

**A FUZZY ASSOCIATION RULE MINING EXPERT-DRIVEN
APPROACH TO KNOWLEDGE ACQUISITION**

BY

OLADIPUPO, OLUFUNKE OYEJOKE
(CUGP040086)

*B.Sc. (Hons) Computer Science (University of Ilorin, Ilorin),
M.Sc. Computer Science (Obafemi Awolowo University, Ile-Ife)*

**A THESIS SUBMITTED TO THE DEPARTMENT OF COMPUTER AND
INFORMATION SCIENCES, COLLEGE OF SCIENCE AND TECHNOLOGY,
COVENANT UNIVERSITY.**

**IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD
OF DOCTOR OF PHILOSOPHY DEGREE IN COMPUTER SCIENCE**

2012

CERTIFICATION

I hereby certify that this is an original research work carried out by Olufunke Oyejoke OLADIPUPO in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ogun State, Nigeria, under my supervision.

1. Name: Professor Charles Onuwa Uwadia
(Supervisor)

Signature _____ Date _____

2. Name: Prof. Charles Korede Ayo
(Co-Supervisor)

Signature _____ Date _____

3. Name: Prof. Charles Korede Ayo
(Head of Department)

Signature _____ Date _____

4. Name: Prof. A. Osofisan
(External Examiner)

Signature _____ Date _____

DECLARATION

It is hereby declared that this research was undertaken by Olufunke Oyejoke OLADIPUPO. The thesis is based on her original study in the Department of Computer and Information Sciences, College of Science and Technology, Covenant University, Ota, under the supervision of Prof. C.O. Uwadia and Prof. C.K. Ayo. Ideas and views of this work are products of the original research undertaken by Olufunke Oyejoke Oladipupo and the views of other researchers have been duly expressed and acknowledged.

Prof. C.O. Uwadia

(Supervisor)

Signature _____

Date _____

Prof. C.K. Ayo

(Co-Supervisor)

Signature _____

Date _____

DEDICATION

This work, to the glory of God, is dedicated to my loving, ever supporting,
indefatigable husband

ROTIMI OLANRELE OLADIPUPO

who sacrificed his time and pleasure for the fulfilment of this dream

and

my LATE mother, a rare GEM, *Abia-omo tooto*

MRS. OLUYEMISI DORCAS OJENIYI

whose desire was to see this dream fulfilled when she was alive.

ACKNOWLEDGEMENT

Foremost, I bless the name of the Lord Jesus Christ for His consistent mercies, goodness and generous endowments of grace towards me through my doctoral studies. He is indeed a covenant keeping God.

My earnest gratitude goes to my supervisors, Prof. C.O. Uwadia and Prof. C.K. Ayo, for their moral support, fatherly advice, and painstakingly reading through my thesis, despite their busy schedules.

I also want to acknowledge Prof. Ezekiel Adebisi, the former head of the department of Computer and Information Sciences, for his support and motivation. Big thanks to Prof. O.O. Olugbara, Dr. J.O. Daramola, all members of software engineering cluster and my colleagues in the department for their critical and useful contributions toward the success of this work. I will also like to acknowledge the Doctors in Covenant University Health Centre and Lagos State University Teaching Hospital (LASUTH) for their assistance in data description.

Special thanks go to my amiable and supporting pastor and his wife, Pastor & Pastor (Mrs) Andrew Folorunsho. They are always there for me. I will also like to say thanks to other pastors and their wives for their concerns: Pastor Blessing & Bunmi Imoroa, Pastor Jose & Helen Ighalo, Pastor Godwin & Toyin Usifo, A/P Emmanuel & Toyin Igeleko, A/P Olalekan & Pat. Olurotimi, A/P Akeem & Folake Oyedele. We are one big family.

Also worthy of appreciation are some friends, brothers and sisters that stood with me during the course of this project, Barrister & Dr (Mrs) Steve Folorunsho, Pastor and Pastor (Mrs) Tunde Oladimeji, Ibukun Afolabi, Oyelade O.J, Joy, Faith and Lizzy. You are indeed a plus to my life. Knowing you is a blessing.

Finally, my profound gratitude goes to my fatherly husband, I call him Daddy-Love; without his support and love this work would not have become a testimony. I thank God for His grace upon his life. My Father, Pa. Benjamin Abioye Ojeniyi, I thank God that he is alive to witness this success. Daddy, your desire has finally become a testimony. Papy, as your days so shall your strength be. My siblings and in-laws, I say thank you for your cares and concern. I am proud to belong to your happy family. You are more than Silver and Gold.

TABLE OF CONTENTS

Title	Page
Title Page	i
Certification	ii
Declaration	iii
Dedication	iv
Acknowledgement	v
Table of Contents	vii
List of Tables	xiv
List of Figures	xv
Abbreviations	xviii
Abstract	xix

CHAPTER ONE: INTRODUCTION

1.1	Background Information	1
1.2	Statement of the Problem	7
1.3	Aim and Objectives of the Study	8
1.4	Methodology	9
1.5	Significance of the Study	15
1.6	Motivation of the Study	16
1.7	Contribution to knowledge	18
1.8	Delimitation of the scope of the study	19
1.9	Thesis Organisation	19

CHAPTER TWO: LITERATURE REVIEW

2.1	Introduction	21
2.2	What is an Expert System?	21

2.2.1	Advantages and disadvantages of Expert Systems	23
2.2.1.1	Advantages of Expert Systems	23
2.2.1.2	Disadvantages of Expert Systems	23
2.2.2	Comparison of Expert System, with Conventional System and Human Expert	24
2.3	Approaches for Modeling an Expert System	25
2.3.1	The rule-based approach	26
2.3.2	The blackboard system approach	27
2.3.3	The frame-based approach	27
2.3.4.	The Open-based Expert System (OES) approach	28
2.3.5	The object-oriented approach	28
2.4	Rule-Based Expert System (RBES)	29
2.4.1	The RBES component	32
(a)	The user interface	32
(b)	The Database	32
(c)	Explanation facilities	33
(d)	Inference engine	34
(e)	The Knowledge-base	35
2.4.2	Rule-base Conflict Resolution	36
2.4.3	Conventional Programs vs Rule Based Systems	36
2.4.4.	Advantages of Rule Based Systems	37
2.4.5	Drawbacks of Rule Based Systems	38
2.4.6	Good Domains for Rule Based Systems	40
2.5	Overview of Fuzzy Theory	40
2.5.1	What is fuzzy Logic?	40

2.5.2	Observations about Fuzzy Logic?	41
2.5.3	A Fuzzy Set has Fuzzy Boundaries	42
2.5.4	Fuzzy set Operations	45
2.5.5	Membership functions	46
	(a) Triangular membership function (trimf)	47
	(b) Trapezoidal membership function (tramf)	48
	(c) Gaussian function (gaussmf)	49
	(d) Generalized Bell membership function (gbellmf)	49
	(e) Sigmoidal membership function (sigmf)	50
2.6.6	Linguistic Variables and Terms	52
2.6	Fuzzy Rule-based Expert system	54
2.6.1	Fuzzy Inference	54
	(a) Fuzzification	55
	(b) Rule Evaluation	56
	(c) Rule Aggregation	58
	(d) Defuzzification	68
2.7	Quantitative Measures of a Fuzzy Expert System	60
2.7.1	Accuracy measure	60
2.7.2	Comprehensibility Measure	61
	(a) Compactness:	61
	(b) Linguistic similarity:	62
	(c) Inconsistency of Fuzzy Rules:	63
2.7.3	Incompleteness Measure	64
2.8	Knowledge Acquisition and the building of Expert Systems	65
2.8.1	Knowledge Acquisition	71

(a)	Manual method	71
(b)	Semi-automated method	72
(c)	Automated knowledge acquisition	72
2.9	State-of-the-art in Medical Fuzzy Expert System Knowledge Acquisition Research	74
2.10	Data Mining in Knowledge Acquisition	75
2.10.1	CRISP-DM Model	76
2.10.2	Mining Quantitative Attributes with Association Rule Mining	79
2.10.3	Quality Measures	85
(a)	Fuzzy Support Value	85
(b)	Fuzzy Confidence Value	86
(c)	Interestingness Measure	87
2.11	Sharp Boundary Problem in Rule-Based Expert System in Medical Domain	88
2.12	The Context of this Research	90
2.13	SUMMARY	91

CHAPTER THREE: FUZZY ASSOCIATION RULE MINING EXPERT-DRIVEN (FARME-D) KNOWLEDGE ACQUISITION APPROACH

3.1	Introduction	93
3.2	Overview of the proposed solution: FARMED Approach	94
3.2.1	Limitation and Assumptions	94
3.3	Components of FARME-D Knowledge Acquisition	96
3.3.1	Historical Database	97
3.3.2	Domain expert	97

3.3.3	Expert-Driven Fuzzification process	97
3.3.4	Data Mining Engine	98
	(A) Data Pre-processing	98
	(i) Data Cleaning	99
	(ii) Data integration	99
	(iii) Data Transformation	100
	(iv) Data Reduction	100
	(B) Mining Process	100
3.3.5	Rule Interpretation/Knowledge Representation	102
3.4	Integration of FARME-D approach to standard Fuzzy Expert System architecture	102
3.4.1	Summary of how FARMES works	103
3.5	Tool Support for FARME-D	104
3.6	Application Scenarios	105
3.7	Validation Approach	106
3.8	FARME-D in Medical domain	106
3.9	Summary and Discussion	108
CHAPTER FOUR: PRACTICAL APPLICATION OF FARME-D IN MEDICAL DOMAIN		
4.1	Introduction	110
4.2	On Sharp Boundary Problem (SBP) in Medical Expert System	111
4.2.1	Investigation Process	111
	(a) Data Sets	111
	(b) Quantitative binary partition	112
	(c) Fuzzy Interval partition	115
	(d) Rule Generation	120

(e)	Quantitative Binary Expert System (QBES)	121
(f)	Fuzzy Expert System (FES)	122
(g)	Investigation result and Recommendation	123
4.3	FARME-D Process Life Cycle in Building a FES in Medical domain	126
4.3.1	The Prototype User Interface	126
4.3.2	Fuzzification Process	128
(i)	Domain Analysis	128
(ii)	Fuzzy model construction and Implementation	129
4.3.3	FARME-D approach	131
(i)	Historical database	132
(ii)	Rule Generation	132
(iii)	Rule Selection	133
4.3.4	Knowledge-base for Prototype System	136
4.3.5	Fuzzy Inference Process for Prototype System	137
4.3.6	Application Result and Discussion	138
4.4	Implementation Components and Tools	140
4.5	Summary and Discussion	141
CHAPTER FIVE: EVALUATION OF FARME-D APPROACH		
5.1	Introduction	142
5.2	Evaluation Overview	142
5.3	Motivation for Quantitative Measure of Evaluation	143
5.3.1	Quantitative evaluation	143
5.3.2	Discussion	144
5.4	Statistical Evaluation	148
(a)	t-Test	148

(b)	ANOVA Test	149
5.5	Possibilities for Generalization of Result	150
5.6	Summary and Conclusion	150
CHAPTER SIX: SUMMARY OF FINDINGS, CONCLUSION AND FUTURE WORK		
6.1	Summary	152
6.2	Conclusion	154
6.3	Future Work	155
REFERENCES		156
APPENDIX		
Appendix A.	ATP III Guidelines At-A-Glance quick Desk Reference	166

LIST OF TABLES

Title	Pages
Table 2.1 Apriori algorithm notation	84
Table 4.1 Linguistic variables and their fuzzy sets	115
Table 4.2 ATPIII, FES and QBES CHD risk value according to 2+ risk factor CHD for non-smoking men	123
Table 4.3 Instances showing the effect of SBP on medical expert system	124
Table 4.4 Description of Coronary Heart Disease determinant factors	128
Table 4.5 Confidence value against the number of rules	134
Table 4.6 The extracted rules from Mining system	135
Table 4.7 Non-Smoking men Test Case	139
Table 5.1 Tabular report of Quantitative evaluation	145
Table 5.2 t –Test result for ATP III & FARME-D	148
Table 5.3 ANOVA result for ATP III, FUMES & FES	149

LIST OF FIGURES

Title	Page
Figure 1.1	A model conceptualization of the methodology of this thesis 14
Figure 2.1	Complete structure of a rule-based expert system 31
Figure 2.2	Fuzzy logic allows overlapping of categories 44
Figure 2.3:	Non-fuzzy sets: with Sharp boundary problem 45
Figure 2.4:	Fuzzy sets: void of Sharp boundary problem 45
Figure 2.5	Triangular MFs 48
Figure 2.6	Trapezoidal membership function 49
Figure 2.7	<i>Gaussian</i> membership function 49
Figure 2.8	Generalized Bell membership function 50
Figure 2.9	Sigmoidal membership function 51
Figure 2.10	The slope of membership functions 52
Figure 2. 11:	Basic architecture of a fuzzy expert system 54
Figure 2.12	Structure of MFs within a universe of discourse for a crisp input 56
Figure 2.13	Mamdani-style rule evaluation 57
Figure 2.14	Clipped (a) and scaled(b) membership functions 58
Figure 2.15	Aggregation of rule consequents 58
Figure 2.16	Defuzzifying the solution variable's fuzzy set 59
Figure 2.17	The process of knowledge engineering 69
Figure 2.18	Extended process of knowledge engineering 70
Figure 2.19	The CRISP-DM Model 77
Figure 2.20	An algorithm for mining Fuzzy Association Rules 85
Figure 3.1.	The Structure of FARME-D knowledge acquisition approach 96
Figure 3.2	FARMES architecture 103

Figure 4. 1	Input and Output variable partitioning for (a) Age, (b) Cholesterol (c) HDL-C (d) Blood pressure (e) CHD % risk	113
Figure 4. 2	Binary partition for Age	113
Figure 4.3	Binary partition for Cholesterol	114
Figure 4. 4	Binary partition for HDL-C	114
Figure 4.5	Binary partition for Blood Pressure	114
Figure 4.6	Binary partition for %CHD risk	115
Figure 4.7	The membership function for Age	118
Figure 4.8	The membership function for Cholesterol	118
Figure 4.9	The membership function for (HDL-C)	119
Figure 4.10	The membership function for Blood pressure	119
Figure 4.11	The membership function for CHD %risk	119
Figure 4.12	The snapshot for standard rule-base formulation process	120
Figure 4.13	QBES CHD risk for the value Age=48, Cholesterol = 260, HDL-C=33, Bloodpressure = 120 with CHD risk % = 1.44	121
Figure 4.14	FES CHD risk for the value Age=48, Cholesterol = 260, HDL-C=33, Bloodpressure = 120 with CHD risk % = 10.9	123
Figure 4.15	ATPIII, FES and QBES CHD % risk value diagrammatic representation	125
Figure 4.16	The linguistic % CHD risk diagrammatic representation for ATP, FES and QBES	126
Figure 4.17	The snapshot of the main interface	127
Figure 4.18	The snapshot of the fuzzification process (crisp values)	130
Figure 4.19	The snapshot of the fuzzification process (fuzzified values)	130
Figure 4.20	The snapshot for FARME-D output	133

Figure 4.21	The number of rules against the confidence at a constant support of zero.	135
Figure 4.22	The snapshot of the rule-base on SQL server platform	136
Figure 4.23	The snapshot of inference process	138
Figure 5.1	Qualitative measures and their models	143
Figure 5.2	ATPIII, FES with 108 rules and FES with 79 rules CHD % risk value diagrammatic representation	144
Figure 5.3	The linguistic values for CHD risk diagrammatic representation for ATP, FES with 108 rules and FES with 79 rules	145
Figure 5.4	The snapshot FES (108 rules) with 16 rules fired.	147
Figure 5.5:	The snapshot FES (79 rules) with 8 rules fired.	147

ABBREVIATIONS

FES	Fuzzy Expert System
FARME-D	Fuzzy Association Rule Mining Expert-Driven
FARMES	Fuzzy Association Rule Mining Expert System
CHD	Coronary Heart Disease
ES	Expert System
ARM	Association Rule Mining
FARM	Fuzzy Association Rule Mining
KDD	Knowledge Discovery in Databases
MF	Membership Function
ANOVA	Analysis of variance
SBP	Sharp Boundary Problem
IS	Intelligent systems
AI	Artificial Intelligence
OES	Open-based Expert System
OO	Object oriented
RBES	Rule-based Expert System
QBES	Quantitative-based Expert System
WM	Working Memory
ANFIS	Adaptive Neuro-fuzzy Inference Systems
KE	Knowledge Engineering
CRISP-DM	Cross Industry Standard Process for Data Mining
ATP II	Adult Treatment Panel III
DBMS	Database Management Systems
IDE	Integrated Development Environment

ABSTRACT

This study tackles two concerns of knowledge engineers in designing and developing a fuzzy rule-based expert system (FES). First is to acquire a knowledge-base that emulates human perception of application domain concept in order to avoid sharp boundary problems. Second is the need for modelling a comprehensive fuzzy rule-based expert system which eliminates redundant rules in order to solve the problem of rule-base unwieldiness and provide for knowledge-base instant updates.

This thesis introduces an expert-driven knowledge discovery approach- Fuzzy Association Rule Mining Expert-Driven (FARME-D) approach to knowledge acquisition. In doing this, the Apriori-like Fuzzy Association Rule Mining algorithm was adopted for mining historical databases based on expert-driven approach (where the interval boundaries, fuzzy sets membership function model and fuzzy rules consequences are determined by the expert's opinion about the domain data). The fuzzy models were constructed using trapezoidal (trapmf) and triangular (trimf) membership functions based on the domain expert description of the database and literature. The implementation was done using C# programming language. The novelty of this approach was demonstrated by developing a prototype fuzzy expert system with mining generated rules using a case study of Coronary Heart Disease (CHD) as a cardiovascular disease in medical domain.

FARME-D approach generated 79 rules as against 108 rules by standard rule-base formulation approach. Using a test case approach of validation, it was observed that FARME-D approach saved 20% of memory size utilized by the knowledge-base and achieved 27 % rule deduction while the accuracy is maintained. The statistical

analysis of the result, using t-test and ANOVA revealed that decision making by FARME-D approach is significantly not different from the result by standard rule-base formulation and the domain expert at 95% confidence.

In conclusion, adopting FARME-D automated knowledge acquisition in modelling fuzzy expert system enhances the system comprehensibility by eliminating redundant rules and save memory usage. The rules generated based on expert-driven approach correspond to human perception of the application domain as compared to data-driven approach. Also, the integration of FARME-D approach to standard fuzzy expert system architecture provides for knowledge-base instant updates and resulted in a novel architecture called Fuzzy Association Rule Mining Expert System (FARMES). In future research, the mining process could be extended to involve text mining, image mining, voice mining and web mining in order to extend the scope of knowledge acquisition which will turn out to enrich the knowledge-base. Also, the knowledge representation could be extended beyond production rule to semantic net and case bases representations.

