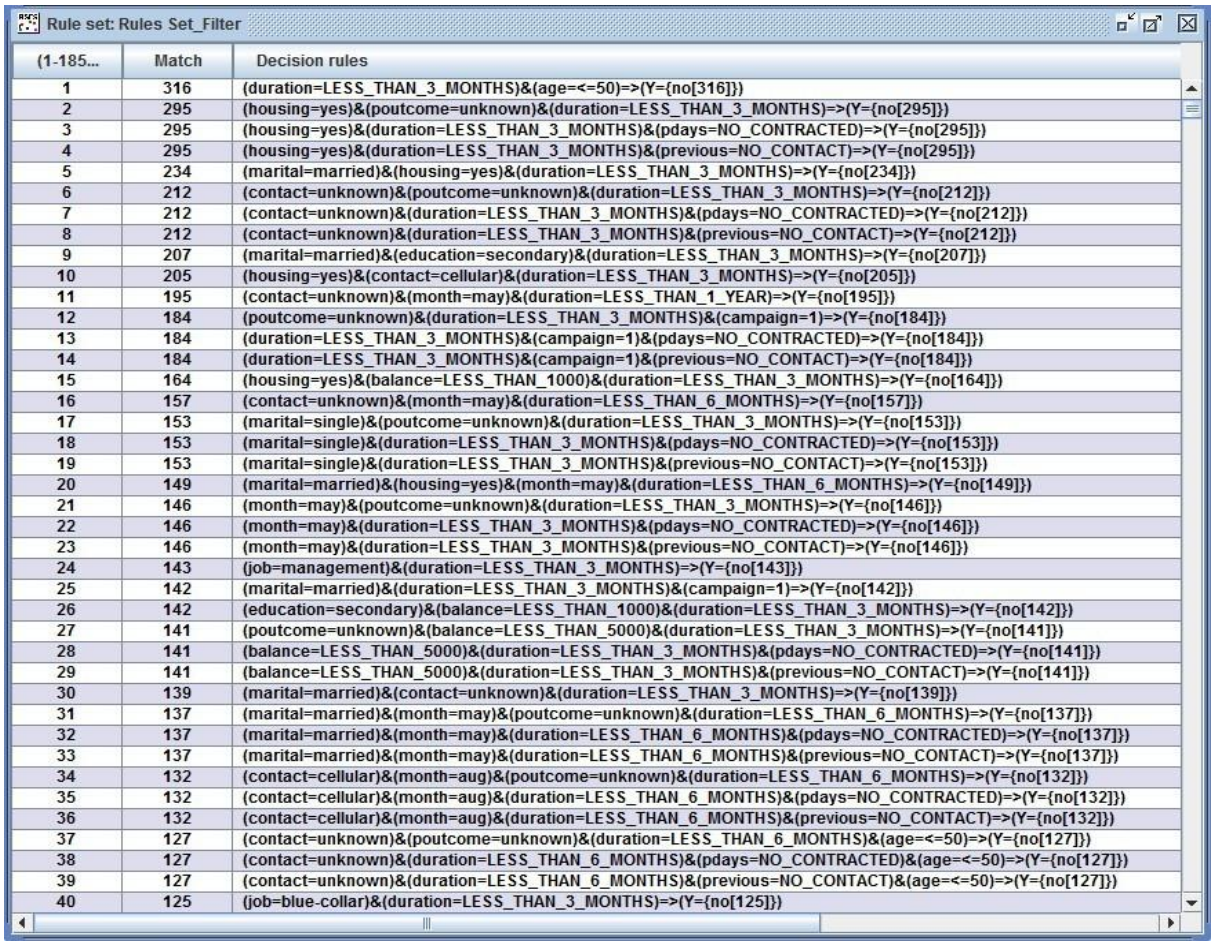


Figure 4.7: Chart of rules lengths for rule set after filtration

The reason why filtrate the decision rules is because the data set is relatively big and a lot of rules only have 1 or 2 supports, they may interfere the classification process. So the final rules used in classify testing data is the following figure.



