

**Proposed Architectural Model for Optimal Transformation of
Decision Table & Decision Tree into Knowledge Base**

T06653



**Research Dissertation Submitted By,
Muhammad Shuaib Qureshi
(419-FBAS / MSCS / S08)**

MS (Computer Science)

Supervisor

Muhammad Imran Saeed
Assistant Professor,
Department of Computer Science,
International Islamic University, Islamabad.

Co-Supervisor

Syed Muhammad Saqlain
Assistant Professor,
Department of Computer Science,
International Islamic University, Islamabad.

**Department of Computer Science, Faculty of Basic & Applied Sciences,
International Islamic University, Islamabad.**



Accession No 7H6653

DATA ENTERED

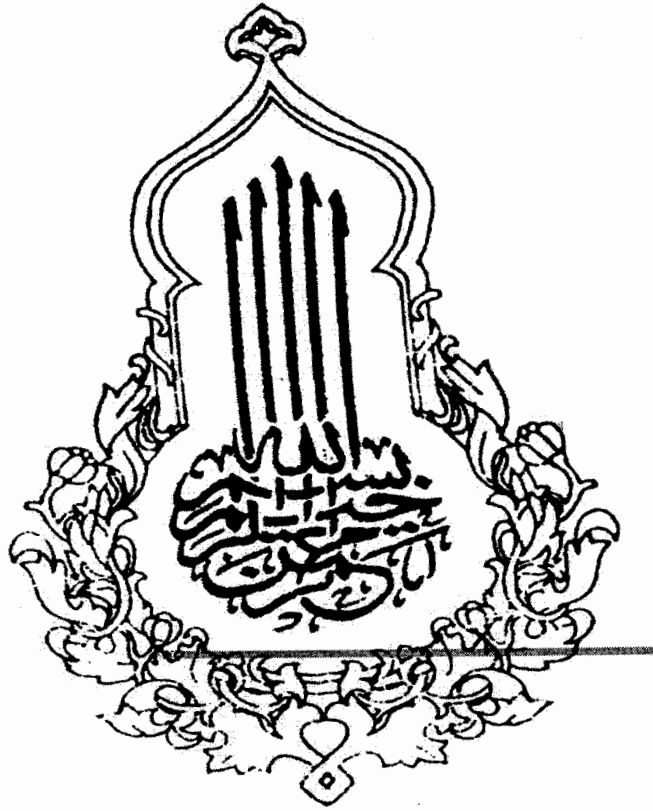
8/07/2010

to 40-introd.

Index
med

MS -
005.74.
QUP.

- 1- Data mining - Computer programs
- 2- Decision trees - " "



With the Name of

Allah,

*The most merciful and compassionate the most gracious and beneficent whose help and guidance we
always solicit at every step, and every moment.*

**A dissertation submitted to the
Department of Computer Science,
International Islamic University, Islamabad
as a partial fulfillment of the requirements
for the award of the degree of
Masters of Science in Computer Science.**

DEDICATION

*To my **parents** who are like cool shade in the noontide of my life, particularly to my **mother** whose hands get tired of praying for my success and to those who pray for me and encouraged me throughout my educational career.*

**Department of Computer Science
International Islamic University Islamabad**

Date: 07.05.2010

Final Approval

It is certified that we have examined the thesis report submitted by **Muhammad Shuaib Qureshi**, Reg No: 419-FBAS/MSCS/S08, and it is our judgment that this research project is of sufficient standard to warrant its acceptance by the International Islamic University, Islamabad for the Degree of Master of Science in Computer Science.

Committee:

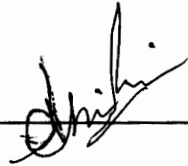
External Examiner

Dr. Nasro Min-Allah
HoD, Department of Computer Science
COMSATS Institute of Information Technology
Islamabad.



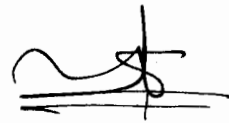
Internal Examiner

Mr. Asim Munir
Assistant Professor
Department of Computer Science
International Islamic University,
Islamabad.




Supervisor

Mr. Muhammad Imran Saeed
Assistant Professor
Department of Computer Science
International Islamic University,
Islamabad.



Co-Supervisor

Mr. Syed Muhammad Saqlain
Assistant Professor
Department of Computer Science
International Islamic University,
Islamabad.



ACKNOWLEDGMENT

I offer heartiest “DROOD-O-SALAM” to Holy Prophet MUHAMMAD (PEACE BE UPON HIM). I am grateful to almighty ALLAH who is merciful and beneficent, and who enable me to work on this research successfully. Accomplishment of a research thesis requires the help of many people who steer, guide, give confidence and help you. I have also been supported and guided by many people who were always there to help me out in the time of need. First of all I would like to express my sincere gratitude to my supervisor, *Mr. Muhammad Imran Saeed Qureshi* and my co-supervisor, *Mr. Syed Muhammad Saqlain* for their esteemed supervision, encouragement and guidance for successful completion of this thesis. Secondly I am really grateful to my brother *Mr. Muhammad Bilal Qureshi* for his sincere and unconditional help throughout my educational career. I am also thankful to my friends who always encouraged me to complete this research work. I am heartedly grateful to my parents for their gracious, unconditional support and encouragement throughout my study.

Muhammad Shuaib Qureshi

Declaration

I hereby declare that this thesis, neither as a whole nor as a part thereof has been copied out from any source. It is further declared that no portion of the work presented in this report has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning.

Muhammad Shuaib Qureshi

Abstract

Knowledge is one the most precious resource of an organization. Every organization wishes to preserve and fully utilize its knowledge. Within an organization knowledge is present in various forms, may be in the minds of workers or in documented form. In documented form the knowledge have various representation schemes such as frames, scripts, lists, decision trees and decision tables etc. The author of the thesis proposes a transformation method to transform the knowledge present in decision trees and decision tables to knowledge base. According to the author's proposal, the knowledge present in these two representation schemes should first be converted into corresponding set of human interpretable rules by using some existing transformation algorithms. For decision trees, the author proposes that it should be directly converted into set of human interpretable rules but for decision table the author proposes two ways that either it should be converted first into decision tree and then to set of rules, or it should be directly converted into set of rules. Once set of rules is obtained then it will be optimized by using existing optimization algorithms and unnecessary conditions in these rules will be omitted. After optimization of the rules it must be compared with existing rules present in the knowledge base. If the obtained optimized rules are not present in the knowledge base it should be added to the knowledge base. If some rules in the knowledge base need updation then after updating, the rules should be added to the knowledge base. During comparison, those rules should be omitted which already exist in the knowledge base.

Keywords: Data mining, Knowledge discovery, Decision Table, Decision Tree, Rules, Knowledge Base, Transformation.

Table of Contents

CHAPTER # 1	1
1. INTRODUCTION	2
1.1 Problem Statement	3
CHAPTER # 2	4
2. LITERATURE REVIEW	5
2.1 Data Mining	5
2.1.1 Reasons for using Data mining	6
2.1.2 Applications of Data Mining	7
2.1.3 Data Mining Project Steps	7
2.1.4 Methods of Data Mining	9
2.1.5 Data Mining Techniques	11
2.1.6 Limitations of Data Mining	12
2.1.7 Data Mining Issues	12
2.1.8 Research Challenges	14
2.2 Decision Table	14
2.3 Decision Tree	16
2.4 Advantages of Decision Table over Decision Tree	17
2.5 Advantages of Decision Tree over Decision Table	17
CHAPTER # 3	20
3. RESEARCH METHODOLOGY	21
3.1 Research Approach	21
CHAPTER # 4	24
4. PROPOSED FRAMEWORK	25
4.1 Advantages of the Proposed Framework	27
4.2 Working Flow of the Proposed Framework	27
4.3 Case Study	31

CHAPTER # 5	53
5. FUTURE WORK	54
REFERENCES	55

List of Figures

Figure No	Figure Description	Page No
Figure 2.1:	Decision Table	15
Figure 2.2:	Decision Table for book order	16
Figure 2.3:	Decision Tree	16
Figure 2.4:	Decision Tree for book order	17
Figure 3.1:	Research Approach	23
Figure 4.1:	Proposed Framework	26
Figure 4.2:	Decision Tree Algorithm	28
Figure 4.3:	Rules Covering Algorithm	29
Figure 4.4:	Ant Miner Algorithm	30
Figure 4.5:	Diabetes Data Set Results	51
Figure 4.6:	Sonar Data Set Results	51
Figure 4.7:	Zoo Data Set Results	52

List of Tables

Table No	Table Description	Page No
Table 2.1:	Data Mining Applications	7
Table 4.1:	Results of the proposed framework	50

List of Abbreviations and Symbols

- | | | |
|-----|-------|--|
| 1) | OLAP | Online Analytical Processing |
| 2) | KDD | Knowledge Discovery in Databases |
| 3) | NLP | Natural Language Processing |
| 4) | KDT | Knowledge Discovery in Textual databases |
| 5) | TDM | Text Data Mining |
| 6) | IE | Information Extraction |
| 7) | IM | Information Mining |
| 8) | IF | Intermediate Form |
| 9) | S & T | Science and Technology |
| 10) | N | any positive integer number |
| 11) | AI | Artificial Intelligence |

CHAPTER # 1

INTRODUCTION

1. INTRODUCTION

Knowledge is one the most precious resource of an organization. Every organization wishes to preserve and fully utilize its knowledge. Once the knowledge is acquired, it must be organized in an applications knowledge base for later use. The collection of knowledge related to a problem is organized and is called Knowledge base. Knowledge can be organized into different configuration to facilitate the inferencing from knowledge. There are different ways to organize knowledge base [18]. Many different knowledge representations, such as Semantics Nets, frames etc, have been proposed for use over the years. Within an organization knowledge is present in various form, may be in the minds of workers or in documented form. In documented form the knowledge have various representation schemes such as frames, scripts, lists, decision tree and decision tables etc. Every knowledge representation scheme has certain inherent strengths and weaknesses.