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND INFORMATION

The ultimate challenge of life is problem solving. Problem solving is the process of looking for a way out. In solving real-world problems, heuristic problem solving strategies and algorithmic strategies are not sufficient because of their limitations. The heuristic strategy is problem specific and could not absolutely guarantee the provision of the best solution. To this effect, an algorithmic problem solving strategy was introduced. Algorithms can be simply defined as straightforward procedures that are guaranteed to solve problems every time, for they are fully determinate and time invariant. However, many real-world problems especially in the medical domain cannot be reduced to algorithms, which lead to the invention of expert systems (Abraham, 2005).

An Expert System (ES) is an intelligent computer program that uses knowledge and inference procedures to solve problems that are difficult enough to require significant human expertise for their solutions (Feigenbaum, 1982). There are different approaches to modelling expert systems: the rule based approach, black-board system approach, the frame-based approach, the open-based expert architecture and object-oriented approach (Aly & Vrana, 2006). However, all of these approaches have their limitations.

The rule-based expert systems collect the small fragments of human knowledge into a knowledge-base in form of if-then rules, used to reason through a problem, by knowledge that is appropriate (Abraham, 2005). Rule-based expert systems are easy to

design; they emulate human cognitive processes and decision-making ability; and finally, they represent knowledge in a structured homogeneous and modular way (Ally & Vrana, 2006). Steps in the rule-based expert systems development process include determining the actual requirements, knowledge acquisition, constructing expert system components, implementing results, and formulating a procedure for maintenance and review (Abraham, 2001).

Knowledge acquisition being a crucial process in modelling expert systems is defined as the process of gathering the relevant information about a domain. The information gathering could be deductively from the human experts or inductively by learning from examples. Usually, the human thinking, reasoning, and perception processes cannot be expressed precisely, because the world of information is surrounded by uncertainty and imprecision. So, this type of human expert experiences can rarely be expressed or measured using statistical or probability theory. Therefore, fuzzy logic has provided a framework to model uncertainty, the human way of thinking, reasoning, and the perception process (Abraham, 2005). Fuzzy systems were first introduced by Zadeh (1965).

Fuzzy rule-based expert system (FES) is simply an expert system that uses collection of fuzzy membership functions and rules instead of the Boolean logic to reason about data in the inference mechanism (Neshat, & Yaghobi, 2009; Schneider et al., 1996). A fuzzy expert system consists of fuzzification process, inference mechanism, knowledge-base, and defuzzification subsystems. Fuzzy if-then rules and fuzzy reasoning are the backbone of fuzzy expert systems, which are the most important modelling tools based on fuzzy set theory. However, there are several limitations to

this system, which include large numbers of rules in the knowledge-base that causes the system to become unwieldy because of the presence of rules that might not be relevant to the application domain (Aly & Vrana, 2006). For instance, in standard rule-base formulation, the input space is divided into multidimensional portions and then actions are assigned to each of the portions.

The standard rule-base formulation is such that given M dimensions where each dimension is partitioned into N subspaces, there exist up to N^M rules in the fuzzy system (Meesad, 2001). The larger the N the larger the number of rules and, according to Meesad, if all the possible rules are used, then the system is not compact because of the redundant rules. These have three negative effects on expert system: 1) it increases the knowledge-base memory usage, since extra space is needed to store the redundant rules; 2) the existence of large number of rules reduces the rule access rate which ultimately slows down the response time of the ES; 3) it makes the knowledge-base unwieldy.

Nowadays, medical databases are growing in an increasingly rapid way with a big amount of quantitative attributes. Analyzing medical data is essential for medical decision making and management (Delgado et al., 2001). It has been widely recognized that analyzing medical data can lead to enhancement of health care by improving the performance of medical expert systems (Lavrac et al., 1996). According to Lavrac et al., (1996) there are two main aspects that define the need for medical data analysis 1) support of specific knowledge-based problem solving activities through the analysis of patients raw data collected from past experience, (2) discovery of new knowledge that can be extracted through the analysis of representative collections of example cases, described by symbolic or numeric descriptors. For these purposes, the

increase in database size makes traditional manual data analysis to be insufficient. Therefore, to fill this gap, knowledge discovery in databases (KDD), has proved sufficient. KDD is concerned with the efficient computer-aided acquisition of useful knowledge from large sets of data like medical database (Delgado et al., 2001

Association Rule Mining (ARM) is said to be one of the models for pattern discovery in the field of data mining (Agrawal, et al., 1993). Association rule mining is used to discover interesting relationships among items with categorical nature in a given database. The bottleneck of this technique is its inability to mine quantitative attributes directly. To achieve these, quantitative attributes have to be transformed into discrete intervals. This makes the mining process not void of sharp boundary problem, where boundary values are either overestimated or underestimated (Verlinde et al., 2006). Fuzzy logic has been proved sufficient for interpretability of discrete intervals (Delgado et al., 2003). Therefore, Fuzzy association rule mining (FARM) is an enhanced ARM technique that extracts interesting and hidden relationship from quantitative database. It relates the value of some attributes with values of some other attributes using fuzzy set concept to partition the attributes into different linguistic terms with membership value. According to Verlinder et al., (2006), the fuzzy interval partition/construction of membership function has been a problem in mining quantitative attribute. The two extreme solutions to this problem are expert-driven approach (an expert manually sets the interval boundaries and/or defines the membership functions) and the data-driven approach (they are generated automatically from the data table).The most common approach in the literature is data-driven approach. The membership functions obtained from data-driven approach may not correspond with the most intuitive human perception of concept. Hence, one may

expect rules obtained using a data-driven approach to be significantly different from the rules obtained using an expert-driven approach (Verlinder et al., 2006). FARM generates rules based on the linguistic term with support and confidence. The semantics of such rules are improved by introducing imprecise terms in both the antecedent and the consequent, as these terms are the most commonly used in human conversation and reasoning. The terms are modelled by means of fuzzy sets defined in the appropriate domains. The mining task is performed on the precise data. So, fuzzy association rules are more informative than rules relating precise values (Delgado et al, 2001).

With the advent of machine learning techniques, several of them have been introduced to knowledge acquisition in developing a fuzzy rule-based expert system in medical domain. This is done to enhance the comprehensibility of the expert system. These techniques include: clustering techniques (Shah et al., 2006), classification based data mining (Harleen & Siri, 2006; Gadaras & Mikhailov, 2009; Arias-Aranda et al., 2010; Ioannis & Ludmil, 2009), hybrid system of fuzzy and neural (Christoph, 1995; Moein et al., 2008), fuzzy evolutionary (Koutsojannis & Hatziygeroudis, 2006), neural network (Yan et al., 2006) and rough set theory (RST). The rules were selected and fuzzified based on information from discretization of numerical attribute (Setiawan et al., 2009). Adeli and Neshat, recently designed a fuzzy expert system for heart disease diagnosis (Adeli & Neshat, 2010).

An evolutionary fuzzy system was also presented (Shi et al., 1999). In most other cases, such as (Allahverdi et al., 2007; Saritas et al., 2003) rules were generated by conventional standard rule-base formulation. Also, fuzzy association rule mining based

on data-driven approach (where data partitions are generated automatically from the data table) was introduced to intrusion detection system (Norbik & Bharanidharan, 2005). What is yet to appear in the literature to the best of our knowledge is Fuzzy Association Rule Mining Expert-Driven approach (FARME-D, where domain expert's opinion is involved in calibrating fuzzy membership functions and determining the mined rules' consequences.) in medical domain (Verlinde et al., 2006). Some of the proposed approaches are not free from sharp boundary problem, rule inconsistency, membership function not corresponding with the intuitive human perception and more importantly having redundant rules in the knowledge-base. The aforementioned reviews pose critical challenges to development of medical expert systems which are (1) how to acquire a knowledge-base that will emulate human perception of medical concept and avoid the sharp boundary problem? (2) how to acquire a complete knowledge-base without redundant rules in order to solve the problem of rule-based expert system unwieldiness and allow for knowledge-base update?

In this thesis, we address these two concerns. Firstly, we investigated the effect of sharp boundary problem on medical expert system. Expert-driven approach for fuzzification process is adopted in tackling the sharp boundary problem and to acquire a knowledge-base that will emulate human perception of the domain problem. Direct interview with the experts in the application domain and literature are employed to determine the appropriate fuzzy models for the expert system determinant factors.

Secondly, we have adopted fuzzy association rule mining technique with incorporation of domain experts' opinion factors (the rules' consequences are determined by the domain experts' opinion) for automated knowledge acquisition in solving the

challenge of unwieldiness in rule-based expert systems. The expert-driven mining system is going to be integrated with the standard fuzzy expert system architecture so as to enhance the knowledge-base update in case of new invented instances by the domain expert. FARME-D has been used to mine the existing patient medical data in the application domain in order to extract useful interesting rules and hidden patterns from the database based on the domain experts' opinion. FARME-D is expected to: 1) minimize the number of rules in the knowledge-base by eliminating rules that are not relevant to the application domain, in order to solve the problem of knowledge-base unwieldiness, 2) generate rules that correspond intuitively with domain experts' perception of the data, 3) generate rules with support and confidence values which could be used to determine the frequent occurrences of each rule for rule rating.

1.2 STATEMENT OF THE PROBLEM

Knowledge acquisition has long been known as a bottleneck to modelling of an expert system in a variety of fields. The difficulty is especially great for medical knowledge-bases because medical fields present a combination of imprecise causal knowledge, very large amount of information, and potentially life-threatening consequences of incorrect conclusion (Fetical et al., 1989; Aly & Vrana, 2006; Delgado et al., 2001). Therefore, there is a need to generate knowledge that is void of sharp boundary problem, corresponding with the most intuitive human perception of concept in the domain, consistent and able to give accurate result (Oladipupo et al., 2010). The fundamental concerns of modelling a rule-based expert system are presence of large number of rules in the knowledge-base, which make the system to become unwieldy, and the difficulty in assigning confidence rating to each rule (Aly & Vrana, 2006). Hence, in order to enhance the comprehensibility of the rule-based expert systems,

reduce knowledge-base space complexity and increase rule access rate which in turn will increase system response time, there is need to tackle these challenges.

The research presented in this thesis is intended to address the challenges of sharp boundary problem, unwieldiness in fuzzy rule-based expert system, and knowledge-base update especially in medical domain. The research questions investigated in this thesis are:

- How do we acquire a knowledge-base that will emulate human perception of application domain concept and void of the sharp boundary problem? And
- How can an expert system developer develop a comprehensive fuzzy rule-based expert system which eliminates redundant rules in order to solve the problem of rule-base unwieldiness and provides for knowledge-base update?

1.3 AIM AND OBJECTIVES OF THE STUDY

The aim of this research work is to establish a fuzzy association rule mining expert-driven approach (where experts' opinion factors are incorporated into the mining process) to expert system knowledge acquisition. This will enhance rule-based system comprehensibility, make the rules to correspond with domain experts' perception of the data and encourage knowledge-base update. To achieve this aim, the following concrete objectives will be pursued:

- Creating a theoretical framework and design-oriented framework from which a comprehensible medical fuzzy expert system can evolve in cycle with the state-of-the-art practice in designing fuzzy rule-based expert systems.
- Investigating the effect of Sharp Boundary Problem (SBP) in medical rule-based expert systems.

- Extracting interesting knowledge in form of rules from application domain historic database using fuzzy association rule mining algorithm based on expert-driven approach.
- Demonstrating the potential of fuzzy association rule mining expert-driven approach in responding to the unwieldiness challenge of rule-based expert systems, by evolving a new automated knowledge acquisition approach for rule-based expert systems knowledge engineering; and
- Validating the credibility of the introduced automated knowledge acquisition approach by using a cardiovascular disease such as Coronary Heart Disease (CHD) as a case study.

1.4 METHODOLOGY

To achieve the aforementioned concrete objectives, we chose to investigate the KDD inductive knowledge acquisition method as a solution to modelling a comprehensible fuzzy rule-based expert system. This was meant to create a mining platform coupled with experts' opinion where hidden knowledge could be discovered from historical database to enhance the comprehensibility of the fuzzy rule-based expert system. In order to achieve this we analyzed the state-of-the art in building expert system through an extensive review of literature, study of existing expert systems, identification of stages in knowledge engineering and study of different approaches to knowledge acquisition. This finally resulted into the proposition of a new approach for knowledge acquisition component of a generic referenced Fuzzy Expert System architecture (FES). The proposed approach creates a potential platform for knowledge acquisition using mining technique coupled with experts' opinion factors.

In order to investigate the effect of sharp boundary problem in medical fuzzy expert system, two different medical expert systems were simulated for CHD risk determination using MatLabTM fuzzy logic toolbox with Mandani inference mechanism, MaxMin method and centroid defuzzification method. The first expert system was simulated based on quantitative binary partition, using distance-based partitioning method according to domain experts' opinion about the data description. Also, the second expert system was simulated based on fuzzy models which were constructed based on the domain experts' opinion about the data description. The rules were generated using standard rule-base formulation (Meesad, 2001) in conjunction with ATP III Guidelines for CHD risk ratio determination by National Cholesterol Education programme based on Framingham risk scoring. Trapezoidal (trmf) and Triangular membership function (trimf) were used for fuzzy partitioning while test case approach was used to determine the effect of the sharp boundary problem on medical expert systems.

Building on FES architecture, this thesis introduces a new approach of Fuzzy Association Rule Mining Expert Driven (FARME-D) approach as a data mining technique which incorporates experts' opinion factors for knowledge acquisition component of FES. This approach extracts knowledge inductively from past experiences. The FARME-D expert system development phases, which include data preprocessing, data transformation, mining process, and knowledge representation were systematically demonstrated to extract the interesting knowledge in form of rules from the medical domain historical database.

Data pre-processing phase includes Data cleaning in which noise and inconsistent data are removed from the historical database; data integration, where multiple data source are combined using Pearson's product moment confident; and data selection, where relevant data for the mining process are retrieved from the source data. All these were well carried out offline based on the established KDD methodology.

Also during data transformation process, the identified attributes, based on the mining requirement, were transformed into the form that is appropriate for mining. The main activity here is fuzzification process. Fuzzification process was based on domain experts' opinion about the data. Unlike data-driven approach where data partitions are generated automatically from the data table, this employed the human expert knowledge about the data description to determine the appropriate membership function to describe each attribute. This is to avoid the sharp boundary problem, enhance the accuracy of the constructing fuzzy models for each concerned attribute and make it correspond to human expert perception of the data. (Aly & Vrana, 2006; Oladipupo et al., 2010). Based on the literature and expert description of the identified attribute in the case study, Trapezoidal (tramf) and Triangular membership function (trimf) were found appropriate for modelling the determinant factors (Allarverdi et al., 2007). The constructed fuzzy models were implemented using C sharp (C#) programming language on Visual studio engine. The output from this phase is a fuzzified database.

Mining process is an essential phase, where intelligent methods are applied in order to discover hidden pattern from historical database. During this process, existing Apriori-like fuzzy association rule mining algorithm proposed by Gyenesei, (2001) was

adopted coupled with experts' opinion factors, so as to allow the rules discovered to correspond with human perception (Delgado et al., 2001; Verlinde et al., 2006). The algorithm was implemented to return rules with 4 attribute antecedents only. This was based on the case study's determinant factors. Also, unlike the data-driven approach the expert's opinion is factored into the existing algorithm to determine each rule consequence so that the rules will correspond intuitively to human expert perception in decision making. This approach is capable of extracting frequent relationships or hidden patterns from a repository of past experiences in form of rules with support and confidence measures for each rule. Beyond the extraction of frequent patterns, rules are also evaluated for interestingness based on the interestingness measure of certainty factor.

The data-set of 389 records consisting of 8 attributes of non-smoking men with no diabetics history of Cleveland Clinic Foundation database and Hungarian database from University of California, Irvine (UCI), online machine learning repository was used for the mining process. The implementation was done using C sharp (C#) programming language on Microsoft Visual Studio 2008.

During data representation, the extracted interesting rules were transformed into relational structure so as to enhance the rules accessibility by the expert system inference engine. SQL Server, Management Studio Express was used as a database management system. This was determined by the platform upon which the expert system was modelled

To demonstrate the potential of fuzzy association rule mining expert-driven approach in responding to the unwieldiness challenge in rule-based expert system, a fuzzy

association rule mining expert system was developed and the components were instantiated with the Coronary Heart Disease (CHD) requirement. This was used to generate a prototype fuzzy association rule mining expert system for determining CHD risk ratio in medical domain. The prototype was modelled to validate the credibility of the introduced FARME-D knowledge acquisition approach. The system was developed based on Mandani inference mechanism with MaxMin operator and centroid defuzzification method using C sharp (C#) programming language. The knowledge-base was evolved with interesting rules from the mining process.

Lastly, the performance of the prototype expert system was examined using test case approach and evaluated using quantitative measure of fuzzy expert system with a view to determining the capability of FARME-D knowledge acquisition approach in fulfilling its set objectives. Thereafter, the results of the test cases were compared with other two approaches and analyzed with t-test and ANOVA statistical analysis based on the following hypothesis:

- (1) The null hypothesis H_0 is that the mean difference $(\mu_1 - \mu_2) = 0$ or in other words the means are the same.
- (2) The alternative hypothesis H_a is that the mean difference $\neq 0$ or in other words the means are not the same.

This was done to establish a basis for the generalization of our results. A schematic model of the methodology of this thesis using a UML activity diagram is shown in Figure 1.1.