Rule set: Rules Set_Filter

(1-185...	Match	Decision rules
1	316	(duration=LESS_THAN_3_MONTHS)&(age=<=50)=>(Y={no[316]})
2	295	(housing=yes)&(poutcome=unknown)&(duration=LESS_THAN_3_MONTHS)=>(Y={no[295]})
3	295	(housing=yes)&(duration=LESS_THAN_3_MONTHS)&(pdays=NO_CONTRACTED)=>(Y={no[295]})
4	295	(housing=yes)&(duration=LESS_THAN_3_MONTHS)&(previous=NO_CONTACT)=>(Y={no[295]})
5	234	(marital=married)&(housing=yes)&(duration=LESS_THAN_3_MONTHS)=>(Y={no[234]})
6	212	(contact=unknown)&(poutcome=unknown)&(duration=LESS_THAN_3_MONTHS)=>(Y={no[212]})
7	212	(contact=unknown)&(duration=LESS_THAN_3_MONTHS)&(pdays=NO_CONTRACTED)=>(Y={no[212]})
8	212	(contact=unknown)&(duration=LESS_THAN_3_MONTHS)&(previous=NO_CONTACT)=>(Y={no[212]})
9	207	(marital=married)&(education=secondary)&(duration=LESS_THAN_3_MONTHS)=>(Y={no[207]})
10	205	(housing=yes)&(contact=cellular)&(duration=LESS_THAN_3_MONTHS)=>(Y={no[205]})
11	195	(contact=unknown)&(month=may)&(duration=LESS_THAN_1_YEAR)=>(Y={no[195]})
12	184	(poutcome=unknown)&(duration=LESS_THAN_3_MONTHS)&(campaign=1)=>(Y={no[184]})
13	184	(duration=LESS_THAN_3_MONTHS)&(campaign=1)&(pdays=NO_CONTRACTED)=>(Y={no[184]})
14	184	(duration=LESS_THAN_3_MONTHS)&(campaign=1)&(previous=NO_CONTACT)=>(Y={no[184]})
15	164	(housing=yes)&(balance=LESS_THAN_1000)&(duration=LESS_THAN_3_MONTHS)=>(Y={no[164]})
16	157	(contact=unknown)&(month=may)&(duration=LESS_THAN_6_MONTHS)=>(Y={no[157]})
17	153	(marital=single)&(poutcome=unknown)&(duration=LESS_THAN_3_MONTHS)=>(Y={no[153]})
18	153	(marital=single)&(duration=LESS_THAN_3_MONTHS)&(pdays=NO_CONTRACTED)=>(Y={no[153]})
19	153	(marital=single)&(duration=LESS_THAN_3_MONTHS)&(previous=NO_CONTACT)=>(Y={no[153]})
20	149	(marital=married)&(housing=yes)&(month=may)&(duration=LESS_THAN_6_MONTHS)=>(Y={no[149]})
21	146	(month=may)&(poutcome=unknown)&(duration=LESS_THAN_3_MONTHS)=>(Y={no[146]})
22	146	(month=may)&(duration=LESS_THAN_3_MONTHS)&(pdays=NO_CONTRACTED)=>(Y={no[146]})
23	146	(month=may)&(duration=LESS_THAN_3_MONTHS)&(previous=NO_CONTACT)=>(Y={no[146]})
24	143	(job=management)&(duration=LESS_THAN_3_MONTHS)=>(Y={no[143]})
25	142	(marital=married)&(duration=LESS_THAN_3_MONTHS)&(campaign=1)=>(Y={no[142]})
26	142	(education=secondary)&(balance=LESS_THAN_1000)&(duration=LESS_THAN_3_MONTHS)=>(Y={no[142]})
27	141	(poutcome=unknown)&(balance=LESS_THAN_5000)&(duration=LESS_THAN_3_MONTHS)=>(Y={no[141]})
28	141	(balance=LESS_THAN_5000)&(duration=LESS_THAN_3_MONTHS)&(pdays=NO_CONTRACTED)=>(Y={no[141]})
29	141	(balance=LESS_THAN_5000)&(duration=LESS_THAN_3_MONTHS)&(previous=NO_CONTACT)=>(Y={no[141]})
30	139	(marital=married)&(contact=unknown)&(duration=LESS_THAN_3_MONTHS)=>(Y={no[139]})
31	137	(marital=married)&(month=may)&(poutcome=unknown)&(duration=LESS_THAN_6_MONTHS)=>(Y={no[137]})
32	137	(marital=married)&(month=may)&(duration=LESS_THAN_6_MONTHS)&(pdays=NO_CONTRACTED)=>(Y={no[137]})
33	137	(marital=married)&(month=may)&(duration=LESS_THAN_6_MONTHS)&(previous=NO_CONTACT)=>(Y={no[137]})
34	132	(contact=cellular)&(month=aug)&(poutcome=unknown)&(duration=LESS_THAN_6_MONTHS)=>(Y={no[132]})
35	132	(contact=cellular)&(month=aug)&(duration=LESS_THAN_6_MONTHS)&(pdays=NO_CONTRACTED)=>(Y={no[132]})
36	132	(contact=cellular)&(month=aug)&(duration=LESS_THAN_6_MONTHS)&(previous=NO_CONTACT)=>(Y={no[132]})
37	127	(contact=unknown)&(poutcome=unknown)&(duration=LESS_THAN_6_MONTHS)&(age=<=50)=>(Y={no[127]})
38	127	(contact=unknown)&(duration=LESS_THAN_6_MONTHS)&(pdays=NO_CONTRACTED)&(age=<=50)=>(Y={no[127]})
39	127	(contact=unknown)&(duration=LESS_THAN_6_MONTHS)&(previous=NO_CONTACT)&(age=<=50)=>(Y={no[127]})
40	125	(job=blue-collar)&(duration=LESS_THAN_3_MONTHS)=>(Y={no[125]})