A decision table is composed of rows and columns. In problem specification, the corroboration and confirmation checking like contradiction, incongruity, redundancy etc is allowed by the decision table [21].

A decision tree is a tool using model or graph for taking different decisions [19]. Goals are represented by the nodes of the decision tree while as decisions are represented by its links. Conversion of decision tree into other formats can be done easily.

It is not necessary that decision tables, rules and decision trees in knowledge gaining or completion stage will be similar. So using of only one depiction during the knowledge development cycle is not needed. The obtaining of rule-based stipulation in the knowledge accomplishment stage is however possible while beginning from decision tables in the knowledge acquirement phase [4].

Knowledge usually changed from one shape to another suitable layout to get the response quicker and to decrease the quantity of computation. One knowledge depiction format is appropriate for one type of computation and other is appropriate for other type. Therefore mapping of knowledge from one depiction to another representation is needed. This mapping gives earlier response and decreases the computation amount. The decision table may be transformed into decision tree and decision tree might be transformed into set of human interpretable rules. Any client can recognize and alter a rule set without any difficulty than he/she can recognize and alter a decision table or decision tree.

This Thesis describes a transformation architectural model that transform decision table and decision tree into set of rules using some already existing sort of transformation functions or algorithms, and then optimized/refined these rules by using existing rules optimization algorithms to get the response faster. After getting the accurate, non redundant, optimized/refined rules, knowledge base is updated.

1.1 Problem Statement

Classification algorithm is a kind of important technology in Data Mining. At present there are a variety of classification modules. some of which are, Value Reduction, Decision Tree, Decision Tables, Neural Network, Statistical Model, Bayesian Classifier, etc.

Decision Tree, Decision Tables and Rules are commonly used methods [1]. The major issues when produce complete knowledge based application is, "how to apply the decision logic" [3]. In knowledge growth life cycle decision tables, decision trees and decision rules are used in different phases, starting from knowledge gaining till the completion phase [4].

Decision Tables and Decision Trees are optimized before its conversion into Rules, i.e. in earlier stages. *Construction of fully optimized Decision Tables and Decision Trees are NP-Hard problems.*

So "how to optimally transform decision table and decision tree into knowledge base and how to store the optimized knowledge in the knowledge base?", What strategy should be adopted?.

CHAPTER # 2

LITERATURE REVIEW

2. LITERATURE REVIEW

Literature review is particularly valuable for the understanding of a problem. Without literature review no one is able to realize, understand the problem and give appropriate solution to the problem. This chapter has two sections. First section gives detail review of data mining and the second section is about Classification Rules Mining (Decision Table, Decision Tree and Rules).

2.1 Data Mining

It does not give any benefits to an organization just to stock the data in data warehouse. Actually the use of that information to take some decision and obtaining some benefits gives value to the stored information. [2]

Bill Inman, the father of data warehousing define data ware housing as:

“A subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process” [7]

Data mining is one of the excellent ways to mine interesting patterns from a vast quantity of data. Data mining is of considerable importance in discovering of information from Data warehouse. [2]

According to R. J Brachman et al [8] data mining is:

“The process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions”

Data mining uses refine data analysis tools to find out up to that time unknown, appropriate samples and associations in huge data sets [9]. These tools consist of arithmetical models, machine learning methods and mathematical algorithms. Data mining is extra than gathering and management of data. It also consists of guess and analysis of data.

Multimedia and textual data can be mined through data mining. Most of data is in textual and multimedia forms. Data mining uses parameters of various types for analysis of data. It includes classification, forecasting, clustering, association and sequence of path selection. [10]

W. Frawley et.al (1992) defined data mining as:

“The nontrivial extraction of implicit, previously unknown, and potentially useful information from data” [11].

Another name given to data mining is Knowledge discovery in databases (KDD). Functionally, data mining discover knowledge from multiple data storages like databases or data warehouses. [12]

Data mining is related with investigation of data and exploit of software techniques for discovery of unseen unpredicted patterns and associations in data sets. The data mining main theme is to expose information that is unexpected [2].

2.1.1 Reasons for using Data mining

There are some reasons and factors, which shows the need and importance of data mining. Some of them are discussed as under. According to P.S.Bradley et.al (1998):

1. The analysis and discovery in OLAP is entirely done by a human analyst. Contrasting to OLAP, data mining permits the option of discovering data through computer. This option opens the opportunity to interact with databases in a new fashion [13].
2. Query formulation is another problem that has not gain consideration to a considerable extent in database research: if a client does not identify how to express objective in the form of a particular query then how to grant access to data [13].
3. One more problem, which can be improved through data mining, is the reality that it is extremely hard for humans to visualize and recognize a huge data. In data sets the growth of data is along two dimensions rows and columns. Rows show number of cases while columns show number of attributes. The human ability of visualization and analysis does not balance between high dimensions and a huge amount of data. The best way to deal high dimensionality is to transform it into a low dimensional space and then models of low dimensions should be made. A successful way to visualize data is the use of data mining algorithms to achieve suitable reductions in dimensionality. An alternative factor that revolves data mining into a prerequisite is that the growth rates of data sets go beyond the limit which can be handled through usual "manual" analysis techniques. Traditional data analysis system can't handle a situation where someone uses data on time bases. In fact it means that the majority of the data would remain unused [13].

2.1.2 Applications of Data Mining

Data mining is applied for gaining multiple aims in private as well as in public zone. Many industries like insurance, retailing / marketing, medicine and banking usually use data mining for cost reduction, research improvement and sales boosting. Table 2.1 shows few data mining applications.

Retail / Marketing
Identifying buying patterns of customers
Finding associations among customer demographic characteristics
Predicting response to mailing campaigns
Market basket analysis
Banking
Detecting patterns of fraudulent credit card use
Identifying loyal customers
Predicting customers likely to change their credit card affiliation
Determining credit card spending by customer groups
Insurance
Claims analysis
Predicting which customers will buy new policies
Medicine
Characterizing patient behavior to predict surgery visits
Identifying successful medical therapies for different illnesses

Table 2.1: Data mining applications [2]

2.1.3 Data Mining Project Steps

The steps of data mining are a procedure for mining buried knowledge from data storehouse, catalog or any other data file. The procedure or steps for data mining project is given in [14].

2.1.3.1 Objective Identification

Prior to start, it should be clear that what hope to be achieve from analysis. So first identify objectives of data mining. Find out whether the objective or goal is measurable or not. Some of the

goals are, identification of particular buying patterns over time, finding tendency of sales product and recognition of possible categories of customers.

2.1.3.2 Data Selection

Once objective is cleared, the subsequently stair is to choose data to accomplish the objective. May be this is a part of data mart or data warehouse that holds particular information. Divide the range of data to be mine, up to the possible extent. Some of the key points are sufficient data is there or not to illustrate the trend what the data mining analysis is going to model, are the data constant i.e. will the extracted aspects be consistent after analysis, while integrating databases, is a common field there for connecting them, are the data is up to date and related to objectives etc.

2.1.3.3 Data Preparation

After selection and collection of data, alteration of attributes to utilizable formats is of considerable importance. Some of the issues which may be under consideration are absent data handling, identification of redundant variables and decision about the exclusion of fields.

2.1.3.4 Data Auditing

Explain the formation of data in order to decide suitable tools. Some of the issues concerned under the topic are the makeup and nature of a database, the overall state of data set, allotment of data set, and the ratio of categorical (binary) attributes etc.

2.1.3.5 Tools Selection

For selection of appropriate data mining tool, two concerns are there, which are business goals and data structure. Both of them lead to the same tool. For assessment of potential tools some of the issues are platforms of candidate tool, heavily categorical or not, and data format etc. Some tools combine numerous tools in a group of neural network, arithmetical investigation programs and a figurative classifier.

2.1.3.6 Solution Format

Along with data audit, business goals and choice of tool decide the layout of resolution. The main inquiries are best possible format of the solution, on hand format choices, solution goal, and presentation of data to end users etc.

2.1.3.7 Model Construction

The data mining process actually begins from here. Generally the gateway is to apply an arbitrary figure to divide the data into a test set and training set, creates and assesses an architecture. Some of the matters are acceptable levels of error rates and its improvement, irrelevant attributes and their removal, need of additional data or methodology, training and testing of new data set.

2.1.3.8 Justification of Results

Propagate and converse about the outcomes of the investigation with end users and domain specialist. Make sure that results are correct and suitable to goals. Some points of discussion are the results seem sensible or not; returning to any previous steps is needed to get better outcomes.

2.1.3.9 Convey the Results

Present the ultimate statement to manufacturer or user. The whole data mining progression together with data training, source rules and code should be included in a report. Several problems are supplementary data ~~enhance scrutiny or not~~, what suggestions can be obtained from data mining analysis etc.

2.1.3.10 Solution Integration

Share conclusion with all concerned clients inside a particular business elements. The data-mining project is bringing to an end by integrating outcomes of analysis into company's business procedures.

2.1.4 Data Mining Methods

Data mining practices are composed of five practices [13]. These five methods are explained in this section. Numerous of these practices are not giving concentration to integrate data that is present in memory and database. They just only explain how it works over data present in memory. For huge databases handling these practices gives a foundation for scaling to work on it. For examples decision trees [15] included in classification, association rules [16] included in summarization and also in clustering [17].