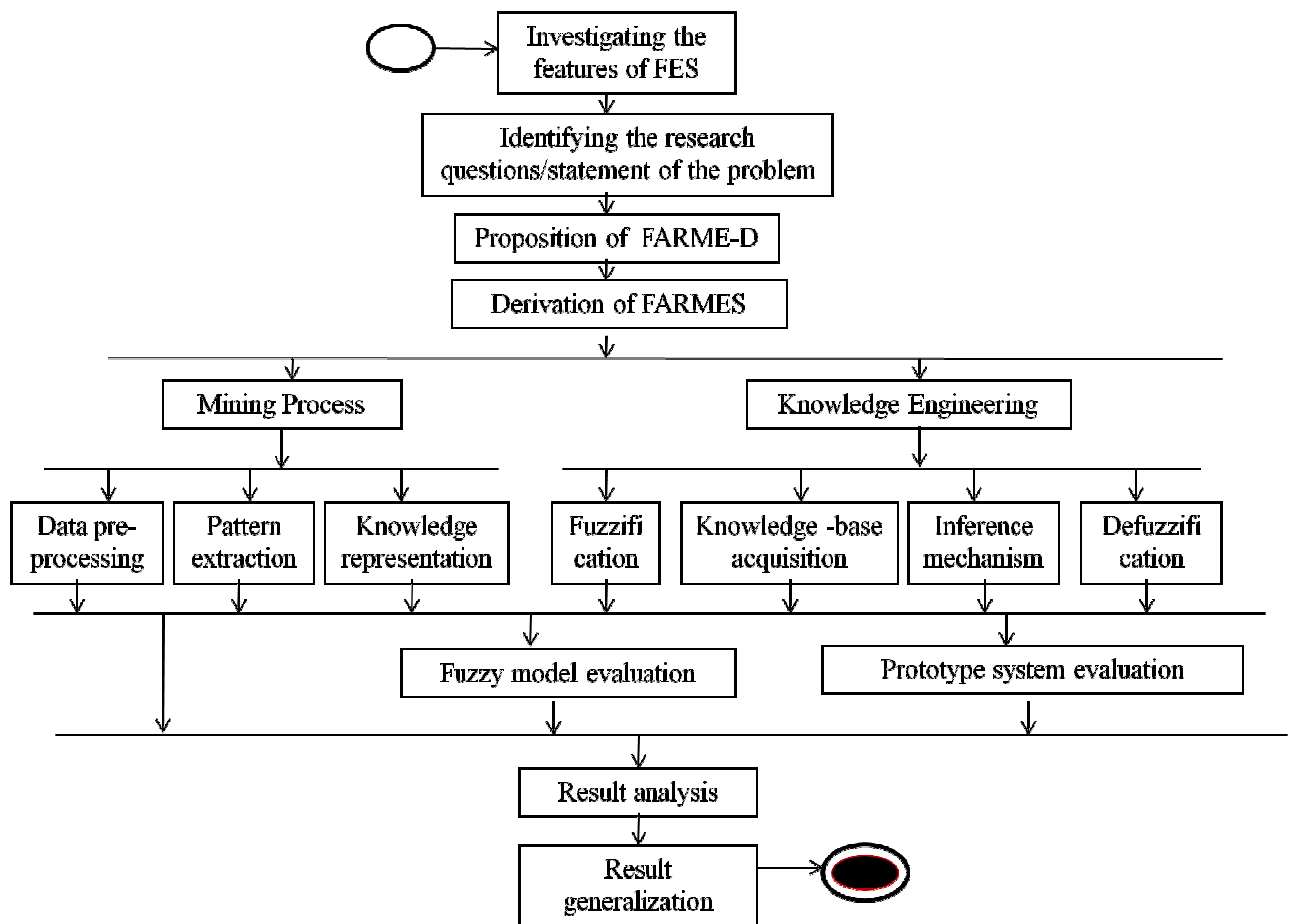


Figure 1.1: A model conceptualization of the methodology of the thesis

1.5 SIGNIFICANCE OF THE STUDY

This research work is bi-directional, with one direction in rule-based expert system theory and the other in the field of medicine being the application domain. It is significant for the following reasons:

1. The study demonstrates the feasibility of fuzzy association rule mining expert-driven approach to knowledge acquisition component of a medical rule-based expert system knowledge engineering as none is yet to be reported in the literature. The approach provides a better way to minimize rule redundancy in the knowledge base.
2. This approach also makes room for knowledge-base flexibility, such that new innovated patient instances could be used to enhance the strength of the knowledge-base time to time since the mining engine is integrated with the expert system.
3. The fuzzy concept and expert-driven approach has a significant implication in the medical rule-based expert system because of the imprecision of the medical domain expert knowledge. These guide against the sharp boundary problem and enable the extracted knowledge to correspond to human perception in the application domain.
4. The approach introduced will provide actual knowledge needed to replicate expert knowledge even when the expert is no more.
5. The Fuzzy Association Rules confidence value defined the importance rating of each rule. This can also serve as a rating weight in rule storage to enhance the accessibility of the most frequent rules which in turn enhance the system response time.

6. The study provides a platform for promoting intelligent e-medicine as a viable tool for reducing death rate in rural areas where there is no access to the human expert.

1.6 MOTIVATION FOR THE STUDY

In most existing fuzzy rule-based systems, the fuzzy rules are generated by the domain experts, especially for control problems with only a few inputs. With an increasing number of variables, the possible number of rules for a system increases exponentially, which makes it difficult for experts to define a complete rule set for good system performance (Shi, 1999; Pirnau & Maioreescu, 2008). The case is very common in medical domain where a disease could take several variables for it to be diagnosed. To attend to this issue an automated way of knowledge acquisition is considered preferable (Shi, 1999; Pirnau & Maioreescu, 2008). The advent of knowledge discovery, where hidden patterns could be extracted from a historical data storage of past domain expert decisions in form of knowledge is also a motivation for automated knowledge acquisition (Delgado et al., 2001). As rule-based expert systems are easy to formulate, emulate human cognitive process and decision making ability, they also have the limitation of large number of rules in the knowledge-base which cause the system to become unwieldy and complicates its maintenance especially in the case of subtle updates (Aly & Vrana, 2006).

In considering the issues about the rule-based systems above, the motivation for this work is two-fold. The first stem comes from trends in automated knowledge acquisition with a view to have limited number of rules and knowledge-base instant update. The second is derived from crave for knowledge-base void of sharp boundary

problem and correspond with human perception of application domain in the medical expert system context.

One of the greatest difficulties in designing a convenient expert system is acquiring the knowledge-base being the back bone of a good knowledge-based system. The more compact the systems are the more understandable they become (Meesad, 2001). The prospective user of rule-based expert systems desires a more comprehensible and compact system, that emulates and corresponds with the most intuitive human perception of concept, in order words there is quest for high comprehensible system to enhance the understanding of the expert system. One of the ways to enhance rule-based expert system comprehensibility is by minimizing the number of rules in the knowledge base while system accuracy is maintained and making the rules to correspond with human perception of the domain concepts, which is still an open issue to which this thesis is making a contribution (Aly & Vrana,2006; Pirnau & Maiorescu, 2008).

Secondly, in building an expert system there is need for the knowledge engineer to watch after sharp boundary problem which could be caused by quantitative attributes. Most important in medical field where there is combination of imprecise causal knowledge, potentially life-threatening consequences of incorrect conclusion and a big amount of quantitative attributes. Machine learning data-driven techniques might not do very well because of their tendency to overestimate or underestimate boundary data values which resulted into Sharp Boundary Problem (SBP) (Verlinde et al., 2006). Hence, this thesis seeks to investigate the effect of sharp boundary problem in medical rule-based expert system. The outcome of the investigation thereby informed the

introduction of fuzzy association rule mining expert-driven approach as a unified approach to acquire knowledge for modelling fuzzy expert systems in this thesis. It involves the domain expert knowledge for fuzzy set membership calibration to avoid sharp boundary problem and have a knowledge-base that correspond with domain expert perception.

1.7 CONTRIBUTION TO KNOWLEDGE

The contributions of this work apply to the rule-based expert system knowledge engineering in broad and medical rule-based expert systems in specific.

To the best of our knowledge the incorporation of domain experts' opinion factors into the existing fuzzy association rule mining process (where experts determine the data interval partitions, fuzzy membership function models and the fuzzy rules consequences) is being attempted for the first time to knowledge acquisition in the area of Expert System. Hence, this study presents fuzzy association rule mining expert-driven approach (FARME-D) as a viable solution approach to solving the problem of large number of rules in rule-based expert systems, especially in medical domain.

Secondly, this work has introduced experts' opinion factors into the existing fuzzy association rule mining technique for automated knowledge acquisition which allows the rules to emulate human cognitive process of decision making ability. This will also alleviate the effect of the sharp boundary problem in medical expert system.

Thirdly, thus far, to the best of our knowledge the automated knowledge acquisition processes are not integrated with the expert system, which makes it difficult for

knowledge-based update even when there are new inventions. To this effect the approach proposed in this work is integrated with the expert system in order to enhance instant update of the knowledge-base as new instance is invented by the domain experts.

Lastly, The integration of FARME-D as a component to standard fuzzy expert system architecture has resulted into a derived Fuzzy Association Rule Mining Expert System (FARMES) architecture. This enhances knowledge-base instant update.

1.8 DELIMITATION OF THE SCOPE OF THE STUDY

The main focus of this thesis is to demonstrate the feasibility of FARME-D automated knowledge acquisition approach in modelling a comprehensible fuzzy expert system. This is validated in medical domain using as a case study, Coronary Heart Disease risk determination fuzzy expert system. Although, the theoretical concepts pictured in this work are applicable to traditional rule-based systems, the prototype design and implementations in this work are based on medical fuzzy expert systems.

1.9 THESIS ORGANISATION

Chapter One of this thesis presents a general introduction, highlighting the motivation for the research, the methodology used, the aim and specific objectives of the work and the research contribution to knowledge.

Chapter Two undertakes a critical review of the expert system domain and the challenges of rule-based expert systems. The chapter presents a critical review on knowledge engineering component of an expert system. A review of related works so

as to identify the gaps that exist in literature in order to situate the context of the research undertaken in this work, is also presented in this chapter. To conclude this chapter the proposal of fuzzy association rule mining expert-driven approach of knowledge acquisition for modelling medical rule-based expert system is presented.

Chapter Three introduces the Fuzzy Association Rule Mining Expert-Driven approach: a unified solution platform to solving the research questions raised in this thesis.

Chapter Four presents the detail of a case study of automated knowledge acquisition that was undertaken to validate the proposed approach. Specifically, the details of FARME-D components are discussed. In Chapter Five, the details of the evaluation procedure for FARME-D approach are discussed.

Finally, in Chapter Six, the summary, conclusion and a discussion of the future research outlook of this thesis are presented.

CHAPTER TWO

LITERATURE REVIEW

2.1 INTRODUCTION

In building an expert system one of the greatest challenges is how to obtain the knowledge-base and the representation of the rules in the knowledge-base. As the size and scope of the problem domain increases, knowledge acquisition and knowledge engineering becomes more challenging. Knowledge acquisition is termed the bedrock of a solid knowledge engineering because it determines the effectiveness and the accuracy of an expert system. Hence, the issue of knowledge acquisition becomes crucial and continues to attract interest in knowledge engineering research.

Our approach in this thesis explores fuzzification process from expert-driven point of view as a basis for employing domain expert knowledge in knowledge discovery context for knowledge acquisition. The historic database in the context of this research represents the domain expert decision making experiences in form of structured information. In this chapter, comprehensive review of expert system concept, fuzzy logic concept, knowledge acquisition and data mining is presented.

2.2 WHAT IS AN EXPERT SYSTEM?

An Expert System (ES) is an application of Intelligent Systems (IS). ES is one of the sub-disciplines of Artificial Intelligence (AI). It is used and applied more than any other AI technology (Turban et al. 2001). ES is used interchangeable as knowledge-based system in text. Expert system is a branch of AI that makes extensive use of specialized human expertise to solve semi or ill-structured problems for which there is

no exact guaranteed solving algorithm (Aly & Vrana, 2006). An expert system is a computer application that solves complicated problems that would otherwise require extensive human intelligence. To do so, it simulates the human reasoning process by applying specific knowledge and interfaces. These expert systems represent the expertise knowledge as data or rules within the computer. These rules and data can be called upon when needed to solve problems (Turban & Arason, 2001). It is a computer program designed to model the problem solving ability of a human expert (Durkin, 1994). Feigenbaum, (1982) also defines ES, as “an intelligent computer program that uses knowledge and inference procedures to solve problems that are difficult enough to require significant human expertise for their solutions”

Within the context of this thesis ES is defined as an intelligence system which uses extracted knowledge from past domain expert decision making reasoning in form of rules to solve problems that ordinarily require human expertise for their solution, and has the capability to update its rule-base as new knowledge is discovered. There are several major application areas of expert system such as agriculture, education, environment, law manufacturing, medicine, power systems, etc. In contrast to conventional computer program where the knowledge base is often embedded in the program code, so that as the knowledge changes, the program has to be rebuilt, the knowledge-based expert systems collect the small fragments of human knowledge into a knowledge-base, which is used to reason through a problem, using knowledge that is appropriate (Abraham, 2005). An important advantage here is that within the domain of the knowledge-base, a different problem can be solved using the same program without programming efforts. Also, expert systems have the ability to explain the

reasoning process and handle levels of confidence and uncertainty that conventional algorithms could not handle (Giarratano & Riley, 1989).

2.2.1 Advantages and Disadvantages of Expert System

Some important advantages and disadvantages of expert system are as follows (Abraham, 2005; Feigenbaum, 1982):

2.2.1.1 Advantages

Ability to:

- capture and preserve irreplaceable human experience.
- develop a system more consistent than human experts, it provides consistent answers for repetitive decisions, processes and tasks.
- minimize human expertise needed at a number of locations at the same time (especially in a hostile environment that might be dangerous to human health).
- proffer solution faster than human experts.
- reduce employee training costs.
- provide centralized decision making process.
- combine multiple human expert intelligences.
- reduce the amount of human errors.
- give strategic and comparative advantages creating entry barriers to competitors
- review transactions that human experts may overlook.
- create efficiencies and reduce the time needed to solve problems.

2.2.1.2 Disadvantages of Expert Systems

- Inability to provide common sense needed in some decision making.
- Inability to give the creative responses that human expert can give in unusual circumstances.

- Inability to clearly explain their logic and reasoning.
- Challenges of automating complex processes.
- Lack of flexibility and ability to adapt to changing environments.

2.2.2 Comparison of Expert System, with Conventional System and Human Expert

According to Negnevitsky (2005), the comparison of expert system with conventional system and human expert is summarized below in:

Human Experts

- Have knowledge in a compiled form in their brain.
- Are capable of explaining lines of reasoning and providing the details.
- Use inexact reasoning and can deal with incomplete, uncertain and fuzzy information.
- Can make mistakes when information is incomplete or fuzzy.
- Can enhance the quality of problem solving via years of learning and practical training.
- Can experience slow process, inefficient and expensive.
- Can reason through human brain.

Expert Systems

- Process knowledge expressed in the form of rules and use symbolic reasoning to solve problems in a narrow domain.
- Provide a clear separation of knowledge from its processing.
- Trace the rules fired during a problem-solving session and explain how a particular conclusion was reached and why specific data was needed.
- Permit inexact reasoning and can deal with incomplete, uncertain and fuzzy data.

- Can make mistakes when data is incomplete or fuzzy.
- Enhance the quality of problem solving by adding new rules or adjusting old ones in the knowledge-base. When new knowledge is acquired, changes are easy to accomplish.
- Has reasoning. It reasons through the inference engine.

Conventional programs

- Process data and use algorithms- a series of well-defined operations to solve general numerical problems.
- Do not separate knowledge from the control structure to process this knowledge.
- Do not explain how a particular result was obtained and why input data was needed.
- Work only on problems where data is complete and exact.
- Provide no solution at all, or a wrong one, when data is incomplete or fuzzy.
- Enhance the quality of problem solving by changing the program code, which affects both the knowledge and its processing, making changes difficult.
- has no reasoning facility, except for a specific if-then statement within the program

2.3 APPROACHES FOR MODELLING AN EXPERT SYSTEM

Building an expert system is known as *knowledge engineering* and its practitioners are called *knowledge engineers*. The knowledge engineer must make sure that the computer has all the knowledge needed to solve a problem. The knowledge engineer chooses one or more forms in which to represent the required knowledge as symbol patterns in the memory of the computer; that is, he (or she) must choose a *knowledge representation*. He must also ensure that the computer can use the knowledge

efficiently by selecting from a handful of *reasoning methods*. There are different approaches for modelling an ES based on knowledge representation and presentation for solving problems as reviewed by Ally & Vrana, (2006); Kaula & Lander, (1995). The approaches are explained to include their advantages and limitations.

2.3.1 The rule-based approach

Traditional, expert system engineering is based on the production systems approach (rule-based systems) which emphasizes building a single monolithic knowledge-base. Production rules are written in form of IF-Then rules:

if premise (condition) Then consequent(action)

The major advantages of rule-based systems are many: they are easy to formulate, they emulate human cognitive process and decision making ability, and represent knowledge in a structured homogeneous and modular way. However, there are several limitations associated with those systems: control structures contained in the order of rules cause loss of flexibility, large number of rules in the knowledge-base causes the system to become unwieldy and complicates its maintenance especially in the case of subtle updates and the difficulty in assigning confidence rating to each rule (Kaula & Lander, 1995). The second limitation has great influence on the ES comprehensibility. This has motivated researchers to finding solution to the large number of rules in the rule-based systems knowledge-base in order to enhance the system understandability. The simplicity of this approach has encouraged its common usage in modelling a rule-based expert system.

2.3.2 The blackboard system approach

The blackboard organizes and stores the intermediate problem solving data. A set of independent domain-specific modules called knowledge sources produce changes to the blackboard that lead incrementally to a solution of the problem. Communication between knowledge sources is conducted solely through changing the blackboard. The blackboard model is advantageous in that it provides a very flexible control structure for solving the problem and also provides for modularity. One of the limitations of the model is that such a system does not specify how the specific piece of knowledge should be handled by other knowledge sources (Chi et al., 2001).

2.3.3 The frame-based approach

A frame is another approach used to capture and store knowledge in a knowledge-base. It relates an object or item to various facts or values. A frame-based representation is ideally suited for object-oriented programming techniques (Abraham, 2005). Expert systems making use of frames to store knowledge are also called *frame-based expert systems*. This approach provides a more structured representation in the form of frames. A frame describes an object, consisting of slots containing default values, pointers to other frames, sets of rules, or procedures. Frames are linked to provide for inheritance and communicate by passing messages. However, the modularity of knowledge represented in frames cannot be defined clearly, and the representation lacks flexibility. Also, frame-based systems do not provide a way of defining unalterable slots (Kaula & Lander, 1995).

2.3.4. The Open-based Expert System (OES) Approach

The OES approach proposes an expert system consisting of a number of independently developed smaller autonomous expert subsystems (AESs) that communicate during problem solving. Every participating AES exists with its own self-knowledge, i.e., rules. No AES controls directly the knowledge of another AES, thereby making communication and negotiations essential for problem solving. An AES is made aware of the other AESs by accessing the concept dictionary. Communication between AESs is facilitated by a single communication dictionary, which contains the procedures for implementing communication protocols called message acts. The open system approach is suitable for the development of large expert systems.

One advantage of such open system is that it emphasizes direct communication of knowledge between expert subsystems from within their background knowledge. Furthermore, the use of human communication mechanisms makes passing of messages more representative of the working environment. Moreover, it is possible to add or remove subsystems (AESs) with minimal impact on the environment. Also, there is no global schema or global consistency. However, one limitation of such system is the problem of inconsistency, which may exist.

2.3.5 The Object-Oriented (OO) approach

It is an extension of the frame-based approach. It provides for the development of autonomous objects, which communicate by passing messages to one another during problem solving. The technical benefits of Object-oriented paradigm to the system development are summarized as follows:

- It contributes to modelling and programming productivity.

- Through greater reusability, modularity, inheritance and independence, OO reduces development cost and time.
- It simplifies and reduces problem complexity through a hierarchical and systematized modelling.
- It greatly reduces system maintenance cost and facilitates modification through encapsulation.
- It greatly enhances system flexibility through polymorphism.
- It produces a more reliable system, through ease of communication and understandability.
- It also enhances object sharing, which promotes integration, and clarify interfacing.

One of the limitations of OO systems is that the system may be slower in execution. Also, the message-passing mechanism in the OO approach does not include the provisions of how a message has to be handled by the receiving object from a sender object's background and perspective. In addition, as human communication includes context and intention apart from content, merely sending a message content becomes a very restricted form of communication.