Figure 4.8: Rules Set after filtration

Top ten matches rules

Table 4.4: Top ten matches rules

No	Rules	Attributes	Match
1	(marital=married)&(duration=LESS_THAN_3_MONTHS)=> (Y=no[439])	2	439
2	(housing=yes)&(month=may)& (duration=LESS_THAN_6_MONTHS)=>(Y=no[242])	3	242

3	(month=may)&(poutcome=unknown)& (duration=LESS_THAN_6_MONTHS)=>(Y=no[232])	3	232
4	(month=may)&(duration=LESS_THAN_6_MONTHS)& (pdays=NO_CONTRACTED)=>(Y=no[232])	3	232
5	(month=may)&(duration=LESS_THAN_6_MONTHS)& (previous=NO_CONTACT)=>(Y=no[232])	3	232
6	(poutcome=unknown)&(balance=LESS_THAN_1000)& (duration=LESS_THAN_3_MONTHS)=>(Y=no[224])	3	224
7	(balance=LESS_THAN_1000)& (duration=LESS_THAN_3_MONTHS)& (pdays=NO_CONTRACTED)=>(Y=no[224])	3	224
8	(balance=LESS_THAN_1000)& (duration=LESS_THAN_3_MONTHS)& (previous=NO_CONTACT)=>(Y=no[224])	3	224
9	(contact=unknown)&(month=may)& (duration=LESS_THAN_1_YEAR)=>(Y=no[211])	3	211
10	(contact=unknown)&(poutcome=unknown)& (duration=LESS_THAN_3_MONTHS)=>(Y=no[209])	3	201

At last, 2 confusion tables represent the results of two methods.