2.1.4.1 Deviation Detection

Information sequence on the basis of time sequence or other ordering mechanism is explained through these methods. Observation sequence is of utmost important and their arrangement is essential. This should be explained in these types of methods. This is a unique aspect about methods of this group. For discovering those sequence in databases which are regular, these methods are scalable for it. Complexity is the worse case. To perform efficiently in transactional databases these methods give sparseness [18].

2.1.4.2 Dependency Modeling

Approaching to data is frequently achieved by obtaining some underlying structures within the data. Causality model may be deterministic or probabilistic. The example of deterministic in database is functional dependencies among attributes [19]. In general density estimation methods come under this category. Some explicit causal modeling is given in [20] and [21].

2.1.4.3 Data Summarization

To illustrate data subsets, compact patterns should be mined and sometime it is a main objective to gain. For presentation of data two methods are there which are horizontal (cases) and vertical (fields). Data is summarized in two ways. In horizontal (cases) summaries are generated. For finding relationships between attributes second method which is vertical summarization, is used. There is a big dissimilarity between the objectives of the two summarization techniques. Vertical summarization objective is finding a relationship among attributes. The prediction of an attribute is not considered in vertical summarization. Classification is the prediction of an attribute while clustering is combining cases into a single group. An association rule is a general technique used in data summarization [16]. These are rules which shows that what values of a group will come when other group values occurs. Market basket analysis is a common example of rules association.

2.1.4.4 Predictive Modeling

In database some attributes of fields can be predicted, because it depends on other attributes. Prediction is the basic objective of predictive modeling. It is called problem of regression if the value to be predicted is numeric [13]. It is a type of classification, if the value to be predicted is categorical means class. Regression and classification have a range of methods. To answer a broad range of problems, linear regression and nonlinear transformation on inputs is combined. Input space

transformation is naturally complex problem, which require problem knowledge and a little bit of skill. Feature mining is the name given to this kind of transformation in classification problems.

The primary objective in classification is to guess the mainly probable condition of a class. It is usually called problem of density estimation. If one guess probability of vector x , in class $C=c$, and other fields $X=x$, this is only taken from combined bulk of C and X . The guessing of the vector from the combined bulk is very tricky and not known. Some of the examples of these methods are density estimation for example kernel density estimators [22] and joint density graphical representation [21], K- nearest neighbor method [22], division of the attribute space method, where it is divided into decision regions and for each region a guess value is attached with it.

2.1.4.5 Clustering

Segmentation is another name given to clustering. It does not forecast any field but divides data items into sub groups on the basis of similarity. Number of preferred clusters is not known like classification. There are only two stages for algorithms of clustering. Clustering algorithms normally takes a bi-stage searching method, an external round is for feasible clusters and an inner round is used for finding the finest feasible collection for a known amount of clusters. There are three classes of clustering methods:

- Metric distance based
- Model based
- Partitioned based

2.1.5 Data Mining Techniques

- 1 Association Rule Mining
- 2 Cluster Analysis
- 3 Classification Rule Mining (Decision Tree, Decision Table & Rules)
- 4 Vector Support Machines
- 5 Deviation Detection/ Outlier Analysis
- 6 Genetic Algorithms
- 7 Neural networks
- 8 Rough Set Techniques
- 9 Logistic regression
- 10 Fuzzy methods

2.1.6 Limitations of Data Mining

Data mining products are not autonomous applications although it is powerful tools. For flourishing results data mining have need of analytically and technically expert professionals who can explore and infer final output. Thus data mining constraints are mainly related with personnel and data not with technology. Data mining does not notify worth or importance of patterns although patterns and relationships can be explored through data mining. Users should take decisions of these types. Discovered pattern validity is another factor dependent on a situation where it is compared. One more drawback of data mining is that causal relationship can not be discovered through data mining although variables association can be discovered. Its (causal relationship) example is like that airline tickets buying tendency is related with various factors like education, income, use of internet facility and many other factors are there. But authors can not say surely that ticket purchasing is affected just with one or more of the variables [23].

2.1.7 Data Mining Issues

For implementation of data mining and also for supervision, numerous issues should be handled on a first priority. Although here some issues are given but does not mean that data mining have only these issues. Among these issues some are creeping of mission, interoperability, privacy and data quality. Technology also affects data mining but the other factors are extremely important because it gives success or failure to data mining. The output of data mining is of great important. If the output of data mining is not accurate and efficient so using data mining will be just a time consuming job. Some of the data mining issues are given in [23].

2.1.7.1 Quality of data

Quality of data is a complex matter that embodies major challenges for data mining. Completeness and correctness are the prime factors required for data quality. The composition and uniformity of data also affect the quality of data. Data redundancy, lacking in data standards and mistakes which human beings are doing give considerable shock to data mining techniques efficiency. Cleaning of data is required on a time basis for enhancement in data quality. When data is cleaned, automatically results from that data will also be of excellent quality. Data cleaning may involve eliminating duplicate records.

Other factors which also enhance cleaning of data are extra data fields elimination, standardizing data formats, values standardization and consistent points of data.

2.1.7.2 Interoperability

Interoperability is somewhat relevant term with compatibility. In compatibility hardware or software is checking with already existing hardware and software that whether these are working together or not. Interoperability is checking of data or computer system functionality with other system and data. For interoperability some mechanism should be there to communicate. The mechanism will be processes and standards, which should be followed by both new and old systems. Interoperability of data mining is very important with already existing soft wares, tools and systems. Interoperability is very crucial issue for data mining.

2.1.7.3 Privacy

As the amount of sharing information and data mining projects increases, privacy needs higher priority. The privacy focus concerns about actual proposed projects and about applications of data mining. When privacy is considered in data mining, actual work is extended and data mining purpose is deviating from its original one and ultimately mission creeping occurs. Various researchers have various opinions on privacy compromise. Few are giving their arguments that privacy is on the first priority to guarantee security. Other researcher says that on hand rules and policies concerning confidentiality are enough and any risks to the privacy are not caused by these plans.

2.1.7.4 Mission Creep

Mission creep is one of the primary threats to data mining. It is really consumption of time and data other than its original purpose. This may occur apart from whether individuals granted data willingly or it was gathered through other ways. Information accessing for purposes except original one may appear to be a harmless activity. But the use of such information can cause unintentional effects and generate deceptive outcomes. Inexact data is one of the leading causes for deceptive outcomes. If data is inaccurate then automatically it will lead towards unbelievable results. All attempts of data collection may go through accuracy concerns to some extent. Expensive protocols can be required to ensure information accuracy.

2.1.8 Research Challenges

There are alarming challenges for the advancement of this field. A few of the challenges are given in [13]. Majority of methods suppose that data can fit in main memory. Actually data is located mainly on a server or on a desk. It may not fit in main memory and a substitution exists among accuracy and performance. So the need of developing mining algorithms for dependency analysis, clustering, summarization, classification deviation detection is there, that extend to large databases. For mining non homogeneous data i.e. multimedia, need of some mining methods is there. Those not only mine the non-homogenous data but also handle the sparse relations among data. For successful operation of mining algorithms, metadata (data about data) encoding methods are needed, In order to put more data effectively from user to the KDD system. Such methods of data mining are needed that justify previous knowledge of data. For search reduction using such knowledge may give certainty and remove data missing difficulties. Best and effective methods for sampling, dimensionality and data reduction are required for a combination of categorical (class) and numeric data fields. To mine composite relationships between data fields, novel search and mining algorithms are required [13].

Some methods or techniques are required, to represent change and expansion in data. The growth of data is not constant. So for better detection of data growth, some tools may need to handle it (data growth) [24].

Knowledge is one the most precious resource of an organization. Every organization wishes to preserve and fully utilize its knowledge. Once the knowledge is acquired, it must be organized in an applications knowledge base for later use. The collection of knowledge related to a problem is organized and is called Knowledge base. Knowledge can be organized into different configuration to facilitate the inferencing from knowledge. There are different ways to organize knowledge base [18]. Many different knowledge representations, such as Semantics Nets, frames etc, have been proposed for use over the years. Within an organization, knowledge is present in various forms, may be in the minds of workers or in documented form. In documented form the knowledge have various representation schemes such as frames, scripts, lists, decision tree and decision tables etc. Every knowledge representation scheme has certain inherent strengths and weaknesses.

2.2 DECISION TABLE

The decision table is a table with four quadrants consisting of logical events and situation. Decision table is shown in the following figure 2.1.

Condition Stub	Condition Entry
Action Stub	Action Entry

Figure 2.1: Decision Table [20]

All the essential tests are hold by the **Condition Stub**. All the procedures are contained by the **Action Stub**. The catalog of no/yes combinations is hold by the **Condition Entry**. The **Action Entry** is held by the lower right portion. Dots or X's denotes an action that should be taken. Action is specified by X while as no action is denoted by dots. Figure 2.2 shows the example of "Book order".

If order is from book store
 And if order is for 6 copies
 Then discount is 25%
 Else (if order is for less then 6 copies)
 No discount is allowed
 Else (if order is from libraries)
 If order is for 50 copies or more
 Then discount is 15%
 Else if order is for 20 to 49 copies
 Then discount is 10%
 Else if order is for 6 to 19 copies
 Then discount is 5%
 Else (order is for less then 6 copies)
 No discount is allowed

Condition Stub		Condition Entry					
		1	2	3	4	5	6
IF	Customer is bookstore	Y	Y	N	N	N	N
	Order size is 6 or more	Y	N	N	N	N	N
	Customer is library	N	N	Y	Y	Y	Y
	Order size is 50 or more	N	N	Y	N	N	N
	Order size is 20-49	N	N	N	Y	N	N
	Order size is 6-19	N	N	N	N	Y	N
Then	Allow 25% discount	X
	Allow 15% discount	.	.	X	.	.	.
	Allow 10% discount	.	.	.	X	.	.
	Allow 5% discount	X	.
	No discount	.	X	.	.	.	X
Action Stub		Action Entry					

Figure 2.2: Decision Table for "Book order". [20]

2.3 DECISION TREE

Decision tree is a useful technique using architecture of decisions and its outcomes. In such tree nodes represent goal and links represent decision [19]. Conditions are defined as a sequence of tests from left to right in a decision tree. The model of decision tree is demonstrated by Figure 2.3.