2.4 RULE-BASED EXPERT SYSTEM (RBES)

Different approaches for modelling expert systems, which include the rule-based approach (Aly & Vrana, 2006) have been critically reviewed in 2.4. However, all of these approaches have their limitations. In the early 1970s, Newell and Simon from Carnegie-Mellon University proposed a production system model, which brought in the rule-based expert systems (RBES) (Newell & Simon, 1972). The rule-based expert systems are intelligent computer program that collect the small fragments of human

knowledge into a knowledge-base in form of if-then rules, couple with inference procedures to solve problems that are difficult enough to require significant human expertise for their solutions (Feigenbaum, 1980; Abraham, 2005).

A rule-based system consists of *if-then* rules, a bunch of facts, and an interpreter controlling the application of the rules, given the facts. These *if-then* rule statements are used to formulate the conditional statements that comprise the complete knowledge-base. A single *if-then* rule assumes the form 'if x is A then y is B '. The if-part of the rule: ' x is A ', is called the *antecedent* or *premise*, while the then-part of the rule ' y is B ', is called the *consequent* or *conclusion* (Abraham, 2005). In rule-based expert system, the domain knowledge is represented by a set of facts about current situation. The inference engine compares each rule stored in the knowledge-base with facts contained in the database. When the IF(condition) part of the rule matches a fact, the rule is fired and its THEN (action) part is executed. The fired rule may change the set of facts by adding a new fact (Negnevitsky, 2005). A rule-based expert system can adopt the fuzzy concept in other to enhance its functionality. This is called a Fuzzy Expert System (FES) (Abraham, 2005).

Rule-based expert systems are easy to formulate; they emulate human cognitive process and decision-making ability; and finally, they represent knowledge in a structured homogeneous and modular way. The limitations of these systems include: large numbers of rules in the knowledge base that causes the system to become unwieldy because of the presence of rules that might not be relevant to the application domain. This inevitably complicates its maintenance especially in the case of subtle updates and lower comprehensibility of the expert system (Aly & Vrana, 2006).

According to Roubos & Setnes (2001), comprehensible knowledge representation is a key advantage of FESs over black box schemes. Accuracy alone may not be sufficient to show the goodness of an expert system. Therefore, comprehensibility measure is an additional quantitative assessment that indicates whether a rule-based expert system is understandable or not (Meesad, 2001). The structure of a traditional rule-based system is shown in Figure 2.1

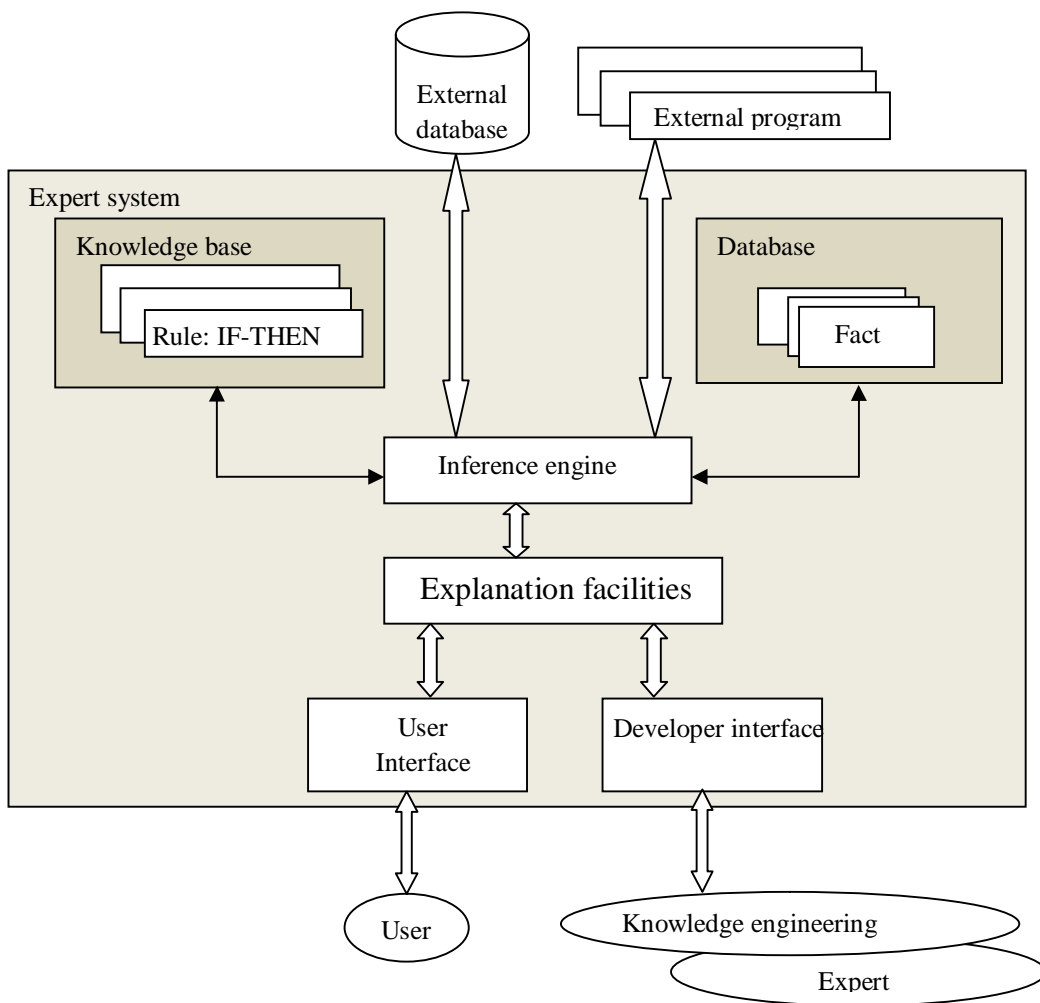


Figure 2.1 Complete structure of a rule-based expert system (Negnevitsky, 2005).

2.4.1 The RBES component

The rule-based expert system, according to Negnevisky (2005), has five components:

- the user interface
- the database
- the explanation facilities
- the inference engine and
- the knowledge-base

Sasikumar et al., 2007 summarizes the five components into three as listed below and identify the rule-base and the working memory as the data structures which the system uses and the inference engine as the basic program which is used.

- the working memory,
- the rule-base, and
- the inference engine

(a) *The user interface*

The User Interface is the means of communication between the user and the ES. The purpose of the user interface is to offer ease of use of the ES for developers, users, and administrators (Abraham, 2005). It is responsible for posing the questions to the user, reading the user's reply and explaining the rules used to reach a conclusion.

(b) *The Database*

The Database includes a set of facts used to match against the IF (condition) parts of rules stored in the knowledge-base. This is called working memory (WM) in some other text (Sasikumar et al., 2007). It represents the set of facts known about the domain. The elements represent the current state of the world. For example, in a

medical domain expert system, the WM could contain the details of a particular patient being diagnosed. The working memory is the storage medium in a rule-based system and helps the system *focus* on its problem solving. It is also the means by which rules *communicate* with one another. The actual data represented in the working memory depends on the type of application. The initial working memory, for instance, can contain *a priori* information known to the system. The inference engine uses this information in conjunction with the rules in the rule-base to derive additional information about the problem being solved. Also, the content of the database is useful for automated knowledge-base acquisition.

(c) Explanation facilities

The explanation facility is one of the most important features of a rule-based expert system. Negnevitsky (2005) views the explanation base as a RBES components that enables the user to ask the expert system how a particular conclusion is reached. The explanation facility allows a user to understand how the expert system arrived at certain results (Abraham, 2005). Turban et al., (2001) as referenced by De Kock (2003) also viewed the explanation facility as a separate ES component where the behavior of the ES can be accounted for to provide answers to questions such as:

- How was a certain conclusion reached?
- Why was a certain question asked?
- What is the plan to reach the solution?
- Why was a certain alternative rejected?

(d) Inference engine

The inference engine is the mechanism that performs the reasoning and searching in RBES. The inference engine matches facts in the working memory against rules in the rule-base, and it determines which rules are applicable according to the reasoning method adopted by the engine (Soe & Zaw, 2008). The engine is activated when the user initiates the consultation session. According to Abraham (2005), the purpose of the inference engine is to seek information and relationships from the knowledge-base and to provide answers, predictions, and suggestions in the way a human expert would. The inference engine must find the right facts, interpretations, and rules and assemble them correctly. Inference engine is viewed by Negnevitsky as a linker that links the rules given in the knowledge-base with the facts provided in the database. It carried out the reasoning whereby the expert system reaches a solution (Negnevitsky,2005). This sequence of steps and the linking of facts and patterns and rules are known as chaining (Klein & Methlie, 1995).

Two basic chaining techniques for inferring facts or conclusions from the knowledge-base are:

- Forward Chaining
- Backward Chaining

De Kock, (2003) in his write up identified the hybrid chaining where both forward and backward chaining could be engaged in case of a complex reasoning.

Forward chaining is known as data-driven reasoning (Negnevitsky,2005). In forward chaining, the current situation supplied by the user is matched with the rules' antecedent in the knowledge-base. If there is a match then the inference engine fires the rule and adds the conclusion to the list of known facts. The match-fire cycle stops when no further rules can be fired. It can be very efficient, especially if many rule

conditions match the data provided by the user. In a *forward chaining* system, the initial facts are processed first, and keep using the rules to draw new conclusions given those facts

The second techniques is ***backward Chaining***, also known as goal-driven reasoning (Negnevitsky,2005). Backward chaining starts with a list of [goals](#) (or a [hypothesis](#)) and works backwards from the [consequent](#) to the [antecedent](#) to see if there is [data](#) available that will support any of these consequents. An [inference engine](#) using backward chaining would search the [inference](#) rules until it finds one which has a consequent (**Then** clause) that matches a desired goal. If the antecedent (**If** clause) of that rule is not known to be true, then it is added to the list of goals.

According to De Kock, the hybrid chaining employs both the forward and backward chaining needed when a large problem domain is involved. A more efficient program is yielded when the two techniques are used in combination. One inference engine will not suit all possible tasks solved by an ES (De Kock, 2003).

(e) The Knowledge-base

The Knowledge-base (also called the rule-base) is the set of rules which represents the knowledge about the application domain (Sasikumar et al, 2007). It stores all relevant information, data, rules, cases, and relationships used by the expert system. A knowledge-base can combine the knowledge of multiple human experts (Abraham, 2005); it is the backbone of the ES. The power and effectiveness of the ES is equal to the knowledge it contains. The acquisition of expert knowledge is crucial and involves the gathering of information about a domain usually from a domain expert, a task which can be difficult. This information is translated, represented and stored as a knowledge-base.

2.4.2 Rule-base Conflict Resolution

The choice of which rule to fire is done by conflict resolution. The most commonly used conflict resolution strategy is the first found strategy where the first applicable rule is executed (Klein and Methlie, 1995) or fired by applying rule deduction or using formal logic. Some other conflict resolution methods are:

- (a) **Specificity:** using this strategy, rules with more antecedents are preferred with fewer conditions, that is, specific rules are selected in preference to general rules.
- (b) **Recency:** with this strategy, every element of the working memory is tagged with a number indicating how recent the data is. When a rule has to be selected from the conflict set, the rule with an instantiation which uses the most recent data is chosen. The idea here is that a rule which uses more recent data is likely to be more relevant than one which uses older data.
- (c) **Refractoriness:** this prevents the same rule from applying again and again. If an instantiation has been applied in a cycle, it will not be allowed to fire again. Refractoriness is important for two reasons. It prevents the system from going into a loop (i.e., repeated firing of the same rule with the same instantiation). It also improves the efficiency of the system by avoiding unnecessary matching.

2.4.3 Conventional Programs vs Rule-based Systems

The difference between conventional programs and rule based systems can be summarized according to Sasikumar et al. (2007) as follows:

- The major feature which differentiates a rule based system from a conventional program is its declarativeness. In a rule based system, the knowledge of the world is stated declaratively in the form of rules and facts. A control mechanism

is used to infer new facts about the world. In a conventional program, such a clean separation does not exist.

- A rule based system can also be given a procedural interpretation. However, the nature of procedure invocation differs from that of conventional programs. The rules do not invoke other rules directly. Rules modify the contents of the working memory. This modification of the working memory causes other rules to become *fireable*. This is unlike procedure invocations in conventional programs.
- A rule based system exhibits a high degree of modularity compared to most conventional programs. Each rule represents an independent piece of knowledge. Therefore the addition of a rule to the rule base need not affect any of the other rules. In contrast, in conventional programs, the addition of a new procedure would involve changes in the existing code.

2.4.4 Advantages of Rule-based Systems

Some of the advantages of rule based systems are sated below according to Sasikumar et al. (2007):

- **Homogeneity**

Because of the uniform syntax, the meaning and interpretation of each rule can be easily analyzed.

- **Simplicity**

Since the syntax is simple, it is easy to understand the meaning of rules. Domain experts can often understand the rules without an explicit translation. Rules therefore can be self-documenting to a good extent.

- **Independence**

While adding new knowledge one need not be worried about where in the rule base the rule is added, or what the interactions with other rules are. In theory, each rule is an independent piece of knowledge about the domain. However, in practice, this is not completely true, as we shall see in the next section.

- **Modularity**

The independence of rules leads to modularity in the rule base. You can create a prototype system fairly quickly by creating a few rules. This can be improved by modifying the rules based on performance and adding new rules.

- **Knowledge is Separated from Use and Control**

The separation of the rule base from the inference engine separates the knowledge from how it is used to solve the problem. This means that the same inference engine can be used with different rule bases and a rule base can be used with different inference engines. This is a big advantage over conventional programs where data and control are intermixed.

- **Procedural Interpretations**

Apart from declarative interpretation, rule based systems have procedural interpretations also, which enable them to be viewed as computational models.

2.4.5 Drawbacks of Rule-based Systems

In spite of the advantages mentioned above, rule based systems have their own drawbacks. Some of the drawbacks are listed below:

- **Lack of Methodology**

There is no methodology (i.e., systematic procedure), yet for creating rule based systems. Most systems are built based on intuition, prior experience, and trial and error.

- **Interaction among Rules**

An advantage of the rule based representation was stated to be the relative independence of the different pieces of knowledge. However, in many systems you cannot assume that the rules do not interact among themselves. In certain cases, ignoring rule interaction could lead to unexpected results.

- **Opacity**

Rule based systems provide no mechanism to group together related pieces of knowledge. This makes any structure/relationships in the domain opaque in the rule base.

- **Lack of Structure**

The simplicity of rules leads to the drawback that all rules are at the same level. In many domains it would be useful to have rules at different levels in a hierarchy, but the pure production system model does not support this.

- **Representing Procedural Tasks**

Some tasks which can be easily represented in terms of procedural representations are not very easy to represent using rule based representations.

- **Inefficiency**

As mentioned earlier a large amount of time is taken in each cycle to match applicable rules in the rule base. For large rule bases, this often leads to inefficiencies. However, there is work going on to reduce the number of rules in the rule-base and structuring the rule-base to increase the efficiency in which this thesis is contributing.

2.4.6 Good Domains for Rule-based Systems

Rule-based systems have been used for a variety of applications such as medical diagnosis and machine fault troubleshooting, etc. It would be difficult to list all such domains. However here are some characteristics of domains which can meaningfully use the rule based framework (Davis and King, 1984).

- **Where the Knowledge is diffuse**

For example, clinical medicine is a good domain because it consists of a large number of facts which are more or less independent of each other. In contrast, mathematics has a strong theoretical base and a set of inter-related principles which need to be applied to solve problems.

- **Where Processes are representable as independent actions**

If a process consists of a set of independent actions, there need not be much communication among the rules and therefore such processes are ideal for the rule-based framework. For example, a domain like medical diagnosis is a good domain, as opposed to a domain like accounting.

- **Where Knowledge can be easily separated from its use**

The periodic table in chemistry provides knowledge about the different elements. This knowledge is independent of how it is used.

2.5 OVERVIEW OF FUZZY THEORY

2.5.1 What is fuzzy Logic?

Fuzzy Logic was initiated in 1965 by Lotfi A. Zadeh, a Professor of Computer Science at the University of California in Berkeley. Fuzzy Logic has emerged as a profitable tool for controlling and steering systems and complex industrial processes, as well as

for household and entertainment electronics, and expert systems and applications. It is the theory of fuzzy sets that calibrate vagueness, and used to describe fuzziness.

Fuzzy logic is a set of mathematical principles for knowledge representation based on degrees of membership. A Fuzzy logic system also has a series of rules comprising of an antecedent and a consequent, combined as if–then semantics. An antecedent is a conjunction of input variables, each as an expressed degree of fuzzy set (membership function). A consequent is a single output variable. It is an expressed degree of some fuzzy set. Fuzzy Logic (FL) unlike two –values Boolean logic is a multi-valued logic, that allows intermediate values to be defined between conventional evaluations like true/false, yes/no, high/low, etc. Fuzzy logic uses the continuum of logical values between 0 (completely false) and 1 (completely true).

2.5.2 Observations about fuzzy logic

Here is a list of general observations about fuzzy logic:

- ***Fuzzy logic is conceptually easy to understand.*** The mathematical concepts behind fuzzy reasoning are very simple. Fuzzy logic is a more intuitive approach without far-reaching complexity.
- ***Fuzzy logic is flexible.*** With any given system, it is easy to layer on more functionality without starting again from scratch.
- ***Fuzzy logic is tolerant of imprecise data.*** Everything is imprecise if closely looked enough, but more than that, most things are imprecise even on careful inspection. Fuzzy reasoning builds this understanding into the process rather than tacking it onto the end.
- ***Fuzzy logic can model nonlinear functions of arbitrary complexity.*** You can create a fuzzy system to match any set of input-output data. This process is

made particularly easy by adaptive techniques like Adaptive Neuro-Fuzzy Inference Systems (ANFIS), which are available in Fuzzy Logic Toolbox software.

- ***Fuzzy logic can be built on top of the experience of experts.*** In direct contrast to neural networks, which take training data and generate opaque, impenetrable models, fuzzy logic lets you rely on the experience of people who already understand your system.
- ***Fuzzy logic can be blended with conventional control techniques.*** Fuzzy systems do not necessarily replace conventional control methods. In many cases fuzzy systems augment them and simplify their implementation.
- ***Fuzzy logic is based on natural language.*** The basis for fuzzy logic is the basis for human communication. This observation underpins many of the other statements about fuzzy logic. Because fuzzy logic is built on the structures of qualitative description used in everyday language, fuzzy logic is easy to use.

2.5.3 A Fuzzy Set has Fuzzy Boundaries

A classical set $A \subseteq X$, is defined as a collection of elements $x \in X$. Then the element x either belongs to A ($x \in A$) or does not belong to A ($x \notin A$). Let X be the universe of discourse and its elements be denoted as x . In the classical set theory, crisp set A of X is defined as function $f_A(x)$ called the characteristic function of A :

$$f_A(x) : X \rightarrow \{0, 1\}, \text{ where } f_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases} \quad (2.1)$$

This set maps universe X to a set of two elements. For any element x of universe X , characteristic function $f_A(x)$ is equal to 1 if x is an element of set A , and is equal to 0 if x is not an element of A .