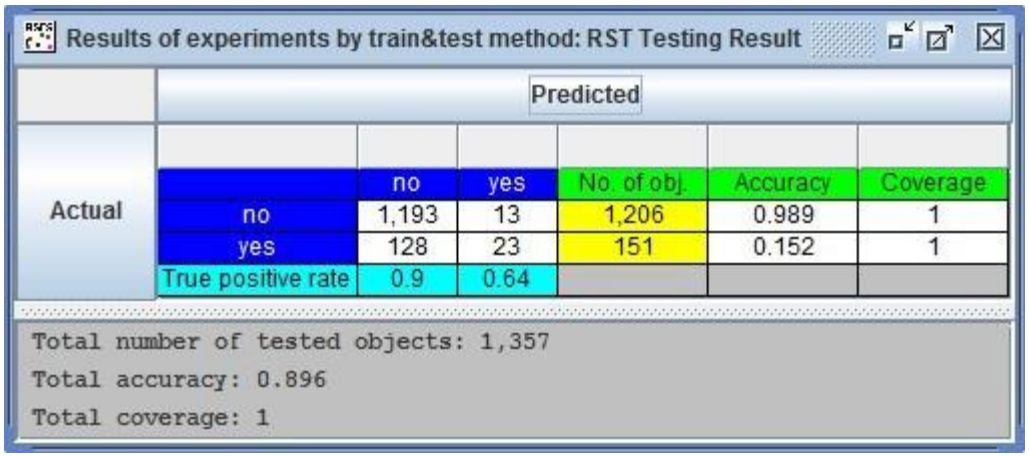


Figure 4.9: Confusion Table of accuracy and coverage using Rough Set Theory

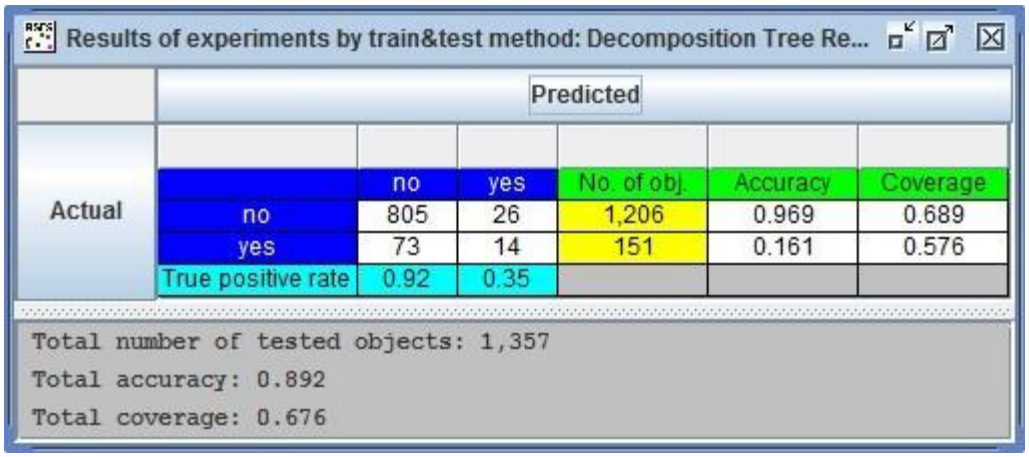


Figure 4.10: Confusion Table of accuracy and coverage using Decomposition Tree

From the results, we may find the accuracy of both methods is really close, but obviously the Rough Set Theory has a high coverage which covers all the testing set even I filter half of the rules.

Next briefly use IBM SPSS Modeler to get the analysis of the other for data mining approaches.