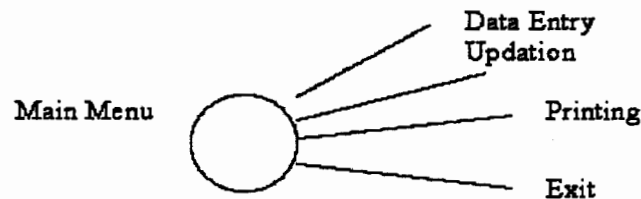


Figure 2.3: Decision Tree [20]

2.3.1 Decision table is transformed by a decision tree into a graph.

This model is stated from left to right. Each fork makes a decision and all forks produce an outcome. Decision tree for the "Book order" is represented by the following figure.

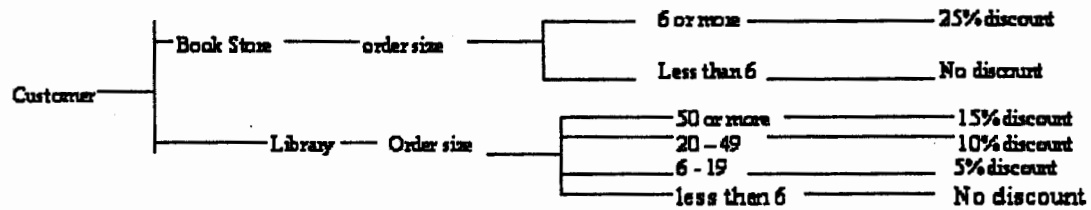


Figure 2.4: Decision Tree for Book Order [20].

2.3.2 Decision Tree transformation algorithms

In a decision tree a test on an attribute is denoted by each internal node. Conclusions of the tests are represented by its branches while as classes are represented by its leaf nodes. The root node is top most node of a tree. A very well known algorithm for decision tree construction is ID3 [33] [34]. Many enhancements to this algorithm have been done and incorporated into C4.5 algorithm [35]. For comparatively minute data sets ID3 and C4.5 algorithms have been soundly recognized. Scalability and efficiency become problems when these algorithms are applied on huge real-world databases because of the millions of training samples in the memory.

2.4 Advantages of decision table over decision tree

Decision Table can create more queries; it is more of multipath / multiflow. Decision Tree follows single path.

Gane and Sarson [22] proposed that in problem-solving phase when there are many mixtures of conditions then decision table is better than the decision tree.

2.5 Advantages of decision tree over decision table

The primary advantage of the decision tree is its branches chronological shape so that the array of examining conditions and executing actions is directly perceptible.

Next benefit is that decision trees actions and conditions are established on some branches but not on others which distinguishes decision tables in which they are all elements of the same table.

Another advantage of decision tree is that, any user can understand it easily. Gane and Sarson [22] claimed that simpler problems are better solved by decision tree.

Classification algorithm is a kind of important technology in Data Mining and the most commonly used technology is Decision Tree learning. The attribute selection of the traditional Decision Tree Algorithm is based on information Theory [1].

At present there are a variety of classification modules, some of which are,

1. Value Reduction
 2. Decision Tree
 3. Decision Tables
 4. Neural Network
 5. Statistical Model
 6. Bayesian Classifier
- etc.....

Decision Tree, Decision Tables are commonly used methods [1].

Decision Tree model can directly reflect the characteristics of data and has many features like,

1. Easy to understand
2. Good classification prediction ability and convenient for the rules extraction

Attribute reduction is core of rough set theory. *Due to the non uniqueness of attribute reduction, the minimum reduction of decision table is NP-hard problem* [1]. A latest technique for data mining and data scrutiny is Pawlak Rough Set Theory [2].

According to Rough Set Theory, construction of all rules from decision table is an NP-complete problem [2]. *The key problems are decision table reduction and NP-completeness.*

The major problem when producing a knowledge based function is, "*how to implement the decision logic*". For this, two alteration techniques are adopted [3],

1. Transforming decision table into tree.
2. Transforming decision table into set of rules.

Decision tables are checked and verified for correctness before its conversion to rules or trees. So it increases the complexity and table maintenance problem is faced [3].

In knowledge growth cycle decision tables, rules and trees are applied in different phases, starting from knowledge gaining upto the accomplishment phase [4]. Decision tables are transformed into either,

1. Decision tree
2. Number of Rules

Decision tables can be changed into a tree form. The resulted tree may be,

1. Balanced tree
2. Un-balanced tree

Un-balanced trees offer more flexibility, but more complex optimization challenges [4].

Confirmation and authentication of knowledge is accomplished in the earlier phases of the knowledge life cycle for avoiding acquiring incorrect knowledge instead of eliminating the fake knowledge in later phases [4].

Hence decision tables are verified and validated before its conversion to rules/ trees, i.e. in earlier stages, due to which *complexity is increased* [4].

In most cases, transformation of tables into tree is not suitable to deal with the knowledge in the knowledge base. In such cases conversion of decision table into rules should be preferred [4].

Reduction of knowledge is one of the important issues in the investigation of rough set theory [5][2]. *It is already proved that the optimal reduction of Decision Table is NP-hard problem.* Thus the people have been trying to search for more efficient heuristic algorithms to get an approximate reduction of decision table all the time [5].

Numerous phases of optimized tree production are NP-hard [32]. The NP-Completeness of constructing optimized decision trees from decision tables is proved by Hyafil and Rivest [30]. Cox showed the NP-completeness of discovering the optimized parent node [31].

The NP-completeness of building of storage optimal trees is proved by Murphy and McCraw, [27]. Naumov also proved the NP-hardness of the construction of optimized tree from decision tables [26]. The NP-Completeness of building optimized tree structured vector quantizers (TSVQ) is discussed by Lin. [28] [29].

Existence of redundant knowledge in knowledge base wastes store space and also prevent people from making decision correctly and concisely [4][5].

Decision tree model complexity and prediction accuracy determine the quality of decision tree. Find the optimal decision tree is an *NP-hard problem* [1].

CHAPTER # 3

RESEARCH METHODOLOGY

3 RESEARCH METHODOLOGY

Qualitative research grants a complete, competent and detailed investigation of research problem. For planning a well defined, brief and requirement oriented questionnaires, qualitative research provides as a basis for the quantitative research. Hence it serves as preliminary step for quantitative research. In other words quantitative research is dependent upon qualitative research [23]. Literature review is used to study research problem in qualitative research method.

For investigating the research problem in depth, different steps have been carried out. The first step is data collection where related data from the literature has been collected. The literature that is collected is comprised upon different research articles of well known conferences, journals and books in the field of Classification Models (Decision Trees, Decision Tables and Rules) and related areas.

The second step is about Decision Tree and Decision Table frameworks. Existing frameworks are exposed and got knowledge about their potential and flexibility.

In the third step data analysis is carried out, where data is scrutinized. From the data analysis, the ideas, concepts and materials related to the problem have been recognized. In the data analysis, data is precious because through data, thoughts are confirmed about what is going on? [23]. The data has been examined to give a synopsis of Knowledge Transformation frameworks, identifying components and interconnections. Finally a Knowledge Transformation framework is proposed.

3.1 Research Approach

The research approach which is adopted is a three phase process for the recognition and resolution of the research problem. The output of each phase is used as active drivers for obtaining the results in the next phase. Output 1 is the overview of exiting Decision Tree and Decision Table frameworks. Output2 consist of two branches: 2a and 2b which are intermediary outputs. In output 2a, components and steps of existing Decision Tree and Decision Table frameworks have been identified based on previous works from the literature. The identification of components has been further analyzed to obtain output 2b i.e. agreed components. After analyzing output 2b, final agreed components (output3) of proposed framework are obtained for designing. Finally proposed framework is designed. The approach used is shown in Figure 3.1.

Phase 1:

Data is gathered and a thorough literature review is made in the first phase. An overview of Decision Tree and Decision Table frameworks is given in this step (output 1) and identification of components and phases (output 2a) having all the components or steps addressed by different authors are acknowledged. These components or steps serve as a basis in phase 2 for obtaining an agreed component set (output 2b). The result of both 2a and 2b is combined to obtain the result 2, which obtains agreed key components.

Phase 2:

The phase 2 also consists of two parts. In the first part, agreed components (result 2b) are obtained by analyzing the components from result 2a. After analyzing output 2b, final agreed components (output3) of proposed framework are obtained for designing.

Phase 3:

An architecture describing steps, interconnections and key components is designed from output 2b (the agreed set of key components) and output3 (final agreed components). Finally output 4 is obtained which is proposed framework.