A fuzzy set is any set that allows its members to have different grades of membership (membership function) in the interval $[0,1]$. In the fuzzy theory, A fuzzy set A in X is defined

as a set of ordered pairs

$$= \{ (x, \mu_A(x)) \mid x \in X \} \quad (2.2)$$

where $\mu_A(x)$ is called the membership function of set A

$$\mu_A(x) : X \rightarrow \{0, 1\}, \text{ where } \mu_A(x) = 1 \text{ if } x \text{ is totally in } A; \quad (2.3)$$

$$\mu_A(x) = 0 \text{ if } x \text{ is not in } A;$$

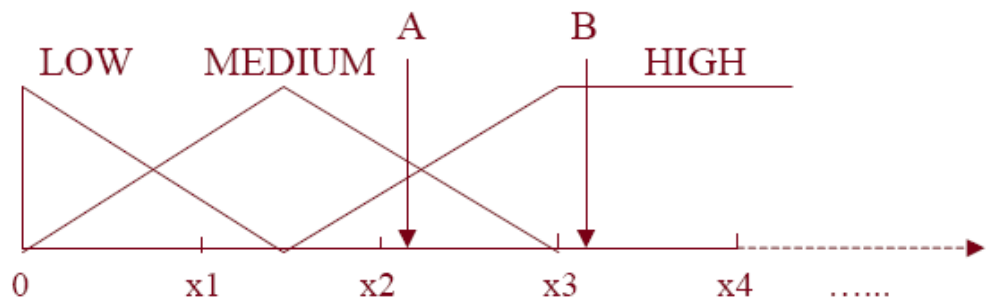
$$0 < \mu_A(x) < 1 \text{ if } x \text{ is partly in } A.$$

This set allows a continuum of possible choices. For any element x of universe X , membership function $\mu_A(x)$ equals the degree to which x is an element of set A . This degree, a value between 0 and 1, represents the degree of membership, also called membership value, of element x in set A .

Fuzzy logic allows overlapping of categories and also avoid sharp boundary problems. This is explained better with the graph shown in Figure 2.2, Figure 2.3 and Figure 2.4 with element A and B . Figure 2.2(a) shows the classical set which categorises element as either a member or not, while the Figure 2.2(b) reflects the fuzzy property of overlapping of categories whereby element “ A ” belongs to “Medium”, to a particular degree, and also belong to a neighbouring category “High” as represented in the figure to a particular degree. Figure 2.3 shows how crisp logic could overestimate or underestimate boundary values such that if A ’s age was 20 years and regarded as being young yesterday, on celebrating his birthday on the following day he clocked 21. Then ‘ A ’ drastically move to another age category (middle age) over the night. This is

obviously not reasonable and leads to a so called sharp boundary problem (Kuok et al., 1998; Oladipupo et al., 2010). Figure 2.4 on the other hand shows a gradual transformation of A's age from young to Middle age in order to avoid sharp boundary problem.

(a) A Non-Fuzzy set partition



(b) A Fuzzy set partition

Figure 2.2 Fuzzy logic allows overlapping of categories



Figure 2.3 : Non-Fuzzy sets: with sharp boundary problem

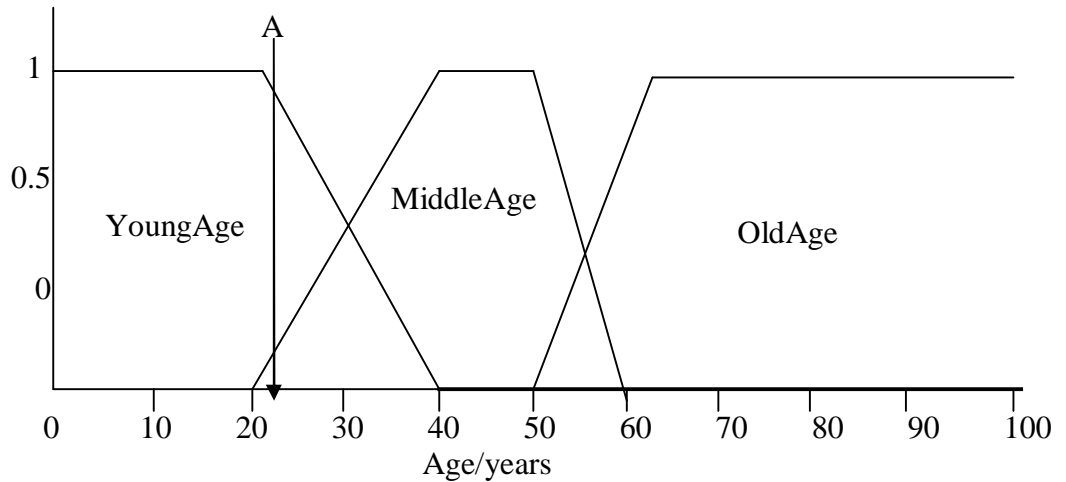


Figure 2.4: Fuzzy sets: void of sharp boundary problem

2.5.4 Fuzzy set Operations

We can introduce basic operations on fuzzy sets. Similar to the operations on crisp sets we also want to intersect and unify fuzzy sets. Zadeh (1965) suggested the minimum operator for the intersection and the maximum operator for the union of two fuzzy sets. The intersection of two fuzzy sets A and B is specified in general by a function $T : [0,1] \times [0,1] \rightarrow [0,1]$, which aggregates two membership grades as follows:

$$\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x)) = \mu_A(x) * \mu_B(x) \quad (2.4)$$

where $*$ is a binary operator for the function T . This class of fuzzy intersection operator is usually referred to as T -norm operators (Jang et al., 1997). Four of the most frequently used T-norm operators are

$$\text{Minimum: } (a, b) = \min(a, b) = a \cap b \quad (2.5)$$

$$\text{Algebraic product: } (a, b) = a \cdot b \quad (2.6)$$

$$\text{Bounded product: } (a, b) = 0 \cup (a + b - 1) \quad (2.7)$$

$$\text{Drastic product: } (a, b) = \begin{cases} a, & b = 1 \\ b, & a = 1 \\ 0, & a, b < 1 \end{cases} \quad (2.8)$$

Like intersection, the fuzzy union operator is specified in general by a function $S: [0,1] \times [0,1] \rightarrow [0,1]$, which aggregates two membership grades as follows:

$$a \cup b = (a, b) = a \bar{\cap} b \quad (2.9)$$

where $\bar{\cap}$ is the binary operator for the function S . This class of fuzzy union operator is often referred to as T -conorm (or S -norm) operators (Jang et al., 1997). Four of the most frequently used T-conorm operators are

$$\text{Maximum: } (a, b) = \max(a, b) = a \cup b \quad (2.10)$$

$$\text{Algebraic sum: } (a, b) = a + b - a \cdot b \quad (2.11)$$

$$\text{Bounded sum: } (a, b) = 1 \cap (a + b) \quad (2.12)$$

$$\text{Drastic sum: } (a, b) = \begin{cases} a, & b = 0 \\ b, & a = 0 \\ 1, & a, b > 0 \end{cases} \quad (2.13)$$

2.5.5 Membership functions

Membership function defines a fuzzy set by mapping crisp values from its domain to the sets associated degree of membership. The degree to which a crisp value is

compatible to a membership function, value from 0 to 1, is referred to as the degree of membership. It is otherwise known as truth value or fuzzy input. A label is the descriptive name used to identify a membership function. The number of labels corresponds to the number of regions that the universe should be divided, such that each label describes a region of behavior. A scope must be assigned to each membership function that numerically identifies the range of input values that correspond to a label.

The type of representation of the membership function depends on the base set. A *membership function* (MF) can be simply viewed as a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. The input space is sometimes referred to as the *universe of discourse*. The only condition a membership function must really satisfy is that it must vary between 0 and 1. There are a number of ways membership function can be represented among which we have:

- Triangular membership function (*trimf*)
- Trapezoidal membership function (*trapezmf*)
- Gaussian function (*gaussmf*)
- Generalized Bell membership function (*gbellmf*)
- Sigmoidal membership function (*sigmf*)

(a) ***Triangular membership function (trimf)***: This is specified by three parameters (a,b,c) with (a<b<c) determining the x coordinates of the three angles. Variable x is the crisp value that its membership function is to be determined within the universe of discourse. The graphical representation is shown in Figure 2.5. *trimf* can be represented mathematically by either of these two mathematical models:

$$(i) \quad \text{triangle}(x; a, b, c) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \quad (2.14)$$

$$(ii) \quad \text{triangle}(x; a, b, c) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ \frac{c-x}{c-b}, & b \leq x \leq c \\ 0, & c \leq x \end{cases} \quad (2.15)$$

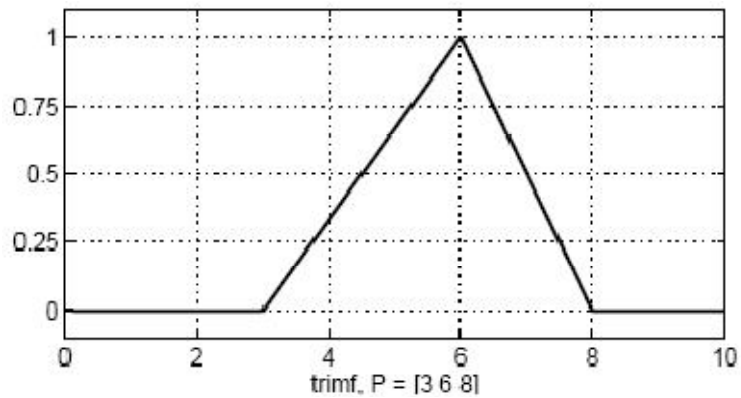


Figure 2.5 Triangular MF.

(b) **Trapezoidal membership function (trapmf):** This is specified by four parameters {a,b,c,d} with (a<b<=c<d) determine the x coordinates of the four angles of the underlying trapezoidal membership function. Figure 2.6 shows the graphical representation of trapmf. It can be represented with either of the following mathematical models:

$$(i) \quad \text{trapezoid}(x; a, b, c, d) = \max\left(\min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), 0\right) \quad (2.16)$$

$$(ii) \quad \text{trapezoid}(x; a, b, c, d) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & b \leq x \leq c \\ \frac{d-x}{d-c}, & c \leq x \leq d \\ 0, & d \leq x \end{cases} \quad (2.17)$$

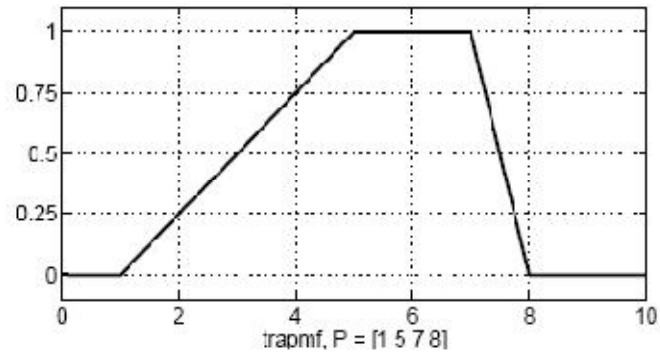


Figure 2.6 Trapezoidal MF.

- (c) **Gaussian function:** This takes three parameters x , c and σ , x is a crisp value. c is the center of the graph while σ is the width. Figure 2.7 shows the graphical representation of `gaussmf`. The mathematical representation is

$$\text{gaussian}(x; \sigma, c) = e^{\left(\frac{-(x-c)^2}{2\sigma^2}\right)} \quad (2.18)$$

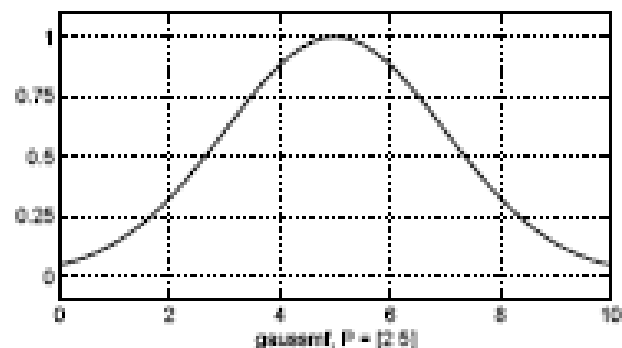


Figure 2.7 Gaussian MF.

- (d) **Generalized Bell membership function:** The *generalized bell* membership function is specified by three parameters and has the function name `gbellmf`. The bell membership function has one more parameter (3 parameters) than the Gaussian membership function, so it can approach a non-fuzzy set if the free parameter is tuned. Because of their smoothness and concise notation, Gaussian

and bell membership functions are popular methods for specifying fuzzy sets. Both of these curves have the advantage of being smooth and nonzero at all points. If b is negative the shape becomes an upside-down bell. We can adjust c and a to vary the center and width of the mf and then use b to control the slopes at the crossover points. Figure 2.8 shows the graphical representation of gbellmf. The mathematical model is:

$$bell(x; a, b, c) = \frac{1}{1 + \left| \frac{x-c}{a} \right|^{2b}} \quad (2.19)$$

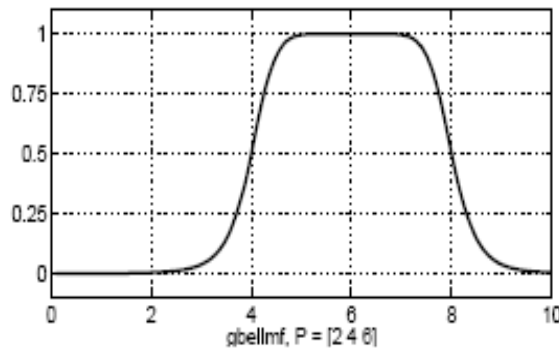


Figure 2.8 Generalized Bell MF

- e) **Sigmoidal membership function:** The sigmoidal function, $\text{sigmf}(x, [a \ c])$, is a mapping on a vector x , and depends on two parameters a and c . Depending on the sign of the parameter a , the sigmoid membership function is inherently open to the right or to the left, and thus is appropriate for representing concepts such as "very large" or "very negative." Where a controls the slope at the crossover point $x = c$. Figure 2.9 shows the graphical representation. The mathematical model is:

$$\text{sig}(x; a, c) = \frac{1}{1 + e^{-a(x-c)}} \quad (2.20)$$

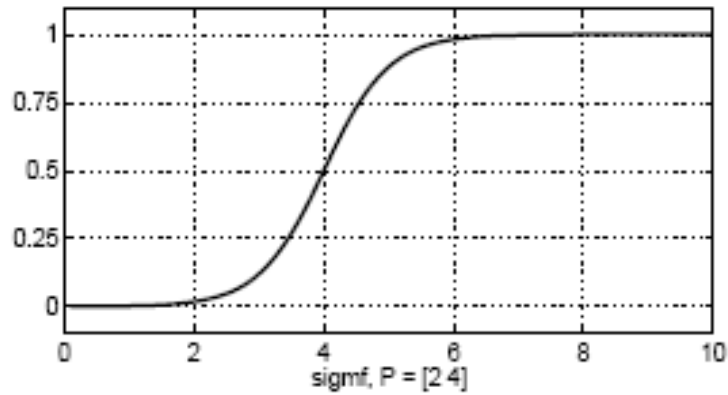


Figure 2.9 Sigmoidal MF

When considering the number of membership functions to exist within the universe of discourse, one must consider that:

- i) Too few membership functions for a given application will cause the response of the system to be too slow and fail to provide sufficient output control in time to recover from a small input change. This may also cause oscillation in the system.
- ii) Too many membership functions may cause rapid firing of different rule consequents for small changes in input, resulting in large output changes, which may cause instability in the system.

These membership functions should also be overlapped. No overlap reduces a system based on Boolean logic. Every input point on the universe of discourse should belong to the scope of at least one but no more than two membership functions. No two-membership functions should have the same point of maximum truth, (1). When two membership functions overlap, the sum of truths or grades for any point within the overlap should be less than or equal to 1. Overlap should not cross the point of maximal truth of either membership function. There are two indices to describe the overlap of membership functions quantitatively. These are overlap ratio and overlap robustness. Figure 2.10 shows the slope of membership functions.

$$= \frac{\int_{L}^{\infty} \mu_1(x) dx}{\int_{L}^{\infty} \mu_2(x) dx} \quad (2.21)$$

$$= \frac{\int_{L}^{\infty} \mu_1(x) dx}{\int_{L}^{\infty} \mu_2(x) dx} = \frac{f(L)}{f(U)} \quad (2.22)$$

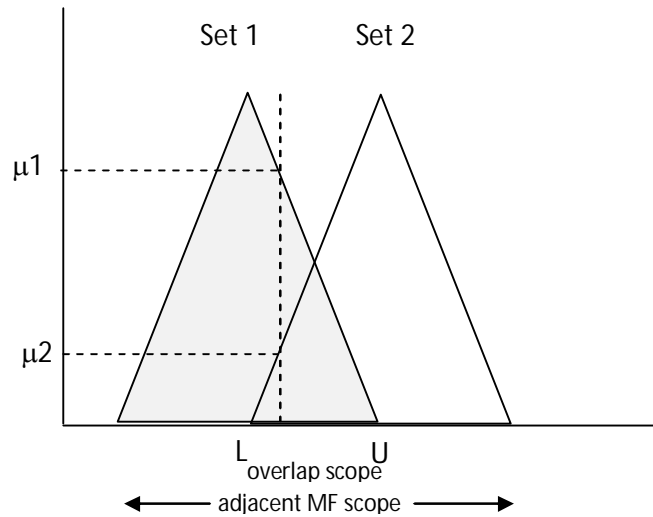


Figure 2.10 The slope of membership functions

2.5.6 Linguistic Variables and Terms

Fuzzy Linguistic Variables are used to represent qualities spanning a particular spectrum. At the root of fuzzy set theory lies the idea of linguistic variables. Each linguistic variable may be assigned one or more linguistic values, which are in turn connected to a numeric value through the mechanism of membership functions. According to Zadeh, (1975) ,

A linguistic variable is characterized by a quintuple denoted by $\langle X, T(X), \mu, \nu, \rho \rangle$ in which:

X is the name of the variable

T(X) is the term set of X whose elements are labels of linguistic values of X,

G is generally a grammar for generating the names of X ,

M is a semantic rule for associating with each label $L \in T(X)$ its meaning

$M(L)$, is a fuzzy set on the universe X whose base variable is x

For example, consider a linguistic variable named temperature, that is, $X =$ temperature, with $T = [0,50]$ and base variable $t \in T$. The terms set associated with temperature could be $T(\text{temperature}) = \{\text{very low, low, medium, high, very high}\}$ where each term in $T(\text{temperature})$ is a label of a linguistic value of the temperature.

2.6 FUZZY RULE-BASED EXPERT SYSTEM

A rule-based expert system can adopt the fuzzy concept in order to enhance its functionality. This is called Fuzzy Rule-based Expert System. A fuzzy rule-based expert system is simply referred to as Fuzzy Expert System (FES). FES is an expert system, which consists of fuzzification, inference, knowledge-base, and defuzzification subsystems. It uses collection of fuzzy membership functions and rules, instead of Boolean logic, to reason about data in the inference mechanism (Aly & Vrana, 2006, Schneider et al., 1996). This approach is used to solve decision making problems, for which no exact algorithm exists. The problem relies on human expertise in form of If-Then rules. FES is well suited to the problem, which exhibits uncertainty, which resulted from inexactness, vagueness or subjectivity.