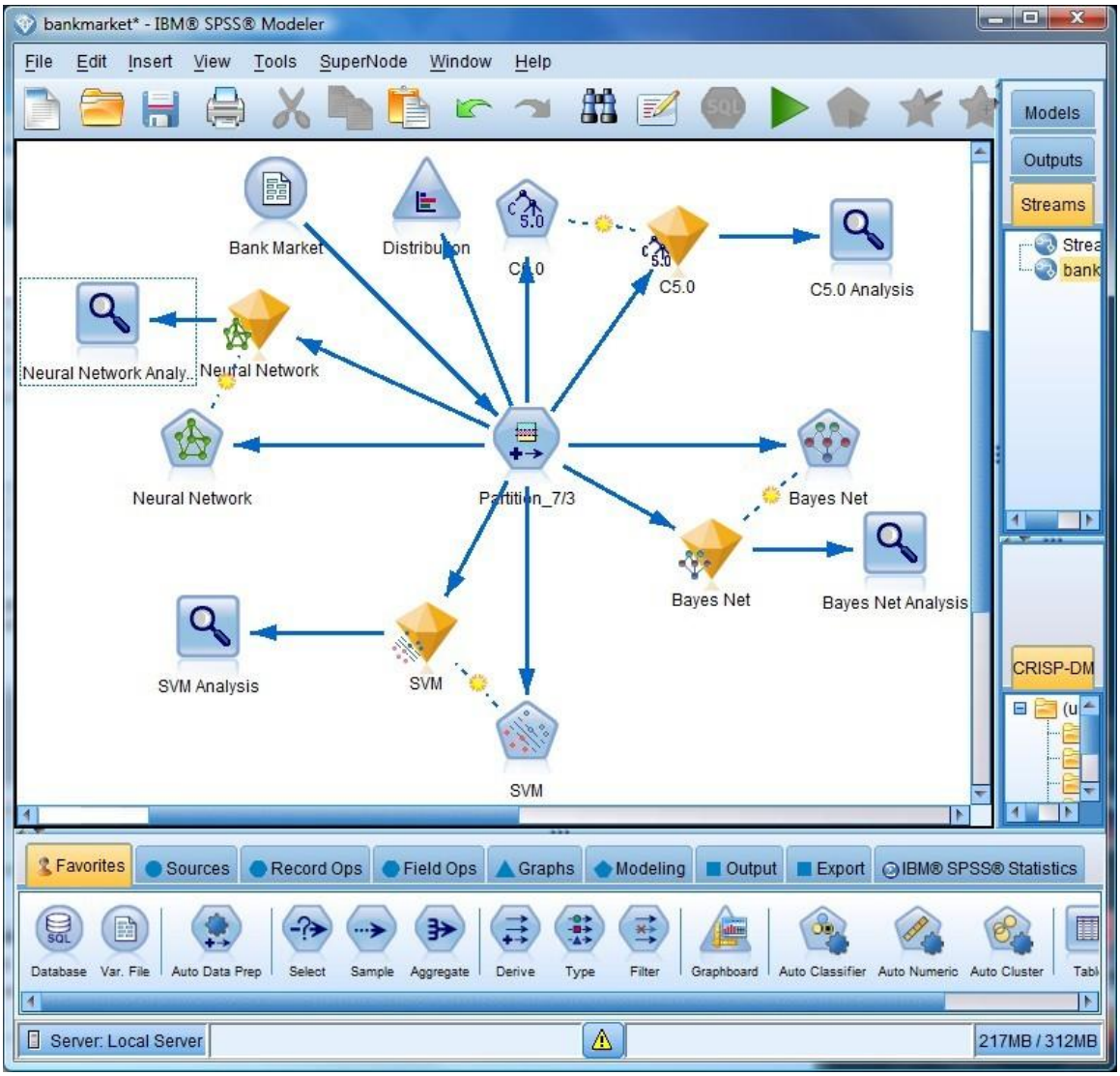


Figure 4.11: Data mining desing model of Bank Market data set in IBM SPSS Modeler

After the calculation, we can find the results as below tables.

C5.0 Analysis

Table 4.5: Table of accuracy of C5.0 Tree

Partition	1_Training		2_Testing	
Correct	2855	90.24%	1212	89.31%
Wrong	309	9.76%	145	10.69%
Total	3164		1357	

Bayes Net Analysis

Table 4.6: Table of accuracy of Bayes Net

Partition	1_Training		2_Testing	
Correct	2923	92.4%	1182	87.13%
Wrong	240	7.6%	175	12.87%
Total	3164		1357	

SVM Analysis

Table 4.7: Table of accuracy of SVM

Partition	1_Training		2_Testing	
Correct	3150	99.55%	1158	85.31%
Wrong	14	0.45%	199	14.69%
Total	3164		1357	

Neural Network Analysis

Table 4.8: Table of accuracy of Neural Network

Partition	1_Training		2_Testing	
Correct	2854	90.21%	1202	88.58%
Wrong	310	9.79%	155	11.42%
Total	3164		1357	

So we can find the Rough Set Theory approach has the highest accuracy compared with other approaches.

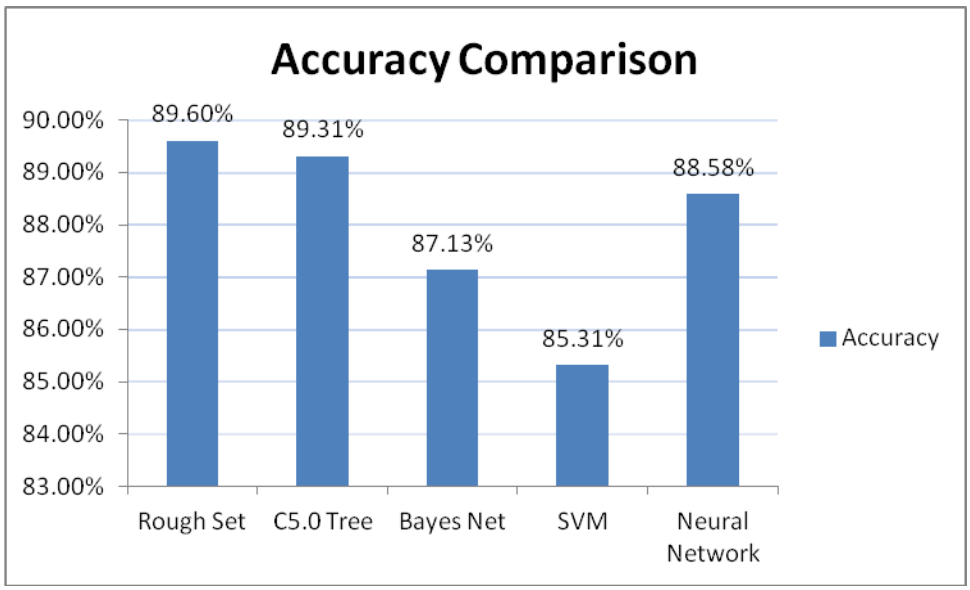


Figure 4.12: Chart of accuracy comparison of five methods

Chapter:5 Conclusion

In this time's research, we used the classification technique to classify the rules based on the decision attribute y (which represents whether a person will subscribe a term deposit).

From the experiment running result we may find Rough Set Theory approach has some advantage over other approaches by comparison.

In this kind of Multi-attributes and Multi-value data set data mining, Rough Set or Decision Tree approach may get a more accurate result but Rough Set Theory can maintain a high coverage of the testing data in the same time. Which stand for its prediction has a higher practicality.

On the other hand, using rough sets is advantageous over black box-type classification schemes such as neural network. With the rough-set approach, the generated rules are visible and available for the user, making the extraction of more information from the rules possible. Since the strength of a rule can also be calculated, this makes the rough set a valuable tool for classification.