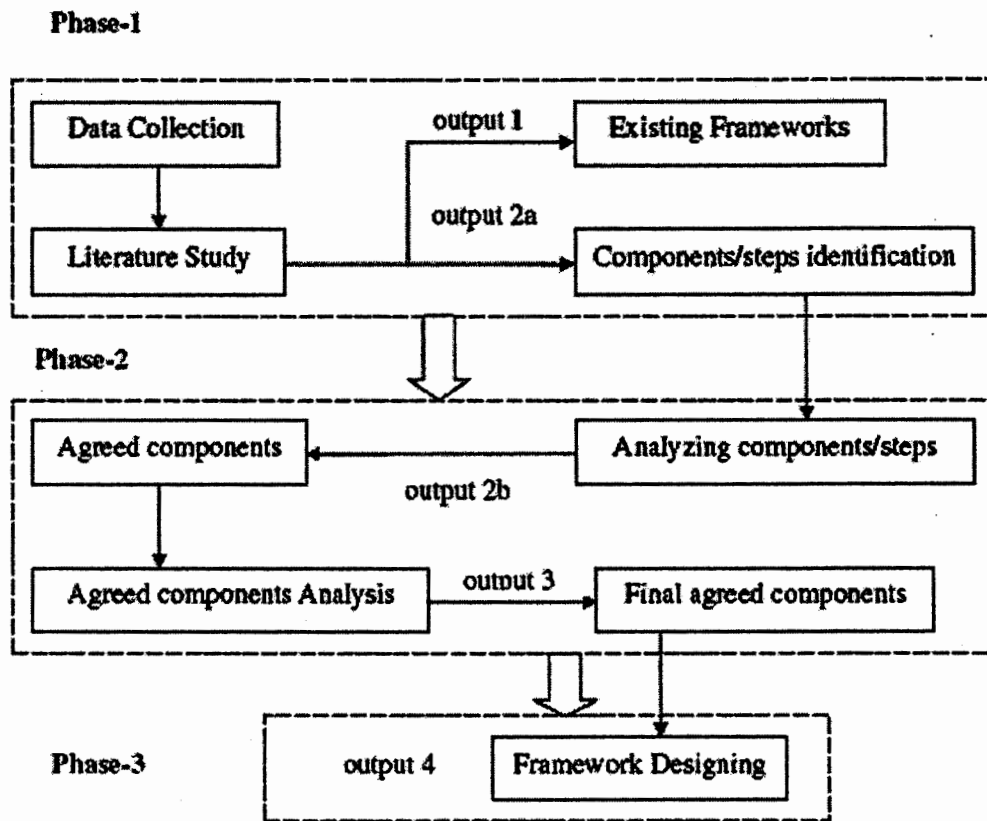


Figure 3.1: Research Approach

CHAPTER # 4

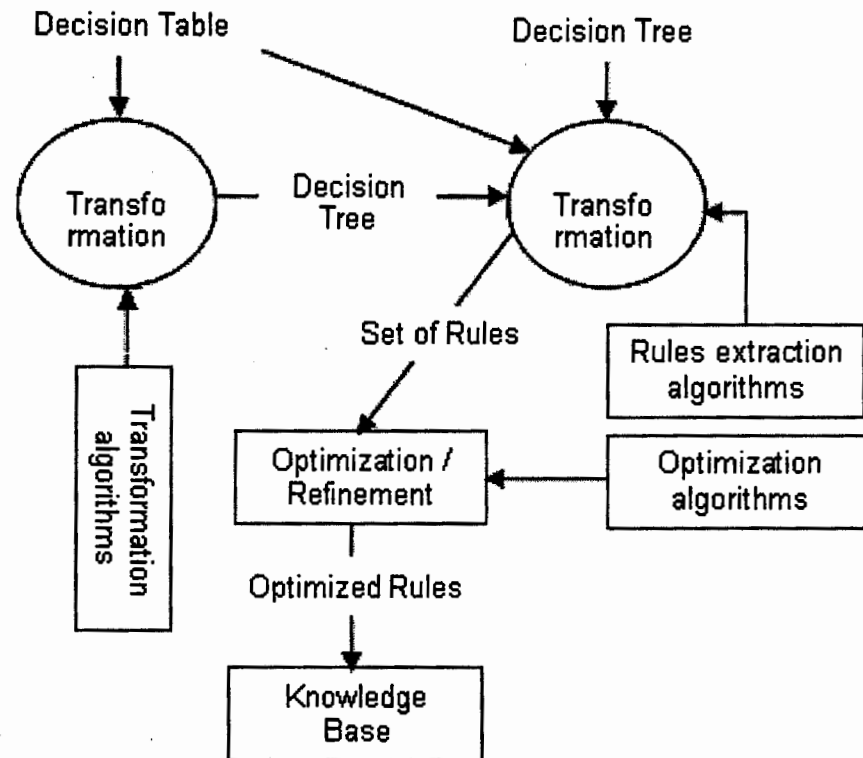
PROPOSED FRAMEWORK

4. PROPOSED FRAMEWORK

The decision tables, rules and decision trees in knowledge gaining or completion stage, however, do not inevitably remain the same. So there is no need to use one and only one representation throughout the entire knowledge development life cycle. For occasion, it might be possible to begin from decision tables in the knowledge acquirement process and obtain rule-based stipulation in the accomplishment stage [4].

Knowledge often transformed from one form to another appropriate format to get the response faster and to reduce the amount of computation. One knowledge representation scheme is suitable for one type of computation and other is suitable for other type. Thus there is a need to map knowledge from one representation to another. This mapping gives faster response and reduces the computation amount.

A new transformation method/framework is proposed for the transformation of knowledge present in decision trees and decision tables to knowledge base. According to the author's proposal, the knowledge present in these two representation schemes should first be converted into corresponding set of human interpretable rules by using some existing transformation algorithms. For decision tree, it is proposed that it should be directly converted into set of human interpretable rules but for decision table, two ways are proposed i.e. either it should be converted first into decision tree and then to set of rules, or it should be directly converted into set of rules. Once set of rules is obtained then it will be optimized (comprehensive rules without redundancy and with higher reliability) and unnecessary conditions in these rules will be omitted by using some existing optimization algorithms. After optimization of the rules, it must be compared with existing rules present in the knowledge base. If the obtained optimized rules are not present in the knowledge base, it should be added to the knowledge base. If some rules in the knowledge base need updation then after updating, the rules should be added to the knowledge base. During comparison those rules should be omitted which already exist in the knowledge base.

**Figure 4.1: Proposed Framework**

4.1 Advantages of the proposed Framework

- Reduce the amount of computation.
- Make the response faster.
- Remove the redundant knowledge.
- Optimization and accuracy is increased.
- Time and Space complexities are reduced.

1.2 Working Flow of Proposed Framework

1. Decision Tree is generated from the training data.
2. Decision Rules are generated from the Decision Table and Decision Tree.
3. Optimization of the generated Rules Set.
4. Storage of the Optimized Rules in the Knowledge Base.

For the whole process some of the existing algorithms are used, which are given below.

Algorithm: Generate_decision_tree. Generate a decision tree from the training tuples of data partition D .

Input:

- ⊛ Data partition, D , which is a set of training tuples and their associated class labels;
- ⊛ *attribute_list*, the set of candidate attributes;
- ⊛ *Attribute_selection_method*, a procedure to determine the splitting criterion that “best” partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and, possibly, either a *split point* or *splitting subset*.

Output: A decision tree.

Method:

- (1) create a node N ;
- (2) if tuples in D are all of the same class, C then
- (3) return N as a leaf node labeled with the class C ;
- (4) if *attribute_list* is empty then
- (5) return N as a leaf node labeled with the majority class in D ; // majority voting
- (6) apply *Attribute_selection_method*(D , *attribute_list*) to find the “best” *splitting_criterion*;
- (7) label node N with *splitting_criterion*;
- (8) if *splitting_attribute* is discrete-valued and
 multiway splits allowed then // not restricted to binary trees
- (9) *attribute_list* ← *attribute_list* – *splitting_attribute*; // remove *splitting_attribute*
- (10) for each outcome j of *splitting_criterion*
 // partition the tuples and grow subtrees for each partition
- (11) let D_j be the set of data tuples in D satisfying outcome j ; // a partition
- (12) if D_j is empty then
- (13) attach a leaf labeled with the majority class in D to node N ;
- (14) else attach the node returned by *Generate_decision_tree*(D_j , *attribute_list*) to node N ;
- endfor
- (15) return N ;

Figure 4.2: Decision Tree Algorithm [25]

Algorithm: Sequential covering. Learn a set of IF-THEN rules for classification.

Input:

- ⊗ *D*, a data set class-labeled tuples;
- ⊗ *Att_vals*, the set of all attributes and their possible values.

Output: A set of IF-THEN rules.

Method:

- (1) *Rule_set* = {}; // initial set of rules learned is empty
- (2) for each class *c* do
- (3) repeat
- (4) *Rule* = Learn_One_Rule(*D*, *Att_vals*, *c*);
- (5) remove tuples covered by *Rule* from *D*;
- (6) until terminating condition;
- (7) *Rule_set* = *Rule_set* + *Rule*; // add new rule to rule set
- (8) endfor
- (9) return *Rule_Set*;

Figure 4.3: Rules Covering Algorithm [25]

TH6653

Algorithm : High level pseudo-code of Ant-Miner.

```

begin Ant-Miner
  training_set ← all training examples;
  rule_list ← ∅;
  while |training_set| > max_uncovered_training_examples do
    τ ← initializes pheromones;
    rule_best ← ∅;
    i ← 1;
    repeat
      rulei ← CreateRule();
      ComputeConsequent(rulei);
      Prune(rulei);
      UpdatePheromones(τ, rulei);
      if Q(rulei) > Q(rule_best) then
        | rule_best ← rulei;
      end
      i ← i + 1;
    until i ≥ max_number_rules OR convergence ;
    rule_list ← rule_list ∪ rule_best;
    training_set ← training_set \ CorrectlyCoveredExamples(rule_best);
  end
end

```

Figure 4.4: Ant Miner Algorithm [43]

4.3 CASE STUDY

In this case study three examples of different data sets are taken. These data sets are passed through the proposed framework. It is shown that how optimized rules are generated and transformed into knowledge base.