The advantages of FES over conventional production rule-based expert systems are characterized by Shah et al., (2006) as follows: (a) fuzzy sets symbolize natural language terms used by experts; (b) since the expert knowledge captured in “If...Then” statements is often not naturally true or false, fuzzy sets afford representation of the knowledge in a smaller number of rules; and (c) smooth mapping

can be obtained between input and output data. Figure 2.11 illustrates the basic architecture of a fuzzy expert system. The fuzzy expert system according to Shi et al. (1999):

- 1) Determine the fuzzy membership values activated by the inputs.
- 2) Determine which rules are fired in the rule set.
- 3) Combine the membership values for each activated rule using the AND operator.
- 4) Trace rule activation membership values back through the appropriate output fuzzy membership functions.
- 5) Utilize defuzzification to determine the value for each output variable.
- 6) Make decision according to the output values.

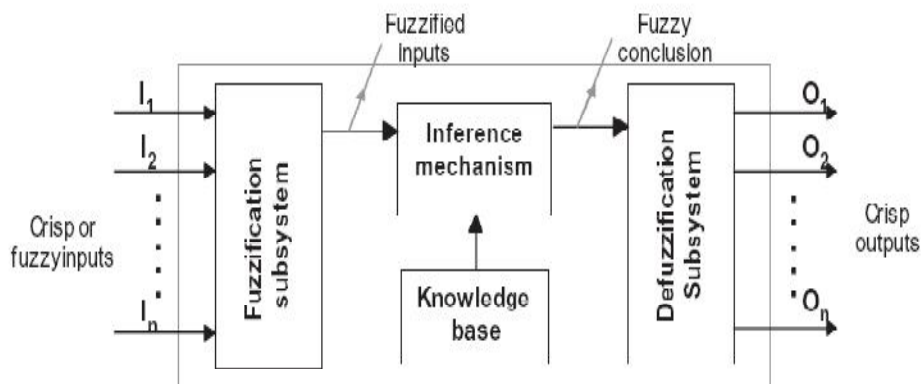


Figure 2. 11: Basic architecture of a fuzzy expert system (Aly & Vrana, 2006)

2.6.1 FUZZY INFERENCE

This can be defined as a process of mapping from a given input to an output, using the theory of fuzzy set. Fuzzy inference can be seen as an evaluation of fuzzy rules to produce an output for each rule (Kosko,1992; Wang & Mendel, 1992). There are different fuzzy inference mechanisms namely:

- Mandani Fuzzy model
- Surgeno Fuzzy model
- Tsukamoto model
- Larsen model

The most commonly used fuzzy inference technique is the so-called **Mamdani** method. The Mamdani-style fuzzy inference process is performed in four steps:

- **Fuzzification:** definition of fuzzy sets, and determination of the degree of membership of crisp inputs in appropriate fuzzy sets.
- **Rule Evaluation:** evaluation of fuzzy rules to produce an output for each rule.
- **Aggregation:** aggregation or combination of the outputs of all rules.
- **Defuzzification:** computation of crisp output.

(a) Fuzzification

Fuzzification is the process of changing a real scalar value into a fuzzy value. This is achieved with the different types of fuzzifiers. Fuzzification is the first step in the fuzzy inferencing process. This involves a domain transformation where crisp inputs are transformed into fuzzy inputs. For instance in medical domain Crisp inputs are exact inputs determined during the laboratory test such as systolic body temperature, age, cholesterol level, etc. and passed into the control system for diagnosis. Each crisp input that is to be processed by the fuzzification inference unit has its own group of membership functions or sets to which they are transformed. This group of membership functions exists within a universe of discourse that holds all relevant

values that the crisp input can possess. Figure 2.12 shows the structure of membership functions within a universe of discourse for a crisp input.

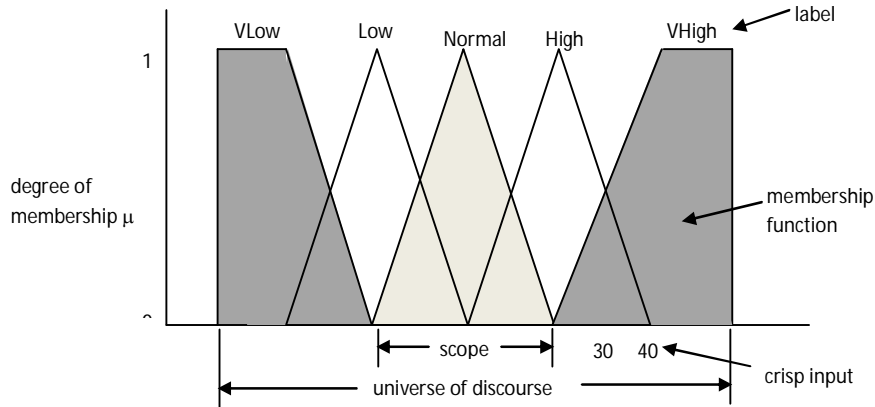


Figure 2.12 Structure of MFs within a universe of discourse for a crisp input.

(b) Rule Evaluation

The second step is to take the fuzzified inputs and apply them to the antecedents of the fuzzy rules. If a given fuzzy rule has multiple antecedents, the fuzzy operator (AND or OR) is used to obtain a single number that represents the result of the antecedent evaluation. This value is then applied to the consequent membership function. To evaluate the disjunction of the rule antecedents, the OR fuzzy operation is performed, such that:

$$\cup () = \max[(), ()] \tag{2.23}$$

Similarly, in order to evaluate conjunction of the rule antecedents, the AND fuzzy operation is performed, such that:

$$\cap () = \min[(), ()] \tag{2.24}$$

The result of the antecedent evaluation can be applied to the membership function of the consequent. This is called rule implication. The most common method of

correlating the rule consequent with the truth value of the rule antecedent is to cut the consequent membership function at the level of the antecedent truth. The example in Figure 2.13 illustrates the evaluation process better according to Negnevitsky (2005).

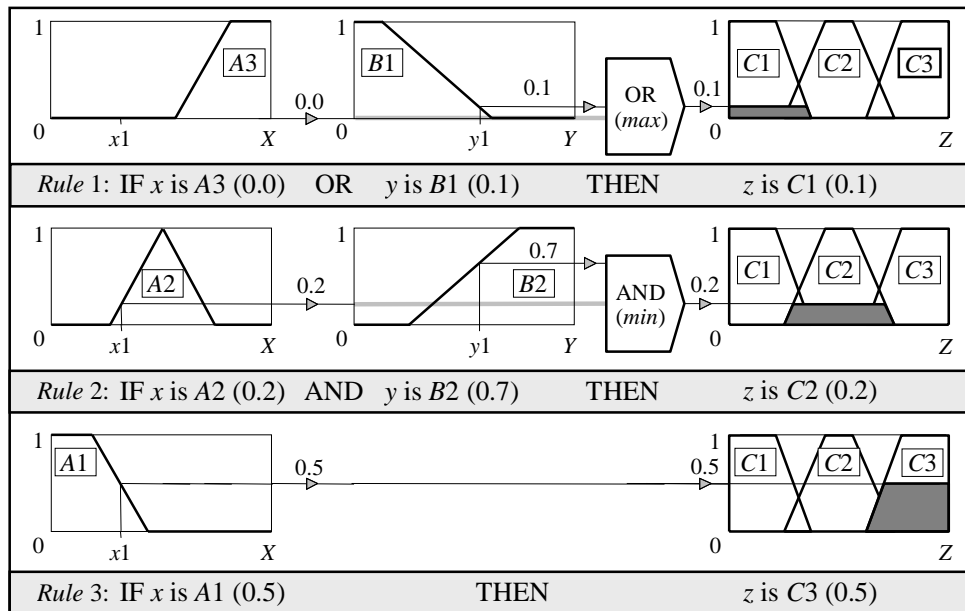


Figure 2.13 Mamdani-style rule evaluation

This method is called **clipping**. Since the top of the membership function is sliced, the clipped fuzzy set loses some information. However, clipping is still often preferred because it involves less complex and faster mathematics, and generates an aggregated output surface that is easier to defuzzify. While clipping is a frequently used method, scaling offers a better approach for preserving the original shape of the fuzzy set. The original membership function of the rule consequent is adjusted by multiplying all its membership degrees by the truth value of the rule antecedent. This method, which generally loses less information, can be very useful in fuzzy expert systems. Figure 2.14 gives a diagram representation of clipping and scaling.

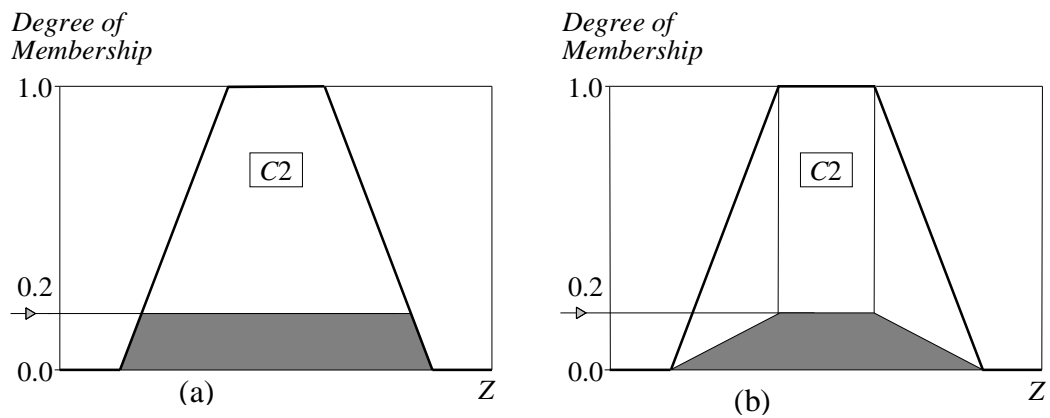


Figure 2.14 Clipped (a) and scaled (b) membership functions

(c) Rule Aggregation

Aggregation is the process of unification of the outputs of all rules. The membership functions of all rule consequents previously clipped or scaled are combined into a single fuzzy set. The input of the aggregation process is the list of clipped or scaled consequent membership functions, and the output is one fuzzy set for each output variable. An example of a rule aggregation process is shown in Figure 2.15.

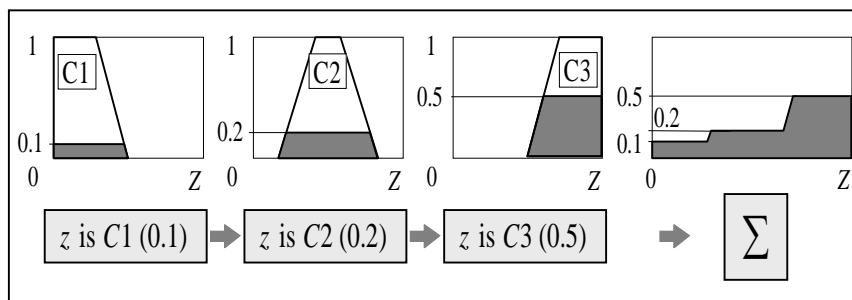


Figure 2.15: Aggregation of rule consequents

(d) Defuzzification

The last step in the fuzzy inference process is defuzzification. Fuzziness helps to evaluate rules, but the final output of a fuzzy system has to be a crisp number. The input for the defuzzification process is the aggregate output fuzzy set and the output is

a single number. There are several defuzzification method, but the most popular one is COG (Center of gravity) simply called centroid technique. It finds the point where a vertical line would slice the aggregate set into two equal masses. This can be represented with this mathematical expression.

$$= \frac{\int f(z) dz}{\int f(z)} \quad (2.25)$$

Centroid defuzzification method finds a point representing the centre of gravity of the fuzzy set, A , on the interval, ab . According to Negnevitsky (2005), in theory, the COG is calculated over a continuum of points in the aggregation output membership function, but in practice, a reasonable estimate can be obtained by calculating it over a sample of points, then the following formular is applied:

$$= \frac{\sum f(z)}{\sum 1} \quad (2.26)$$

This can be illustrated by the example in Figure 2.16

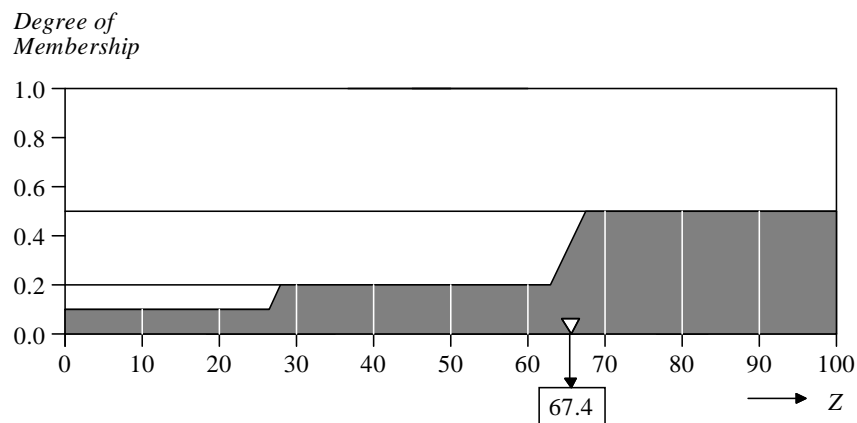


Figure 2.16: Defuzzifying the solution variable's fuzzy set

$$COG = \frac{(0+10+20) \times 0.1 + (30+40+50+60) \times 0.2 + (70+80+90+100) \times 0.5}{0.1+0.1+0.1+0.2+0.2+0.2+0.2+0.5+0.5+0.5+0.5} = 67.4$$

2.7 QUANTITATIVE MEASURES OF A FUZZY EXPERT SYSTEM

Quantitative measures are essential and form the basis for making reliable decisions in software engineering such as fuzzy expert systems (FESs). Quantitative assessment helps us to evaluate the quality of a FES that is not accessible to our intuitive ability. Generally, in constructing a FES, an accuracy measure is a goodness measure that is usually concerned. The accuracy measure implies how good a FES can perform. However, accuracy alone may not be sufficient to show the goodness of FESs (Setnes et al., 1998; Jin, 2000 and Roubos & Setnes, 2001). A comprehensibility measure is an additional quantitative assessment that indicates whether a FES is understandable. Added to these is a completeness measure which is an indicator to whether linguistic variables and rule structure of a fuzzy system cover the entire possible data domain (Jin, et al., 1999, Stamou & Tzafestas, 1999). The three quantitative assessment measures are discuss in the following section.

2.7.1 Accuracy measure

Accuracy is a measure of a predictive (risk determinant) model that reflects the number of times that the model is giving correct classification (output) when it is applied to test data. It measures the probability that the system can correctly determine the risk value of the data. The accuracy measure (AC) can be determined from the following equation.

$$= \frac{\text{Number of correct classifications}}{\text{Total number of test cases}} \quad (2.27)$$

2.7.2 Comprehensibility Measure

Comprehensibility of fuzzy systems involves three important measures: the compactness of fuzzy systems, the similarity between linguistic terms, and the inconsistency of fuzzy rules.

A. *Compactness:*

A compact fuzzy system implies that the fuzzy systems are easy to comprehend. Compactness of fuzzy systems relates to three aspects: a small number of linguistic terms in each dimension, a small number of fuzzy rules in the rule-base, and a small number of conditions in the rule premise or antecedent part. In the first instance, in a scenario where the same fuzzy variable is divided into seven linguistic labels (Extremely Low, Very Low, Low, Medium, High, Very High, Extremely High) and three linguistic labels (Low, Medium, High), it is obvious that the second has fewer linguistic terms. It is relatively easier for the users to discern a fuzzy variable with three than seven linguistic labels.

The second aspect of compactness is the number of fuzzy rules. In a standard structure of a fuzzy system with M dimensions and each dimension partitioned into N subspaces, there exist up to N^M rules in the fuzzy system. For example, a four-dimensional fuzzy system for risk determination has three of the dimensions partitioned into 3 subspaces each and the fourth dimension has 4 subspaces, the number of fuzzy rules would be 108 using standard rule-base formulation. According to Meesad (2001), if all the possible rules are used then the system is not compact. For the same fuzzy system, a more compact fuzzy system could be designed. A compact rule set is easier to comprehend and recognize. Compactness of fuzzy rules is more important when the system has a large number of dimensions (Jin, et al., 1999 ; Wang & Mendel, 1992).

The third aspect of compactness is the number of conditions in the antecedent part of fuzzy rules or the number of features used per rule. If some of the features are not used in rules then the system becomes more compact. The system structure can be easier to comprehend. The compactness of a fuzzy system can be quantified into numerical values as follows.

$$= \frac{h}{h} \quad (2.28)$$

$$= \frac{\dots}{\dots} \quad (2.29)$$

$$= \frac{\dots}{\dots} \quad (2.30)$$

where N_R is the number of rules; N_A is the number of antecedents per rule; N_L is the number of linguistic labels per dimension; and M is the number of dimensions.

B. Linguistic similarity:

Similarity measure for fuzzy sets is used to quantify the comprehensibility of fuzzy knowledge-base. The degree of linguistic similarity is considered the highest when two fuzzy sets are equal. When there are no overlapping fuzzy sets, the degree of linguistic similarity is zeros. The degree of linguistic similarity falls in [0, 1], if there are overlapping fuzzy sets. The degree of linguistic similarity (LS) of a fuzzy system can be determined by the following equations:

$$\left(\dots \right) = \frac{\sum [\dots]}{\sum [\dots]} \quad (2.31)$$

$$= \frac{\dots}{\dots} \sum \sum (\dots); \quad (2.32)$$

f \neq ; $= 1, \dots,$

$$= -\sum \quad (2.33)$$

where $(s_{ij}, s_{kj}) \in [0, 1]$ is the degree of linguistic similarity between linguistic labels s_{ij} and s_{kj} ; k_1 and k_2 are the indexes to linguistic labels; and $\bar{s}_j \in [0, 1]$ is the average of the degree of linguistic similarity in the j th dimension.

C. Inconsistency of Fuzzy Rules:

Inconsistency of fuzzy rules can directly affect the overall decision-making of the system. It can degrade the overall performance of the system. Inconsistency of fuzzy rules should be avoided. Inconsistency of fuzzy rules occurs when there are two or more rules conflicting. Fuzzy rules are conflicting if they have similar antecedents but rather different consequents. Measuring rule inconsistency is equivalent to measuring rule similarity. Degree of fuzzy rule similarity can be measured by using fuzzy similarity measure. Fuzzy rule similarity (RS) is divided into two parts: the similarity of the antecedents (SA) and the similarity of the consequents (SC). The similarity between the j th antecedents of the i th rule and the k th rule ((s_{ij}, s_{kj})) can be determined from the following equation:

$$(s_{ij}, s_{kj}) = \frac{\sum_{j=1}^h [s_{ij} \wedge s_{kj}]}{\sum_{j=1}^h [s_{ij} \vee s_{kj}]} \quad (2.34)$$

Using constant numbers as consequents, the similarity between the consequents of the i th rule and the k th rule ((c_i, c_k)) can be determined from the following equation:

$$(c_i, c_k) = \begin{cases} 1 & \text{if } c_i = c_k \\ 0 & \text{if } c_i \neq c_k \end{cases} \quad (2.35)$$

$$(s_{ij}, c_k) = \frac{1}{h} \sum_{j=1}^h (s_{ij}, c_k) + (s_{ij}, c_k) \quad (2.36)$$

$$\neq ; i = 1, \dots, h-1; k = 2, \dots, h; j = 1, \dots, h$$

$$= \frac{1}{(r_1) (r_2) \dots} \sum \sum (r_i, r_j); \quad (2.37)$$

≠

Where $(r_i, r_j) \in [0,1]$ is the degree of the rule similarity between rules r_i and r_j ; and $RS \in [0,1]$ is the average of the degree of rule similarity.