A guiding philosophy of rough set theory is to let the data material speak for itself. As such, very few assumptions are made about the data attributes. It needs only some notion of inequality defined on the data domain. In particular, rough set theory does not make assumptions about statistical distributions of the data, nor does it need external information such as membership function for fuzzy sets, weighted inputs for neural network or density function in statistical classifications.

The output of a rough set analysis is usually a collection of if-then rules. It is, arguably, as close to a model in natural language as one might expect to obtain and can be read and interpreted by

personnel without expertise in the actual model-induction technique. Construction of the knowledge base is commonly perceived as a major bottleneck in building rule-based expert system. As rough sets can be used to automatically induce if-then rules from empirical data, this offers the possibility to automate, at least in part, the knowledge-acquisition stage for developing such systems.

With the help of Rough Set Theory, business intelligence can generate more commercial opportunity with in less time based on a big data set. Just like the case study showed in this thesis research, prediction can be an easier way by data mining approach than statistical approach especially in the nature value data (with no discipline).

Chapter:6 Future research

90% accuracy is not high enough, we still facing 10% error rate when doing the prediction. It probably caused by the information entropy. So the future research will focus on the suitable and correct granularity of information's partition. Improve the accuracy and deal with more data is the main mission for the future.

Bibliography

1. Pang-Ning Tan, Michael Steinbach and Vipin Kumar. Introduction to Data Mining,2005
2. Kantardzic, Mehmed.Data Mining: Concepts, Models, Methods, and Algorithms,2002
3. David Hand, Heikki Mannila and Padhraic Smyth. Principles of Data Mining ISBN:026208290x 2001
4. Ming-Syan Chen, Jiawei Han, Philip S. Yu. Data Mining: An Overview from a Database Perspective, IEEE Transactions on Knowledge and Data Engineering, Volume 8, NO 6, 1996, 866-883
5. Zdzislaw Pawlak. Rough set approach to knowledge-based decision support, European Journal of Operational Research, Volume 99, Issue 1, 1997, 48-57
6. Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, Chengqi Zhang. Combined Mining: Discovering Informative Knowledge in Complex Data, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE, Volume:41, Issue:3,2011, 699-712
7. Apte,C. Data mining: an industrial research perspective, Computational Science & Engineering, IEEE Volume:4, Issue:2 1997, 6-9
8. Ataee,P. Mining the (data) bank, Potentials, IEEE, Volume:24, Issue:4,2005,40-42
9. Liu, B.; Grossman, R.; Yanhong Zhai. MiningWeb pages for data records, Intelligent Systems, IEEE, Volume:19, Issue:6, 2004, 49-55
10. Kin-Nam Lau; Chongyan Gao. A data mining approach to performance measurement in the banking industry, Services Systems and Services Management ,2005. Proceedings of ICSSSM, Volume:2, 1009-1012
11. Longbing Cao, Vladimir Gorodetsky, Pericles A.Mitkas. Agent Mining: The Synergy of Agents and Data Mining, Intelligent Systems, IEEE,Volume 24, Issue: 3,2009, 64-72

12. Xindong Wu, Xingquan Zhu. Mining With Noise Knowledge: Error-Aware Data Mining, Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions,Volume:38, Issue: 4, 2008, 917-932
13. Longbing Cao. Social Security and Social Welfare Data Mining: An Overview, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions Volume:42, Issue:6,2012, 837-853
14. Zhao Li Ping; Shu Qi Liang.Data Mining application in banking-customer relationship management, Computer Application and System Modeling(ICCASM),2010, Volume:6 124-126
15. Panzzani,M.J. Knowledge discovery from data? Intelligent Systems and their Applications, IEEE, Volume:15, Issue:2,2000, 10-12
16. Yongjiang Fu. Data Mining, Potentials, IEEE, Volume:16, Issue:4,1997, 18-20
17. Olaru,c; Wehenkel,L. Data mining, Computer Applications in Power, IEEE, Volume:12, Issue:3, 1999, 19-25
18. Liu,Gaojun; Zhu, Yan. Credit assessment of contractors: A rough set method, Tsinghua Science and Technology,Volume:11, Issue:4, 2006,357-362
19. Degang Chen; Suyun Zhao; Lei Zhang; Yongping Yang; Xiao Zhang. Sample Pair Selection for Attribute Reduction with Rough Set, Knowledge and Data Engineering, IEEE Transactions,Volume:24,Issue:11,2012,2080-2093
20. Tsang, E.C.C.; Degang Chen; Yeung, D.S.; Xi-Zhao Wang; Lee, J. Attribute Reduction Using Fuzzy Rough Sets, Fuzzy Systems, IEEE Transactions, Volume:16, Issue:5, 2008, 1130-1141