4.3.1. Data Set: Diabetes

i. Decision Table Results:

```
Relation:      pima_diabetes
Instances:     768
Attributes:    9
Test mode:    10-fold cross-validation
```

=== Classifier model (full training set) ===

Decision Table:

```
Number of training instances: 768
Number of Rules : 32
Non matches covered by Majority class.
  Best first.
  Start set: no attributes
  Search direction: forward
  Stale search after 5 node expansions
  Total number of subsets evaluated: 43
  Merit of best subset found: 77.604
Evaluation (for feature selection): CV (leave one out)
Feature set: 1,2,4,9
```

Time taken to build model: 0.55 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	547	71.224 %
Incorrectly Classified Instances	221	28.776 %
Kappa statistic	0.3492	

Mean absolute error	0.3448
Root mean squared error	0.4277
Relative absolute error	75.8525 %
Root relative squared error	89.7294 %
Total Number of Instances	768

=== Confusion Matrix ===

```
  a  b  <-- classified as
405 95 |  a = tested_negative
126 142 |  b = tested_positive
```

Number of Rules: 32

Measure (Accuracy Measure): 71.224 %

ii. Decision Tree Results:

Relation: pima_diabetes
 Instances: 768
 Attributes: 9
 Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

```

-----
plas <= 127
  mass <= 26.4: tested_negative (132.0/3.0)
  mass > 26.4
    age <= 28: tested_negative (180.0/22.0)
    age > 28
      plas <= 99: tested_negative (55.0/10.0)
      plas > 99
        pedi <= 0.561: tested_negative (84.0/34.0)
        pedi > 0.561
          preg <= 6
            age <= 30: tested_positive (4.0)
            age > 30
              age <= 34: tested_negative (7.0/1.0)
              age > 34
                mass <= 33.1: tested_positive (6.0)
                mass > 33.1: tested_negative (4.0/1.0)
          preg > 6: tested_positive (13.0)
  plas > 127
    mass <= 29.9
      plas <= 145: tested_negative (41.0/6.0)
      plas > 145
        age <= 25: tested_negative (4.0)
        age > 25
          age <= 61
            mass <= 27.1: tested_positive (12.0/1.0)
            mass > 27.1
              pres <= 82
                pedi <= 0.396: tested_positive (8.0/1.0)
                pedi > 0.396: tested_negative (3.0)
              pres > 82: tested_negative (4.0)
          age > 61: tested_negative (4.0)
    mass > 29.9
      plas <= 157
        pres <= 61: tested_positive (15.0/1.0)
        pres > 61
          age <= 30: tested_negative (40.0/13.0)
          age > 30: tested_positive (60.0/17.0)
      plas > 157: tested_positive (92.0/12.0)

```

Number of Leaves : 20

Size of the tree : 39

Time taken to build model: 0.15 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	567	73.8281 %
Incorrectly Classified Instances	201	26.1719 %
Kappa statistic	0.4164	
Mean absolute error	0.3158	
Root mean squared error	0.4463	
Relative absolute error	69.4841 %	
Root relative squared error	93.6293 %	
Total Number of Instances	768	

=== Confusion Matrix ===

```
  a  b  <-- classified as
407 93 |  a = tested_negative
108 160 |  b = tested_positive
```

Size of Tree: 39

Measure (Accuracy Measure): 73.8281 %

iii. Classification Rules Results:

Relation: pima_diabetes
 Instances: 768
 Attributes: 9
 Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Decision List

plas <= 127 AND
 mass <= 26.4 AND
 preg <= 7: tested_negative (117.0/1.0)

plas > 154 AND
 mass > 29.8: tested_positive (100.0/14.0)

plas <= 99 AND
 age <= 25 AND
 age <= 22: tested_negative (33.0)

age <= 28 AND
 skin > 0 AND
 skin <= 34 AND
 age > 22 AND
 preg <= 3 AND
 plas <= 127: tested_negative (61.0/7.0)

plas <= 99 AND
 insu <= 88 AND
 insu <= 18 AND
 skin <= 21: tested_negative (26.0/1.0)

age <= 24 AND
 skin > 0 AND
 mass <= 33.3: tested_negative (37.0)

pres <= 40 AND
 plas > 130: tested_positive (10.0)

plas <= 107 AND
 pedi <= 0.229 AND
 pres <= 80: tested_negative (23.0)

preg <= 6 AND
 plas <= 112 AND
 pres <= 88 AND
 age <= 35: tested_negative (44.0/8.0)

age > 61 AND
 preg > 4: tested_negative (11.0)

age <= 30 AND
 pres > 72 AND
 mass <= 42.8: tested_negative (41.0/7.0)

```

plas <= 89 AND
plas > 0: tested_negative (13.0/1.0)

: tested_positive (252.0/105.0)

```

Number of Rules : 13

Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	578	75.2604 %
Incorrectly Classified Instances	190	24.7396 %
Kappa statistic	0.439	
Mean absolute error	0.3101	
Root mean squared error	0.4149	
Relative absolute error	68.224 %	
Root relative squared error	87.0418 %	
Total Number of Instances	768	

=== Confusion Matrix ===

```

a b <-- classified as
422 78 | a = tested_negative
112 156 | b = tested_positive

```

Number of Rules: 13

Measure (Accuracy Measure): 75.2604 %

iv. Optimized Rules Results:

=== Discovered Rules ===

```
IF mass < 26.4 THEN tested_negative
IF plas < 128.0 AND mass < 45.5 THEN tested_negative
IF mass >= 30.0 AND pedi >= 0.435 THEN tested_positive
IF mass >= 34.4 THEN tested_positive
IF plas < 159.0 AND preg >= 4.0 THEN tested_negative
IF insu >= 96.0 THEN tested_positive
IF preg < 2.0 AND mass < 34.3 THEN tested_negative
IF <empty> THEN tested_positive
```

Number of Rules: 8

Measure (Accuracy Measure): 80.5194 %

4.3.2. Data Set: Sonar

i. Decision Table Results:

Relation: sonar
 Instances: 208
 Attributes: 61
 Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Decision Table:

Number of training instances: 208
 Number of Rules : 15

Non matches covered by Majority class.

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 510

Merit of best subset found: 81.25

Evaluation (for feature selection): CV (leave one out)

Feature set: 4,5,11,46,61

Time taken to build model: 1.49 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	144	69.2308 %
Incorrectly Classified Instances	64	30.7692 %
Kappa statistic	0.3802	
Mean absolute error	0.3617	
Root mean squared error	0.4452	
Relative absolute error	72.6665 %	
Root relative squared error	89.2369 %	

Total Number of Instances 208

=== Confusion Matrix ===

```
a b <-- classified as
63 34 | a = Rock
30 81 | b = Mine
```

Number of Rules: 15

Measure (Accuracy Measure): 69.2308 %

=== Summary ===

Correctly Classified Instances	148	71.1538 %
Incorrectly Classified Instances	60	28.8462 %
Kappa statistic	0.422	
Mean absolute error	0.2863	
Root mean squared error	0.5207	
Relative absolute error	57.5045 %	
Root relative squared error	104.3706 %	
Total Number of Instances	208	

=== Confusion Matrix ===

```
  a  b  <-- classified as
69 28 |  a = Rock
32 79 |  b = Mine
```

Size of the tree: 35

Measure (Accuracy Measure): 71.1538 %

iii. Classification Rules Results:

Relation: sonar
 Instances: 208
 Attributes: 61

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Decision List

```

-----
attribute_11 <= 0.197 AND
attribute_1 <= 0.0392 AND
attribute_4 <= 0.0539 AND
attribute_28 <= 0.9578 AND
attribute_27 > 0.2771: Rock (56.0)

attribute_45 <= 0.2611 AND
attribute_36 <= 0.5459 AND
attribute_51 > 0.0125 AND
attribute_25 > 0.5331: Mine (42.0)

attribute_45 > 0.2611: Mine (39.0/1.0)

attribute_36 > 0.4619 AND
attribute_56 <= 0.0117: Rock (21.0)

attribute_58 > 0.0031 AND
attribute_57 <= 0.0058: Mine (20.0)

attribute_59 <= 0.0139 AND
attribute_43 <= 0.2296: Rock (18.0/1.0)

attribute_33 <= 0.7262: Mine (10.0)

: Rock (2.0)

```

Number of Rules : 8

Time taken to build model: 0.37 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	167	80.2885 %
Incorrectly Classified Instances	41	19.7115 %
Kappa statistic	0.6053	
Mean absolute error	0.2045	

Root mean squared error	0.4416
Relative absolute error	41.0846 %
Root relative squared error	88.5028 %
Total Number of Instances	208

=== Confusion Matrix ===

a	b	<-- classified as
79	18	a = Rock
23	88	b = Mine

Number of Rules: 8

Measure (Accuracy Measure): 80.2885 %

iv. Optimized Rules Results:

=== Discovered Rules ===

IF attribute_11 < 0.1989 THEN Rock

IF attribute_27 >= 0.8189 THEN Mine

IF attribute_51 >= 0.0137 THEN Mine

IF attribute_60 < 0.0079 AND attribute_2 >= 0.0275 AND attribute_7 >= 0.0887 THEN
Rock

IF attribute_41 >= 0.2575 AND attribute_37 < 0.5025 THEN Mine

IF attribute_52 < 0.0163 THEN Rock

IF <empty> THEN Mine

Number of Rules: 7

Measure (Accuracy Measure): 0.8571428571428571 = 85.7142 %

4.3.3. Data Set: Zoo

i. Decision Table Results:

Relation: zoo
 Instances: 101
 Attributes: 18
 Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Decision Table:

Number of training instances: 101
 Non matches covered by Majority class.
 Best first.
 Start set: no attributes
 Search direction: forward
 Stale search after 5 node expansions
 Total number of subsets evaluated: 121
 Merit of best subset found: 93.069
 Evaluation (for feature selection): CV (leave one out)
 Feature set: 5,13,14,15,18

Time taken to build model: 0.18 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	87	86.1386 %
Incorrectly Classified Instances	14	13.8614 %
Kappa statistic	0.8127	
Mean absolute error	0.1302	
Root mean squared error	0.2142	
Relative absolute error	59.3758 %	
Root relative squared error	64.9211 %	
Total Number of Instances	101	

=== Confusion Matrix ===

a	b	c	d	e	f	g	<-- classified as
41	0	0	0	0	0	0	a = mammal
0	20	0	0	0	0	0	b = bird
3	0	0	1	0	0	1	c = reptile
0	0	0	13	0	0	0	d = fish
1	0	0	0	3	0	0	e = amphibian
0	0	0	0	0	8	0	f = insect
2	0	1	0	2	3	2	g = invertebrate

Number of Rules: 15

Measure (Accuracy Measure): 86.1386 %

ii. Decision Tree Results:

Relation: zoo
 Instances: 101
 Attributes: 18
 Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

```

feathers = false
|   milk = false
|   |   backbone = false
|   |   |   airborne = false
|   |   |   |   predator = false
|   |   |   |   |   legs <= 2: invertebrate (2.0)
|   |   |   |   |   legs > 2: insect (2.0)
|   |   |   |   |   predator = true: invertebrate (8.0)
|   |   |   |   |   airborne = true: insect (6.0)
|   |   |   |   |   backbone = true
|   |   |   |   |   fins = false
|   |   |   |   |   |   tail = false: amphibian (3.0)
|   |   |   |   |   |   tail = true: reptile (6.0/1.0)
|   |   |   |   |   |   fins = true: fish (13.0)
|   |   |   |   |   |   milk = true: mammal (41.0)
|   |   |   |   |   |   feathers = true: bird (20.0)

```

Number of Leaves : 9

Size of the tree : 17

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	93	92.0792 %
Incorrectly Classified Instances	8	7.9208 %
Kappa statistic	0.8955	
Mean absolute error	0.0225	
Root mean squared error	0.14	
Relative absolute error	10.2478 %	
Root relative squared error	42.4398 %	
Total Number of Instances	101	

=== Confusion Matrix ===

	a	b	c	d	e	f	g	<-- classified as
41	0	0	0	0	0	0	0	a = mammal
0	20	0	0	0	0	0	0	b = bird
0	0	3	1	0	1	0	0	c = reptile
0	0	0	13	0	0	0	0	d = fish
0	0	1	0	3	0	0	0	e = amphibian
0	0	0	0	0	5	3	0	f = insect
0	0	0	0	0	2	8	0	g = invertebrate

Size of the tree: 17

Measure (Accuracy Measure): 92.0792 %

iii. Classification Rules Results:

Relation: zoo
 Instances: 101
 Attributes: 18
 Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Decision List

```

-----
feathers = false AND
milk = true: mammal (41.0)

feathers = true: bird (20.0)

backbone = false AND
airborne = false AND
predator = true: invertebrate (8.0)

backbone = false AND
legs > 2: insect (8.0)

fins = true: fish (13.0)

backbone = true AND
tail = true: reptile (6.0/1.0)

aquatic = true: amphibian (3.0)

: invertebrate (2.0)

```

Number of Rules: 8

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	93	92.0792 %
Incorrectly Classified Instances	8	7.9208 %
Kappa statistic	0.8955	
Mean absolute error	0.0231	
Root mean squared error	0.1435	
Relative absolute error	10.5346 %	
Root relative squared error	43.4854 %	
Total Number of Instances	101	

=== Confusion Matrix ===

a	b	c	d	e	f	g	<-- classified as
41	0	0	0	0	0	0	a = mammal
0	20	0	0	0	0	0	b = bird
0	0	3	1	0	1	0	c = reptile
0	0	0	13	0	0	0	d = fish
0	0	1	0	3	0	0	e = amphibian
0	0	0	0	0	5	3	f = insect
0	0	0	0	0	2	8	g = invertebrate

Number of Rules: 8

Measure (Accuracy Measure): 92.0792 %

iv. Optimized Rules Results:

=== Discovered Rules ===

IF milk = true THEN mammal

IF breathes = true AND feathers = true THEN bird

IF fins = true THEN fish

IF tail = false AND legs >= 6.0 AND breathes = true THEN insect

IF backbone = false THEN invertebrate

IF tail = true THEN reptile

IF <empty> THEN amphibian

Number of Rules: 7

Measure (Accuracy Measure): 100 %

TOTAL RESULTS:

Technique	Diabetes Data Set		Sonar Data Set		Zoo Data Set	
	Number of Rules	Accuracy	Number of Rules	Accuracy	Number of Rules	Accuracy
Decision Table	32	71.224 %	15	69.2308 %	15	86.1386 %
Decision Tree	39 (size)	73.8281 %	35 (size)	71.1538 %	17 (size)	92.0792 %
Rules	13	75.2604 %	8	80.2885 %	8	92.0792 %
Optimized Rules	8	80.5194 %	7	85.7142 %	7	100 %

Table 4.1: Results of the Proposed Framework

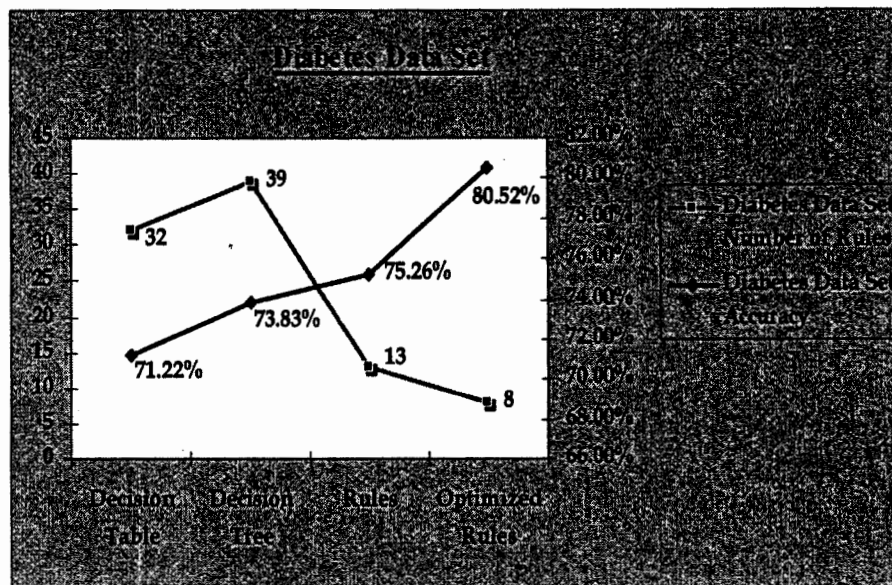


Figure 4.5: Diabetes Data Set Results

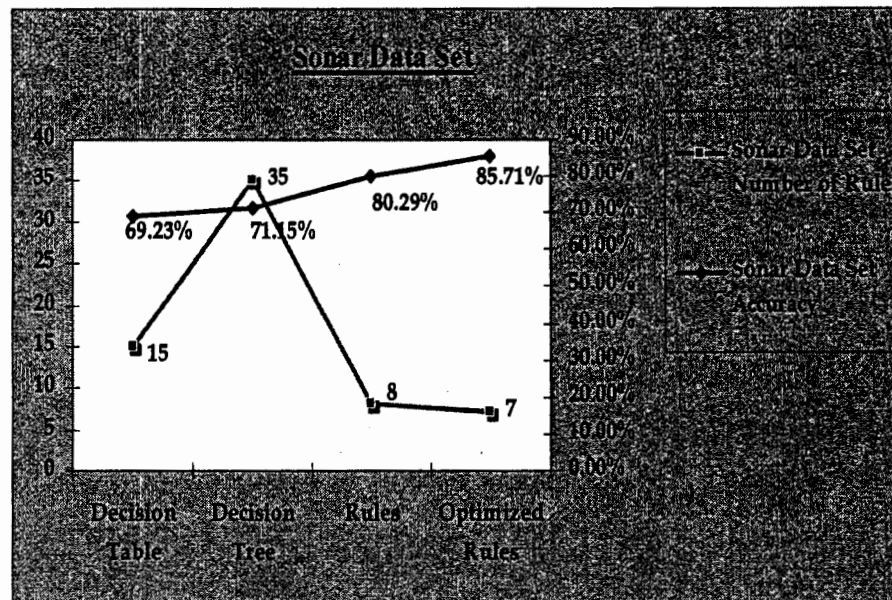


Figure 4.6: Sonar Data Set Results

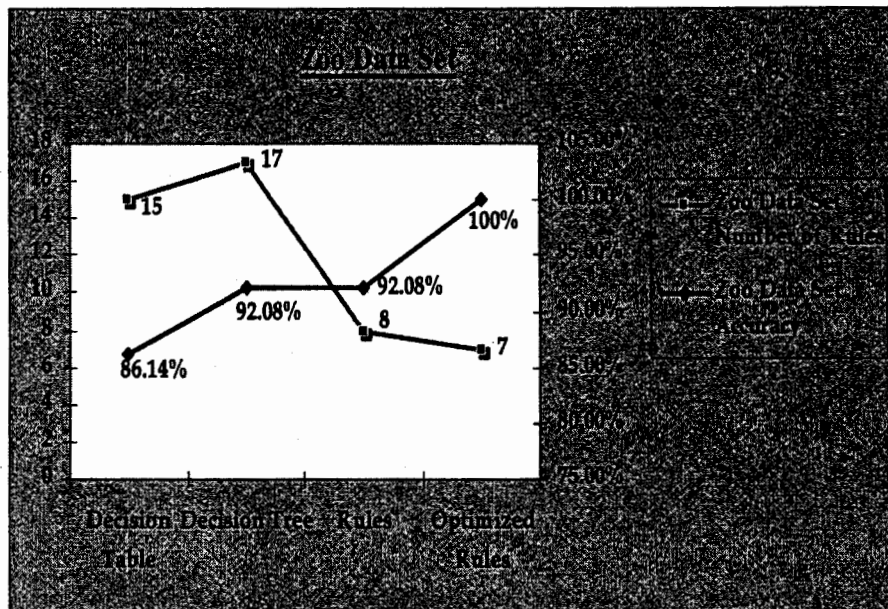


Figure 4.7: Zoo Data Set Results

CHAPTER # 5

FUTURE WORK

5 FUTURE WORK

This research work is qualitative in nature. In the future, it is decided to extend this work. It will be tried to implement the framework, if find suitable resources in terms of time, budget and human expertise.