2.7.3 Incompleteness Measure

Completeness is a property of deductive systems that has been used in the context of artificial intelligence to indicate that the knowledge representation scheme can represent every entity within the intended domain. In a fuzzy system, completeness is a fundamental issue since complete fuzzy systems can respond to any given input. A complete fuzzy system can achieve a proper operation avoiding undesirable situations (Stamoun and Tzafestas, 1999, Oliveira, 1999). The completeness of fuzzy systems consists of two main factors: completeness of fuzzy partitions and completeness of fuzzy rule structure (Jin et. al,1999). Suppose input variable x in the universe of discourse X is divided into N fuzzy partitions represented by membership functions $\mu_i(x)$, for $i = 1 \dots N$. The completeness of the system is satisfied if

$$\forall x \in X, \exists i : 1 \leq i \leq N \text{ such that } \mu_i(x) > 0. \quad (2.38)$$

A certain *level of completeness*, δ , rises to the concept of *strong completeness*, as follows:

$$\forall x \in X, \exists i : 1 \leq i \leq N \text{ such that } \mu_i(x) > \delta. \quad (2.39)$$

A completeness measure of a fuzzy rule structure is defined as the proportion of the complete region and the region of interest. Similarly, an incompleteness measure is defined as the proportion of the incomplete region and the region of interest. Completeness degree in the j th dimension (CD_j) and incompleteness degree in the j th dimension (ID_j) are calculated from the following equations:

$$= \frac{\sum_{j=1}^M \epsilon_j}{M} = \frac{\sum_{j=1}^M (1 - ID_j)}{M} \quad (2.40)$$

$$= 1 - \frac{\sum_{j=1}^M ID_j}{M} \quad (2.41)$$

$$= \frac{\sum_{j=1}^M (1 - ID_j)}{M} \quad (2.42)$$

where ID is the overall incompleteness degree which is the average values of all the incompleteness degrees from each dimension; M is the number of the dimensions. CD_j and $ID_j \in [0, 1]$ are completeness degree and incompleteness degree, respectively, in the j th dimension; CR_j is the length of the complete region in the j th dimension; IR_j is the length of the incomplete region in the j th dimension; and RI_j is length of the region of interest in the j th dimension or the universe of discourse X . $x \in X$ is the input elements. N_x is the number of element x . $m(x)$ the membership degrees of x . $\delta \in [0, 1]$ is the *level of completeness*.

2.8 KNOWLEDGE ACQUISITION AND THE BUILDING OF EXPERT SYSTEMS

According to Feigenbaum (1977) the power of an expert system is the knowledge it possesses. This indicates that most of the emphases in developing expert systems should go to the knowledge-acquisition part of the building process. Hayes-Roth et al., (1983), in classical book on expert-system building describe the expert system building process as the process of knowledge acquisition (Buchanan et al.,1983). The process of building an intelligent system is called knowledge engineering (KE). It has six basic phases as shown in Figure 2.17 (Waterman, 1986; Durkin, 1994):

1. Problem assessment
2. Data and knowledge acquisition

3. Development of a prototype system
4. Development of a complete system
5. Evaluation and revision of the system
6. Integration and maintenance of the system

Turban and Aronson (2001) in their book summarized the six phases of KE process into five phases. This includes knowledge acquisition, validation, representation, inferencing and explanation. Duan et al., (2005) extended the process to include evaluation, implementation, and maintenance as depicted in Figure. 2.18

- Knowledge acquisition. The extraction of domain knowledge from identified sources, such as human experts, books, documents, WWW, sensors, etc.
- Knowledge validation. It is validated and verified against test cases until its quality is acceptable.
- Knowledge representation. The preparation of a knowledge map and encoding of the knowledge in the knowledge base.
- Inferencing. The design of software to allow the computer to make inferences based on the knowledge and the specifics of the problem.
- Explanation and justification. The design and programming of an explanation capability; a Program that allows the system to answer questions about a specific piece of information or how a certain conclusion was derived.

(a) Problem assessment

During this phase the problem's characteristics are determined, the project's participants are determined, the project objectives are specified and the resources needed for the building of the expert system are determined. To characterize the problem, we need to determine the problem type. The problem type may be diagnosis, prediction or risk determination and so on. The problem type influences the choice of tool for building the expert system. Two critical participants are important to be identified, the knowledge engineer and the domain expert. The knowledge engineer should be someone capable of designing, building and testing the expert system. Also the domain expert should be a knowledgeable person capable of solving problems in the problem domain. At this stage the system objectives should clearly specify and determine the resources that would be needed for building the system.

(b) Data and knowledge acquisition

During this phase necessary data and knowledge for building the system are collected and analyzed. The relevant data is identified, extracted and stored appropriately. The choice of the building tool depends on the acquired data. The knowledge that is contained in the system determines the effectiveness of the ES (Feigenbaum, 1981). So this makes this stage more crucial and important in knowledge engineering process. This phase is difficult and time consuming (De Kock, 2003). Knowledge acquisition will be discussed in detail in the next subsection.

(c) Development of a prototype system

At this stage, a small version of the target system is created and tested with a number of test cases. A test case is problem successfully solved in the past for which input

data and an output solution are known. During testing, the system is presented with the same input data and its solution is compared with the original solution. The domain expert takes an active part in testing the system, and as a result becomes more involved in the system's development.

(d) Development of a complete system

At this stage, a plan is developed; schedules and budget for the complete system are developed as soon as the prototype is functioning well. Also, the database and the knowledge-base are populated with complete data and knowledge respectively for the final system.

(e) Evaluation and revision of the system

An expert system is usually designed to solve a particular problem that might not have yet or no solution unlike the conventional program. So, to evaluate an ES, one needs to assure that the system performs intended task to the user's satisfaction. A formal evaluation of the system is normally accomplished with test cases selected by the user. This process of evaluation focuses only on the ES accuracy. According to Meesad (2001), to construct a fuzzy expert system (FES) focusing only on its accuracy without considering the comprehensibility may result in a system that is not easy to understand. Therefore it is important to also measure the ES comprehensibility, even when the accuracy is maintained.

(f) Integration and maintenance of the system

This involves integrating the system into the environment where it will operate and establish an effective maintenance program.

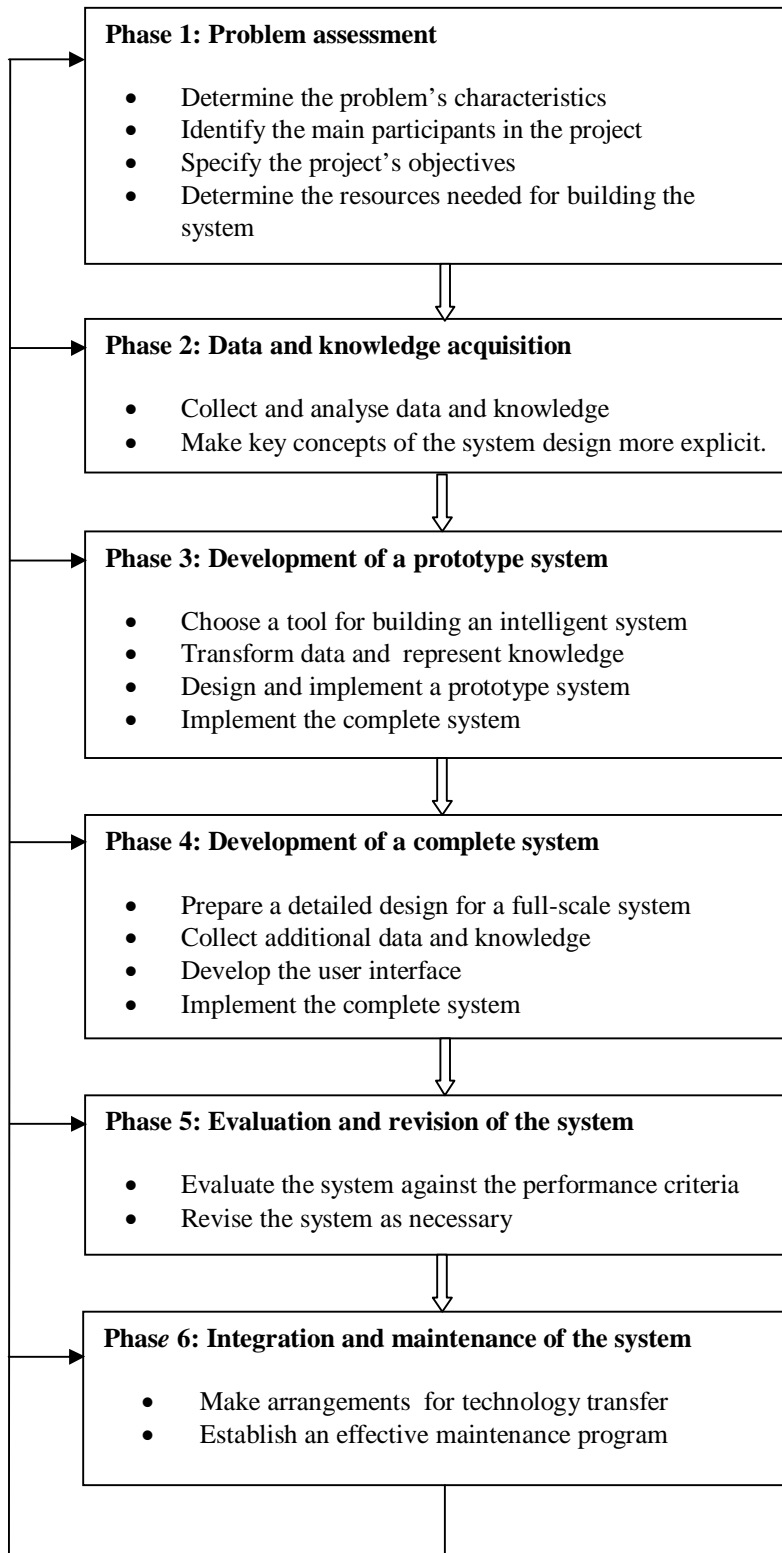


Figure 2.17 The process of knowledge engineering (Negnevitsky,2005)

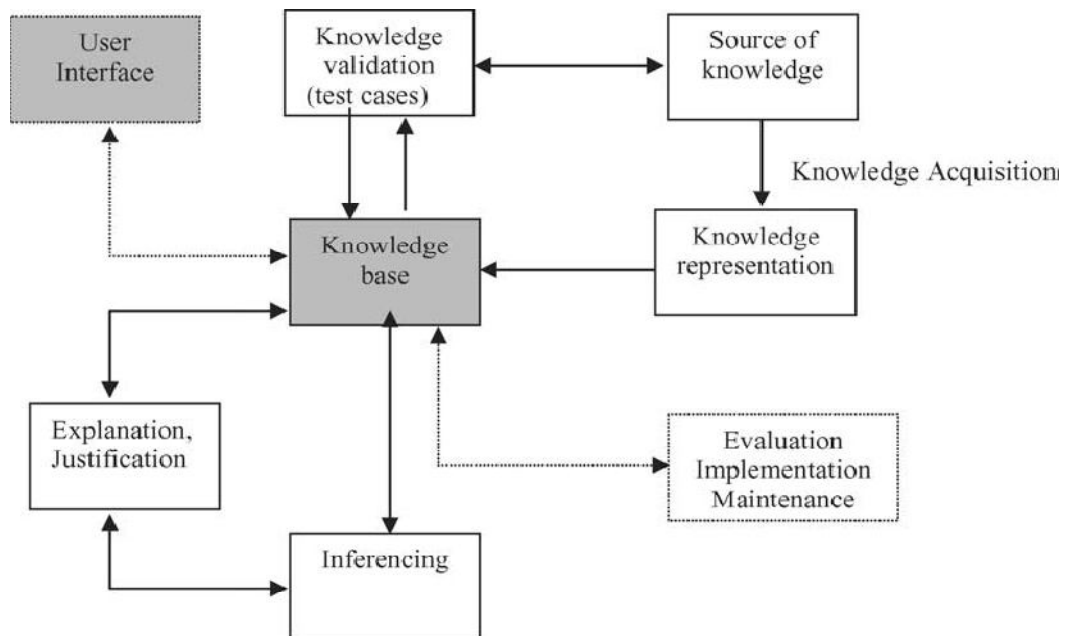


Figure 2.18: Extended process of knowledge engineering (Duan et al., 2005)

2.8.1 KNOWLEDGE ACQUISITION

The knowledge acquisition is part of the knowledge engineering processes. It is the process of acquiring knowledge from a human expert for an expert system, which must be carefully organized into IF-THEN rules or some other form of knowledge representation. The knowledge acquisition represents the extracting, structuring and organizing process of knowledge, out of one or many sources, so that the solving expertise of a matter must be stored in an expert system, in order to be used in solving the issues (Pirnau & Maiorescu, 2008). It is the process of gathering the relevant information from an expert (De Kock, 2003). The method of knowledge acquisition can be divided into manual, semi-automated and automated (De Kock, 2003; Pirnau & Maiorescu, 2008).

a) Manual methods

The manual process is used to deductively extract vital information from the domain expert. The primary manual approach is interview, ranging from complete unstructured to highly structured interview (De Kock, 2003). Interview is the oldest and most usable method of extracting/collecting information from the experts. The unstructured interview is used when the knowledge engineer wants to explore a certain matter. In this manner, the expert has the possibility to answer the questions spontaneously. The structured interview is used when the engineer wants a particular knowledge element, it is an interview orientated on the purpose. The semi-structured interview involves the cognician to ask some certain questions regarding the interest domain and allows the expert to give answers base on the expert's knowledge (Partridge, 1992). Besides the interview process, Pirnau, & Maiorescu (2008) also identified observation during the working process, the brainstorming, the repertoire grid, protocol analysis, nominal

group method, Delphi method and blackboard method, and some other means of extracting vital information from the domain expert.

b) Semi-Automated methods

In semi-automated knowledge acquisition the roles of the expert and the engineer are minimized in the process of knowledge acquisition. They are grouped in two main categories:

- (i) methods that support experts in building a knowledge-base, without cognicians' help, in categorization and implementation phases.
- (ii) methods that support cognicians in executing the specific phases of the knowledge acquisition rapidly and efficiently with less help of the experts.

c) Automated knowledge acquisition

Automated knowledge acquisition uses an induction system with case histories and examples as input to derived knowledge-base. This is also known as machine learning. Automated knowledge acquisition eliminates the role of knowledge engineer and minimizes the role of domain expert in the knowledge extraction (Turban, 1993). In most existing rule-based expert systems, the knowledge-base rules are generated by experts in the area, especially for control problems with only a few inputs. With an increasing number of variables, the possible number of rules for the system increases exponentially, which makes it difficult for experts to define a complete rule set for good system performance (Shi, 1999). An automated way to design fuzzy systems might be preferable (Shi, 1999). Also automated method of knowledge acquisition becomes more important especially where the domain expert is not available and there are case histories and examples.

Different techniques have been introduced in the literature to actualize the automated knowledge acquisition approach. These include clustering (Shah, 2006), classification (Harleen & Siri, 2006; Gadaras & Mikhailov, 2009), neural network (Shi, 1999; Neshat, & Yaghibi, 2009), hybrid system of fuzzy and neural (Chirstoph, 1995, Moein et al., 2008;) rough set (Setiawan et al., 2009), and fuzzy evolutionary (Koutsojannis & Hatzilygeroudis, 2006). In most other cases, such as in (Allahverdi et al., 2007) rules were generated by standard rule-base formulation as described in (Meesad, 2001). In Norbik & Bharanidharan, (2008) the Fuzzy logic concept and data mining approach was implemented to improve intrusion detection system. The improved Kuok fuzzy data mining algorithm which modified apriori algorithm was used to generate the fuzzy rules for the knowledge-base that reflect common way of describing security attacks. The report shows that the approach performed efficiently based on data driven-approach. Because of the peculiarity of the medical expert systems some approaches that have been used to evolve knowledge-base are not free from sharp boundary problem, which could either overestimate the boundary values or underestimate them because they are based on data driven approach or quantitative binary partition (Oladipupo et. al., 2010). Some exhibited rule inconsistency, membership function not corresponding with the intuitive human perception and large number of rules in the knowledge-base. All these shortcomings intimate the context of this research where the fuzzy association rule mining expert-driven knowledge acquisition approach is proposed.

2.9 STATE-OF-THE-ART IN MEDICAL FUZZY EXPERT SYSTEM KNOWLEDGE ACQUISITION RESEARCH

In medical domain, statistics has consistently shown that coronary heart disease is one of the leading causes of death all over the world including African continents (Neshat, & Yaghobi 2009). Many people had fallen victims of such death because they lack knowledge of their heart disease risk status. However, many lives could be saved if an adequate fast response risk determination expert system is made available for people in order to know their status. Coronary Heart Disease (CHD) is a narrowing of the small blood vessels that supply blood and oxygen to the heart. This is also called coronary artery disease. In the domain of heart disease risk, smoke, cholesterol, blood pressure, diabetes, sex and age are the main risk factors that determine heart disease risk (Adeli & Neshat, 2010). In order to reduce the overhead cost of looking for experts in this domain, expert system was introduced for diagnosis and risk determination. This evolves knowledge from human experts and existing knowledge to solve related problems (Feigenbaum, 1982; Abraham, 2005).

In the literature, different approaches have been used to evolve knowledge for fuzzy expert system. An evolutionary fuzzy system was presented by Shi et al., (1999). A hybrid fuzzy-neural based medical diagnosis system was proposed in Moein et al., (2008) and Christoph, 1995. Classification based data mining was used by Harleen & Siri, (2006); Gadaras and Mikhailov, (2009). In some other cases, rules were generated by standard rule-base formulation (Allahverdi et al., 2007; Seritas et al., 2003). All the generated rules were used to build the fuzzy expert system. Though the system maintains its accuracy and coverage but not compact (Meesad, 2001).