21. Mitra, S.; Mitra, M.; Chaudhuri, B.B. A Rough-Set-Based Inference Engine for ECG Classification, Instrumentation and Measurement, IEEE Transactions, Volume:55, Issue:6,2006,2198-2206
22. Lutfi Othman, M.; Aris, I.; Abdullah, S.M.; Ali, M.L.; Othman, M.R, Knowledge Discovery in Distance Relay Event Report: A Comparative Data-Mining Strategy of Rough Set Theory With Decision Tree, Power Delivery, IEEE Transactions, Volume:25, Issue:4,2010,2264-2287
23. Ikno Kim; Yu-Yi Chu; Watada, J.; Jui-Yu Wu; Pedrycz, W. A DNA-Based Algorithm for Minimizing Decision Rules: A Rough Sets Approach, NanoBioscience, IEEE Transactions, Volume:10, Issue:3,2011,139-151
24. Nowicki, R. On Combining Neuro-Fuzzy Architectures with the Rough Set Theory to Solve Classification Problems with Incomplete Data, Knowledge and Data Engineering, IEEE Transactions, Volume: 20 , Issue: 9,2008,1239-1253
25. Maji, P.; Paul, S. Rough Sets for Selection of Molecular Descriptors to Predict Biological Activity of Molecules, Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions, Volume:40, Issue:6, 2010,639-648
26. Huaxiong, Li; Xianzhong, Zhou; Bing, Huang. Method to determine α in rough set model based on connection degree, Systems Engineering and Electronics, Journal of Volume: 20 , Issue: 1,2009,98-105
27. Suyun Zhao; Tsang, E.C.C.; Degang Chen; Xizhao Wang. Building a Rule-Based Classifier—A Fuzzy-Rough Set Approach, Knowledge and Data Engineering, IEEE Transactions, Volume:22, Issue:5

28. Gwanggil Jeon; Donghyung Kim; Jechang Jeong. Rough sets attributes reduction based expert system in interlaced video sequences, Consumer Electronics, IEEE Transactions, Volume:52, Issue:4, 2006, 1348-1355
29. Chen-Fu Chien; Li-Fei Chen. Using Rough Set Theory to Recruit and Retain High-Potential Talents for Semiconductor Manufacturing, Semiconductor Manufacturing, IEEE Transactions, Volume:20, Issue: 4, 2007, 528-541
30. Kusiak, A. Rough set theory: a data mining tool for semiconductor manufacturing, Electronics Packaging Manufacturing, IEEE Transactions, Volume:24, Issue:1,2001, 44-50.
31. Ching-Lai Hor; Crossley, P.A.; Watson, S.J. Building Knowledge for Substation-Based Decision Support Using Rough Sets, Power Delivery, IEEE Transactions, Volume:22, Issue: 3, 2007,1372-1379
32. Vidhya.K.A, G.Aghila. Hybrid text mining model for document classification, Computer and Automation Engineering (ICCAE), Volume: 1,2010, 210-214
33. An Zeng; Dan Pan; Qi-Lun Zheng; Hong Peng. Knowledge acquisition based on rough set theory and principal component analysis, Intelligent Systems, IEEE, Volume:21, Issue:2, 2006,78-85
34. Wu, Q.E.; Tuo Wang; Yong Xuan Huang; Ji Sheng Li. Topology Theory on Rough Sets, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions, Volume:38, Issue: 1, 2008, 68-77
35. Fernandez-Riverola, F.; Diaz, F.; Corchado, J.M. Reducing the Memory Size of a Fuzzy Case-Based Reasoning System Applying Rough Set Techniques, Systems, Man, and

Cybernetics, Part C: Applications and Reviews, IEEE Transactions, Volume:37, Issue: 1,2007, 138-146

36. Peng Chen, Shuang Liu, Rough Set-based SVM Classifier for text Categorization, Natural Computation, ICNC'08 Fourth International Conference, Volume:2,2008, 153-157
37. Wen-Yau Liang. Apply Rough Set Theory into the Information Extraction The Application of the Clustering, INC, IMS and IDC,2009. NCM'09. Fifth International Joint Conference, 2009, 262-266
38. Qiang Li; Jian-Hua Li; Gong-Shen Liu; Sheng-Hong Li. A rough set-based hybrid feature selection method for topic-specific text filtering, Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference, Volume:3 2004, 1464-1468
39. Ilczuk, G.; Mlynarski, R.; Wakulicz-Deja, A.; Drzewiecka, D.; Kargul, W. Rough set techniques for medical diagnosis systems, Computers in Cardiology, 2005, 837-840
40. Wang Chang-long; Qi Yan-ming. Variable Precision Rough Set Weight Calculation Based on Web Text Classification, Wireless Communications, Networking and Mobile Computing, 2009. WiCom '09. 5th International Conference, 2009, 1-4
41. Yu Haiyan; Zhang Xia; Qiao Xiaodong; Zhang Yunliang. Attribute reduction in inconsistent decision tables based on discernible matrix, Intelligent Control and Automation (WCICA), 2010 8th World Congress, 2010, 2760-2763
42. Baowei Zhang; Na Wang. Research of Discernible Matrix-Based Algorithm for Attribute Value Reduction, Intelligent Computing and Cognitive Informatics (ICICCI), 2010 International Conference, 2010, 349-352