Due to the time limitations and resource constraints, the implementation of such architecture has not yet been carried out. The combination of various classification techniques with the proposed architecture for better efficiency is also a choice for future work.

REFERENCES

- [1] Yan Li, Fa-Chao Li, Chen-Xia Jin, Tao Feng, A Rule Extraction Algorithm Based On Attribute Importance, Proceedings of the IEEE Eight International Conference on Machine Learning and Cybernetics, Boading, 12-15 July 2009: 127- 132.
- [2] Jun Zhou, Shu-You Li, Design of the Knowledge System based on Basic Rules, Proceedings of the IEEE Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007: 3704-3707.
- [3] J.Vanthienen, E.Dries, Restructuring and Simplifying Rule Bases, IEEE 1995: 484-485.
- [4] J.Vanthienen, G.Wets, Restructuring and Optimizing Knowledge Representations, : Proc.of IEEE Sixth International Conference on Tools with Artificial Intelligence, 6-9 Nov. 1994. 768 – 771.
- [5] Decai Huang, Lingli Wang, Analysis on the Drawbacks of the Commonly Used Measures of the Significance of Attributes in Decision Table and a New Measure , Proceedings of the First IEEE International Multi-Symposiums on Computer and Computational Sciences (IMSCCS), 2006.
- [6] Marcin Szyrka, Exclusion Rule-based Systems- Case Study, Proceedings of the IEEE International Multiconference on Computer Science and Informational Technology, 2008: 237-242.
- [7] Chien-I Lee, Cheng-Jung Tsai, Jhe-Hao Wu, Wei-Pang Yang, A Decision Tree-Based Approach to Mining the Rules of Concept Drift, Proceedings of the IEEE Fourth International Conference on Fuzzy Systems and Knowledge discovery (FSKD), 2007.
- [8] Hong-Zhen Zheng , Dian-Hui Chu, De-Chen Zhan, Rule Induction for Incomplete Information Systems, Proceedings of the Fourth IEEE International Conference on Machine Learning and Cybernetics,Guangzhou, 18-21 August 2005: 1864-1867.
- [9] Thomas Connolly and Carolyn Begg, P.E.L (2003), DATABASE SYSTEMS, 3rd Ed, Patparganj Dehli India: Pearson Education Limited.
- [10] Inmon, W. H. and Kelley, C. 1993 *Rdb/Vms: Developing the Data Warehouse*. John Wiley & Sons, Inc. New York, NY, USA, ISBN: 0-471-56920-8.
- [11] Brachman, R. J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., and Simoudis, E. 1996. Mining business databases. *Commun. ACM* 39, 11 (Nov. 1996), 42-48. DOI= <http://doi.acm.org/10.1145/240455.240468>
- [12] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third Edition* (Potomac, MD: Two Crows Corporation, 1999); Pieter Adriaans and Dolf Zantinge, *Data Mining* (New York: Addison Wesley, 1996).

-
- [13] SQL Server Definition >Data mining, TechTarget Windows Media, http://searchcrm.techtarget.com/gDefinition/0,294236,sid11_gci211901,00.html, 2008 Sept 10.
- [14] W. Frawley and G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in Databases: An Overview. AI Magazine, Fall 1992, pages 213-228.
- [15] Jiawei Han, Micheline Kamber. Data mining concepts and techniques. Morgan Kaufmann Publishers.
- [16] P.S.Bradley, Usama M. Fayyad, O.L. Mangasarian. "Data Mining: Overview and Optimization Opportunities". Journal of Computing, special issue on Data Mining. January 19, 1998.
- [17] J. Kettenring and D. Pregibon, editors. Statistics and Massive Data Sets, Report to the Committee on Applied and Theoretical Statistics, Washington, D.C., 1996. National Research Council.
- [18] Efraim turban, Jay Eronson, Edition. (2003) Decision support systems and intelligent systems.6th Ed .New Dehli: Prentice –Hall of India Private Limited.
- [19] Decision Tree, Wikipedia, http://en.wikipedia.org/wiki/Decision_tree , Aug 02, 2009.
- [20] Decision Tree and Decision Table, <http://www.ibm.com/>, 2009.
- [21] Nguyen, T.Perkins, W.Laffey, T.Pecora, Knowledge Base Verification, AI Magazine, 1987, pp. 69-75.
- [22] Girish H. Subramanian, John Nosek, Sankaran P. Raghunathan, Santosh S. Kanitkar, A Comparison of the OEI-ISION Table and Tree, Communications of the ACM, January 1992, Vol.35, No.1.
- [23] Uma Sekaram, "Research Methods for Business", 3rd Edition, United State of America, John Wiley and Sons Inc, 2000.
- [24] Meta Group Inc. Data Mining: Trends, Technology, and Implementation Imperatives.Stamford, CT.
- [25] Jiawei Han and Micheline Kamber, Chapter 7 of "Data Mining: Concepts and Techniques Morgan", 2nd Edition. Kaufmann Publishers, San Francisco, CA.
- [26] G. E. NAUMOV. NP-completeness of problems of construction of optimal decision trees. *Soviet Physics, Doklady*, 36(4):270--271, April 1991.
- [27] O. J. MURPHY AND R. L. MCCRAW. Designing storage efficient decision trees. IEEE Transactions on Computers, 40(3):315--319, March 1991.
- [28] JIANHIA LIN AND L.A. STORER. Design and performance of tree structured vector quantizers. Information Processing and Management, 30(6):851--862, 1994.
- [29] JIANHUA LIN, J. A. STORER, AND M. COHN. Optimal pruning for tree-structured vector quantizers. Information Processing and Management, 28(6):723--733, 1992.
- [30] HYAFIL AND RONALD L. RIVEST. Constructing optimal binary decision trees is NP-complete. Information Processing Letters, 5(1):15--17, 1976.

- [31] LOUIS ANTHONY COX, YUPING QIU, AND WARREN KUEHNER. Heuristic least-cost computation of discrete classification functions with uncertain argument values. *Annals of Operations Research*, 21(1):1-30, 1989.
- [32] MICHAEL R. GAREY AND D.S. JOHNSON. *Computers and Intractability: a Guide to the theory of NP-Completeness*. Freeman and Co., San Francisco, CA, 1979.
- [33] J. Quinlan, Induction of decision trees, *Machine Learning*, 1, pp. 81-106, 1986.
- [34] J. Quinlan, Simplifying decision trees, *International Journal of Man-Machine Studies*, 27, pp.221-234, 1987.
- [35] J. Quinlan, *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann, 1993.
- [36] VANTHIENEN, J, ROBBEN, F., Developing Legal Knowledge Based Systems Using Decision Tables, Fourth Int. Conference on A.I. and Law, Amsterdam, 1993, pp. 282-291.
- [37] VANTHIENEN, J., WETS, G., An Expert System Application Generator Based on Decision Table Modeling, Second World Congress on Expert Systems, Estoril, 1993.
- [38] WELLAND, R., *Decision Tables and Computer Programming*, Heyden & Son Ltd, Bury St. Edmunds, Great Britain, 1981, 203 pp.
- [39] VERHELST, M., The Conversion of Limited-Entry Decision Tables to Optimal and Near- Optimal Flowcharts : Two New Algorithms, *Communications of the ACM*, 15(11), Nov. 1972, pp. 974-980.
- [40] Junzhong Ji, Ning Zhang, Chunnian Liu An Ant Colony Optimization Algorithm for Learning Classification Rules, *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. 2006.
- [41] H. B. Duan. *Ant colony algorithm: Theory and Applications*, Science press in China, Beijing, 2005, 119-123.
- [42] Q. B. Zhu, Z. J. Yang. An Ant Colony Optimization algorithm based on mutation and dynamic pheromone updating. *Journal of Software*, 2004, 15 (2): 185-192.
- [43] B. Liu, H.A.Abbass, B. Mckay Classification rule discovery with ant colony optimization. In *Proceeding of the IEEE/WIC International Conference on Intelligent Agent Technology*, Beijing, China, 2003,83-88.
- [44] R. S. Parpinelli, H. S. Lopes, and A. A. Freitas. Data Mining with an Ant Colony Optimization Algorithm. *IEEE Trans on Evolutionary Computation*, special issue on Ant Colony Algorithms, 2002, 6(4): 321-332.
- [45] Z. Q. Wang and B.Q. Feng. Classification Rule Mining with an Improved Ant Colony Algorithm, G.I. Webb and X. Yu (Eds.): *AI 2004, LNAI 3339*, 2004, 357-367.