Appendix

Appendix A:

Data Processing and .tab file of the example 3.1

Data Processing Part

1, For attribute Local?, let Yes=1, and No=2

2, For attribute Marital, let Single=1, Married=2, Divorce=3

3, For attribute Income, let (value<=80K)=1, (value<=100K)=2 others=3

4, For attribute Own a car, let Y=1,and N=2

Prediction.tab file is as follow

TABLE Prediction

ATTRIBUTES 4

Local? numeric 0

Marital numeric 0

Income numeric 0

Car numeric 0

OBJECTS 12

1 1 3 2

2 2 2 1

2 1 1 2

1 2 3 1

2 3 2 2

2 2 1 2

1 3 3 1

1 1 1 2

1 2 1 1

2 3 2 2

2 1 3 1

1 1 2 2

Appendix B

Data Pre-processing in case study.

For attribute Balance:

Table 9: Attribute balance discretion

Numeric	Categorical
<0	Less_than_0
1-9	Less_than_10
10-99	Less_than_100
100-999	Less_than_1000
1000-4999	Less_than_5000
>=5000	More_than_5000

For attribute duration:

Table 10: Attribute duraion discretion

Numeric	Categorical
<90	Less_than_3_months
90-179	Less_than_6_months

180-364	Less_than_1_year
365-729	Less_than_2_years
730-1829	Less_than_5_years
>=1830	More_than_5_years

For attribute campaign:

Table 11: Attribute campaign discretion

Numeric	Categorical
1-10	Keep the number
>10	More_than_10

For attribute pdays:

Table 12: Attribution pdays discretion

Numeric	Categorical
<0	No_contracted
1-89	Less_than_3_months

90-179	Less_than_6_months
180-364	Less_than_1_year
>=365	More_than_1_year

For attribute previous:

Table 13: Attribution previous discretion

Numeric	Categorical
<1	No_contact
1-9	Keep the number
>=10	More_than_10

For attribute age:

Table 14: Attribution age discretion

Numeric	Categorical
<=21	<=21
22-25	<=25

26-35	<=35
36-50	<=50
51-65	<=65
>65	>65

Appendix C

The data file of Bank Market using in RSES

TABLE BANK

ATTRIBUTES 16

job symbolic

marital symbolic

education symbolic

default symbolic

housing symbolic

loan symbolic

contact symbolic

month symbolic

poutcome symbolic

balance symbolic

duration symbolic

campaign symbolic

pdays symbolic

previous symbolic

age symbolic

Y symbolic

OBJECTS 4521

unemployed married primary no no no cellularoct
unknown LESS_THAN_5000 LESS_THAN_3_MONTHS 1
NO_CONTRACTED NO_CONTACT <=35 no
services married secondary no yes yes cellularmay failure
LESS_THAN_5000 LESS_THAN_1_YEAR 1
LESS_THAN_1_YEAR 4 <=35 no
management single tertiary no yes no cellularapr failure
LESS_THAN_5000 LESS_THAN_1_YEAR 1
LESS_THAN_1_YEAR 1 <=35 no
management married tertiary no yes yes unknown jun
unknown LESS_THAN_5000 LESS_THAN_1_YEAR 4
NO_CONTRACTED NO_CONTACT <=35 no
blue-collar married secondary no yes no unknown may
unknown LESS_THAN_10 LESS_THAN_1_YEAR 1
NO_CONTRACTED NO_CONTACT <=65 no
……

Curriculum Vitae

Zhonghua Hu was born on April 2, 1983 in Shanghai, China. He got his bachelor degree in Mechanical Engineering on the summer of 2005 from Tongji University, Shanghai, China. He started to pursue his Master of Science degree in Industrial Manufacturing and System Engineering at University of Texas at El Paso from fall 2008. At UTEP, he worked as a research assistant at Industrial Systems Engineering Laboratory.

Permanent Address: 5600 Star View Dr, El Paso, Texas, 79912