Another work was proposed by using multi layer perceptron to build decision support system for the diagnosis of five major heart diseases (Yan et al., 2006). Also in 2009, neural network was used to generate rules for Hepatitis B intensity rate (Neshat, &

Yaghobi, 2009). Research work on diagnosis of Coronary Artery Disease was also carried out using rough set theory (RST). The rules were selected and fuzzified based on information from discretization of numerical attribute (Setiawan et al., 2009). Adeli & Neshat recently designed a fuzzy expert system for heart disease diagnosis (Adeli and Neshat, 2010). Eleven input linguistic variables and one output linguistic variable were used for disease diagnosis. In all, some of these proposed approaches suffer from sharp boundary problem and inconsistency rules. In some other cases the systems were based on data-driven approach. This makes the membership functions not to correspond with experts' perception, and more importantly, it results in the knowledge-base unwieldiness as a result of a large number of rules in the knowledge-base. All these deficiencies form the basis for this research in order to ensure a comprehensible medical fuzzy expert system with accuracy.

2.10 DATA MINING IN KNOWLEDGE ACQUISITION

Data mining is an Artificial Intelligence (AI) technique for discovery of knowledge in large databases, that could be used to collect hidden information for medical purposes (Siti & Miswan, 1999; Siti & Rogayah, 1999; Neves *et al.*, 1999). According to Delgado et al., (2001) the increase in database size makes traditional manual data analysis to be insufficient. To extract important information from such large databases new research fields such as knowledge discovery in databases (KDD) have rapidly grown in recent years. KDD is concerned with the efficient computer-aided acquisition of useful knowledge from large sets of data. The main step in the knowledge discovery process, called data mining, deals with the problem of finding interesting regularities and patterns in data. One of the main objectives of data mining methods is to provide a clear and understandable description of patterns held in data (Delgado, et al., 2001).

Nowadays, data stored in medical databases are growing in an increasingly rapid way. The discovery of new knowledge by mining medical databases is crucial in order to make an effective use of stored data, for enhancing medical decision making and improving the performance of patient management task (Lavrac, 1996).

There are different data mining techniques that are capable of analyzing and extracting previously unknown hidden pattern from historical database. The nature of the existing database and description of the expected pattern determine the best data mining technique that can be used to extract patterns. One of the best studied models for pattern discovery in the field of data mining is that of association rules (Agrawal, 1993). Association rule mining with fuzzy logic concept has the capability of analyzing and extracting medical database because of the quantitative nature of the medical database (Delgado et al., 2001; He et al., 2006).

2.10.1 CRISP-DM Model

Conceived in 1996, the CRISP-DM (Cross Industry Standard Process for Data Mining) model has evolved as the standard for conducting data mining activities. At that time, many different data mining approaches had been developed and therefore there was a great need for a unified framework. CRSP-DM emerged as a freely available and non-proprietary framework with a standardized process (see Figure 2.19).

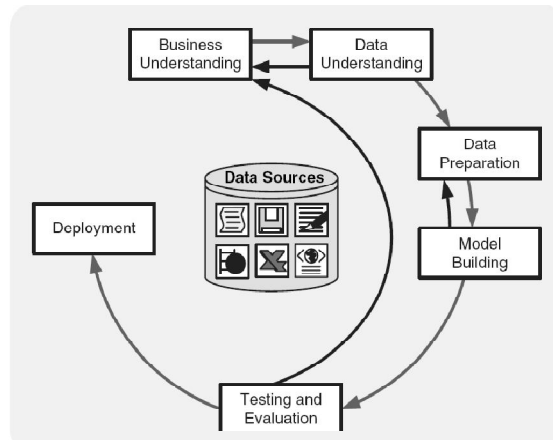


Figure 2.19: The CRISP-DM Model (Chapman et al., 1999)

The model defines six phases to conduct a data mining project which are:

1. Business Understanding

Business understanding is the initial phase of the CRISP-DM model. Most importantly, it focuses on the objectives and requirements of a project from the business perspective. The actual situation in the company is assessed. After this assessment, the acquired knowledge is converted into the data mining problem definition which is a plan that advises the data mining how to deal with these objectives and requirements of the project.

2. Data Understanding

The second phase starts off with a collection of all the available data that might be relevant for the mining project, followed by activities like describing and exploring the data in order to get familiar with them. Another important task is to verify the quality of the data which makes efficient mining possible. This phase helps the participants in getting first insights into the data set.

3. Data Preparation

In the data preparation phase, the relevant data is selected from the overall data set discovered in the previous phase. In a second step, the initial data has to be cleaned,

later integrated and formatted. All those activities aim at constructing the final data set that is adequate for mining. The tasks in this phase are likely to be performed multiple times in order to provide a good data set which is crucial for any data mining project.

4. Modelling

The modelling phase deals with selecting possible models for data mining and calibrating the parameters to optimal values for the specific data mining task. Multiple data mining methods might be adequate for mining all of which should be tested at this step. Some models have specific requirements for the data, making a step back to the preparation phase necessary. Discovered models need to be measured and assessed regarding the data mining goal they have to suit.

5. Evaluation

After having built the models for mining, the degree to which it meets the business objectives needs to be measured. A model may have high quality from a data analysis perspective, but might be deficient in meeting the requirements of the business. To certify the achievement of these goals, the model needs further evaluation. The steps for generating the model are reviewed to assess whether any important task or factor has somehow been overlooked. The phase ends with a decision on the use of the data mining results.

6. Deployment

The process does not end with the creation of a correct model. Instead, the gained knowledge needs to be organized and presented in ways that the customer can understand and use it. In addition, a model has to be monitored and maintained in order to allow future use. A valid model might not be valid at all time because customer behavior changes and thus the model might need adjustments. The complexity of the deployment phase highly relies on the requirements defined at the

beginning of the project. Often, the customer will have to deal with the subsequent steps which will require further explanation by the developer. The end of the process is marked by the generation of the final report including all the previous deliverables and a summarization and organization of the results.

2.10.2 Mining Quantitative Attributes with Association Rule Mining

The discovery of previously unknown, potentially useful and hidden knowledge in databases is called Knowledge Discovery (KD). Data mining is an important process in KD. In data mining, association rule mining (ARM) is an important tool often used to represent and identify dependencies between attributes in a database.

Association rule mining searches for interesting relationships among items in a given dataset. The most popular algorithm for mining rules based on two-valued attribute is APRIORI. This algorithm leads to the problem of categorizing numerical attributes, which the algorithms can only apply to data mining problems with categorical features (He et al., 2006). A quantitative association rule mining algorithm as a solution to this problem was given in (Agrawal & Srikant, 1994) which transforms quantitative variables into a set of binary variables through partitioning the domain variables into discrete intervals. This approach, however, suffered from “sharp boundary problem” the algorithm ignored or over emphasised the elements near the boundary of the interval in mining process. Also, the use of sharp boundary interval is not intuitive with respect to human perception (Verlinde et al.,2006). An alternative solution, according to Kuok, et al., (1999) is using fuzzy logic. Fuzzy logic has demonstrated to be a superior mechanism to enhance interpretability of discrete intervals (Delgado et al., 2001) and offers a smooth transition from one fuzzy set to another.

Fuzzy association rule mining (FARM) was proposed because of the deficiency of quantitative association rule mining (Kuok et al., 1999; Gyenesei, 2001). FARM is the discovery of association rules using fuzzy set concepts, such that the quantitative attributes can be handled. Fuzzy association rule is more understandable because of linguistic terms associated with the fuzzy sets (Kuok et al., 1999). In constructing fuzzy association rule mining algorithm there are two extreme approaches, which are: data-driven (rules are generated automatically from the data) and expert-driven approach (an expert manually determines the membership functions). The expert-driven approach is considered to be more accurate because it corresponds with the most intuitive human perception since an expert in the application domain will be involved (Verlinde et al., 2006).

In literatures different algorithms have been proposed for mining fuzzy rules (Verlinde et al., 2006; Delgado et al., 2001, Gyenesei, 2001). The final output of the algorithm is a set of rules that meet the confidence and support constraints given as input. These constraints are quantitative qualifiers used to evaluate the relevance of an association rule; support confidence. *Support* of a rule is a measure of the fraction of the entire data set for which all predicate terms of the rule hold true. *Confidence* of a rule is a measure of the fraction of the data set for which, if the antecedent holds true, then the consequence holds true. In Ohsaki et al., (2007) the usefulness of rule interestingness measures for medical KDD through experiments using clinical datasets was discussed and, based on the outcomes of these experiments, how to utilize these measures in post processing was also considered. The nature of most dataset in medical domain are quantitative, and thus makes FARM with expert-driven approach appropriate for this

research where rules are generated with avoidance of sharp boundary problem and corresponds with intuitive human perception. FARM could be formally defined as thus:

The formal definition of fuzzy association rule mining, according to Gyenesei, (2001) was given as :

Given a database $T = \{t_1, t_2, \dots, t_n\}$ with attribute $I = \{i_1, i_2, \dots, i_p\}$, and the fuzzy sets $f_{i_1}, f_{i_2}, \dots, f_{i_p}$ associated with attributes in I . We can evolve fuzzy if-then rule such as:

$$\text{If } X \text{ is } A \text{ then } Y \text{ is } B$$

In the above rule, $X = \{x_1, x_2, \dots, x_p\}$ and $Y = \{y_1, y_2, \dots, y_q\}$ are itemsets. X and Y are subsets of I and they are disjoint which means that they share no common attributes.

$f_{x_1}, f_{x_2}, \dots, f_{x_p}$ and $f_{y_1}, f_{y_2}, \dots, f_{y_q}$ contain the fuzzy sets associated with the corresponding attributes in X and Y . For example, an attribute x_k in X will have a fuzzy set f_{x_k} in A such that $f_{x_k}(t) \in [0, 1]$ is satisfied.

A is a fuzzy set in X and B is a fuzzy set in Y . “ X is A ” is the rule antecedent and “ Y is B ” is the rule consequent. The semantics of the rule is when ‘ X is A ’ is satisfied, we can imply that ‘ Y is B ’ is also satisfied. In this context ‘satisfied’ means there are sufficient amount of records which contribute their votes to the attribute fuzzy set pairs and the sum of these votes is greater than a user specified threshold. An itemset $\langle X, A \rangle$ is said to be frequent if its support value satisfies the minimum support threshold.

The frequent itemset obtained is used to generate all possible rules. If the union of antecedent $\langle X, A \rangle$ and consequent $\langle Y, B \rangle$ has sufficient support and the rule has high confidence, then, the rule is said to be interesting. The measure of support and confidence is used to determine the satisfiability of itemsets and rule. In Kork et al.

(1998), instead of support and confidence the two terms are represented by significance and certainty factor respectively.

Some attempts for developing algorithms to discover fuzzy association rules have already been made. In Chan et al. (1998), an algorithm for mining fuzzy association rules in quantitative databases is proposed. The algorithm, called F-APACS, employs linguistic terms to describe the hidden regularities and exceptions rather than splitting up quantitative attributes into fuzzy sets. The linguistic terms are defined by fuzzy set theory; therefore the association rules discovered here are called fuzzy association rules. An objective interestingness measure is used to define whether attributes are related or not. The use of linguistic terms is an attempt to make rules more understandable for the human user. In traditional association rule mining techniques, minimum support and confidence thresholds have to be defined by the user. The F-APACS algorithm addresses this problem by using adjusted difference analysis to identify interesting associations between attributes. In addition, the algorithm can discover both, positive and negative association rules. A negative rule tells us that if a record has a certain characteristic, the associated record will not have another characteristic. The algorithm starts with a data set. The linguistic terms are represented by fuzzy sets L_{pq} , L_{jk} and the degree to which d is represented by L_{pq} , L_{jk} is summarized in $deg_{L_{pq}L_{jk}}$. The interestingness of an association rule is calculated using the adjusted difference measure.

Another algorithm has been suggested in Chen & Wai (2002), which is suitable for mining association rules in fuzzy taxonomic structures. The Apriori algorithm is extended to allow mining fuzzy association rules as well. Fuzzy support and

confidence measures are applied in order to evaluate the interestingness of a rule. The non-fuzzy algorithm of Srikant & Agrawal (1996), decides whether a transaction T supports an itemset X by checking for each item $x \in X$ if the item itself or some descendant of it is present in the transaction. For this reason, all possible ancestors of each item in T are added, forming T' . Now T supports X if and only if T' is a superset of X . A standard algorithm can then be run on the extended transactions to mine the association rules. In the fuzzy case, T' is generated differently. Not only have the ancestors of T had to be added, but also the degree to which the ancestors are supported by the transactions.

A different attempt has been made in Hen et al. (1999), which similarly uses the Apriori algorithm as a basis but incorporates fuzzy sets for mining quantitative values in a database. The algorithm first transforms each quantitative attribute into fuzzy sets and maps items to them via membership functions. An Apriori-like algorithm generates the association rules using the previously collected fuzzy counts.

Another Apriori-like approach is presented in Gyenesi (2000). It addresses the two main steps of association rule mining, namely the discovery of frequent itemsets and the generation of association rules from quantitative databases. The notation in Table 2.1 were used for the algorithm.

Table 2.1 Apriori algorithm notation (Gyenes, 2000)

D	the database
D_f	the transformed database
F_k	set of frequent k -itemsets (having k items)
C_k	set of candidate k -itemsets (having k items)
I	complete itemset
$minsup$	support threshold
$minconf$	confidence threshold
$mincorr$	correlation threshold

The algorithm first searches the database and returns the complete set containing all attributes of the database. In a second step, a transformed fuzzy database is created from the original one. The user has to define the sets to which the items in the original database will be mapped. After generating the candidate itemsets, the transformed database is scanned in order to evaluate the support, and after comparing the support to the predefined minimum support, the items with a too low support are deleted. The frequent itemsets F_k will be created from the candidate itemsets C_k . New candidates are being generated from the old ones in a subsequent step. C_k is generated from C_{k-1} . The following pruning step deletes all itemsets of C_k if any of its subsets does not appear in C_{k-1} . Finally, the association rules are generated from the discovered frequent itemsets. The pseudocode of the algorithm is depicted in Figure 2.20

```

Main Algorithm(minsup, minconf, mincorr, D)
1)  $I = Search(D)$ ;
2)  $(C_1, D_T) = Transform(D, I)$ ;
3)  $k = 1$ ;
4)  $(C_k, F_k) = Checking(C_k, D_T, minsup)$ ;
5) while ( $|C_k| \neq \emptyset$ ) do
6) begin
7)    $inc(k)$ ;
8)   if  $k = 2$  then
9)      $C_k = Join1(C_{k-1})$ ;
10)  else  $C_k = Join2(C_{k-1})$ ;
11)   $C_k = Prune(C_k)$ ;
12)   $(C_k, F_k) = Checking(C_k, D_T, minsup)$ ;
13)   $F = F \cup F_k$ ;
14) end
15)  $Rules(F, minconf, mincorr)$ 

```

Figure 2.20 An algorithm for mining Fuzzy Association Rules (Gyenesei, 2000)

2.10.3 Quality Measures

(a) Fuzzy Support Value

To generate fuzzy association rule, the first step is to find out all large k-itemsets which are itemsets with fuzzy supports value greater than the minimum specified threshold. The fuzzy support value is calculated by first summing all votes of each record with respect to the specified itemset, then dividing it by the total number of records (Kuok et al., 1999). Each record contributes a vote of number which falls in [0,1]. We can express the fuzzy support value with this mathematical expression

$$\langle \cdot \rangle = \frac{\langle \cdot \rangle}{|\cdot|} \quad (2.42)$$

$$\langle \cdot \rangle = \frac{\sum_{\epsilon \in \Pi} \epsilon}{|\cdot|} \quad (2.44)$$

Where

$$= \begin{cases} 0 & \text{if } h \geq \omega \\ h & \text{if } h < \omega \end{cases} \quad (2.45)$$

In the above equation, $\langle X, A \rangle$ represents the itemset-fuzzy set pairs, where X is the set of attributes x_j and A is the set of fuzzy sets a_j . A record satisfied $\langle X, A \rangle$ means that the vote of the record is greater than zero. The vote of a record is calculated by the membership grade of each x_j in that record. The membership grade should not be less than the user specified threshold ω such that low membership values will not be considered. h is used to obtain the value of x_j in the i th records, then transform the value into membership grade by $\mu_{a_j}(x_j)$ which is the membership function of x_j . After obtaining all membership grades of each x_j in a record, $\prod_{j \in X} \mu_{a_j}(x_j)$ is used to calculate the vote of t_i . After taking the sum of all the vote, then the value is divided by the total number of records $|T|$. Besides the multiplication operator (mul, Π) other operators like min and max can also be used. The Mul operator provides the simplest and reasonable results, especially when the fuzzy transactions are not normalized. Mul is more suitable because it takes the degree of all items in a transaction into account (Kuok et al., 1999).

(b) Fuzzy Confidence Value

Fuzzy Confidence (FC) Value is the measure of the degree of support given by the transaction. FC is used to estimate the interestingness of the generated rules. Having discovered the frequent itemsets, the support is known and all subsets of the frequent itemset can also be identified, then the fuzzy confidence values FC for a rule

$\langle A, B \rangle \rightarrow \langle C, D \rangle$ where $U = A \cup B$, $U = C \cup D$ is determined by this mathematical expression:

$$C(A, B) = \frac{\langle A, B \rangle}{\langle C, D \rangle} \quad (2.46)$$

$$\langle A, B \rangle \rightarrow \langle C, D \rangle = \frac{\sum_{A \in \Pi} \Pi(A) \cdot C(A, B)}{\sum_{A \in \Pi} \Pi(A)} \quad (2.47)$$

where

$$C(A, B) = \begin{cases} \frac{\sum_{A \in \Pi} \Pi(A)}{h} & \geq \alpha \\ 0 & \text{otherwise} \end{cases} \quad (2.48)$$

(c) Interestingness Measure

A rule can be considered interesting if the fuzzy set union of antecedent and the consequent has enough significance and the rule has adequate certainty. The measure of interestingness other than support and confidence are required in order to evaluate the quality of fuzzy association rules. The quality measure of a rule to be interesting is called certainty factor (Gyenesei, 2000). The certainty factor is determined by computing the fuzzy correlation of antecedent and the consequent of the rule. The Pearson's product-moment correlation coefficient between attributes could be used (Gyenesie, 2001 and Kuok et al., 1999). This correlation is different from the general statistical usage of correlation because in association rule mining $A \rightarrow B \neq B \rightarrow A$. The correlation (X, Y) between two variables X and Y with expected values $E(X)$ and $E(Y)$ and standard deviation σ_X and σ_Y is defined according to Gyenesei, (2000) as:

$$C(A, B) = \frac{\langle A, B \rangle - \langle A \rangle \langle B \rangle}{\sigma(A) \sigma(B)} \quad (2.49)$$

$$\langle (A, B), (C, D) \rangle = \frac{\langle A, C \rangle \langle B, D \rangle}{\langle A \rangle \langle C \rangle \langle B \rangle \langle D \rangle} \quad (2.50)$$

where

$$(\mu, \nu) = [\langle \mu, \nu \rangle] - [\langle \mu, \nu \rangle] \times [\langle \mu, \nu \rangle] \quad (2.51)$$

$$(\mu) = [\langle \mu, \nu \rangle] - [\langle \mu, \nu \rangle] \quad (2.52)$$

$$(\nu) = [\langle \mu, \nu \rangle] - [\langle \mu, \nu \rangle] \quad (2.53)$$

$$[\langle \mu, \nu \rangle] = \frac{\sum_{\epsilon} \prod_{\epsilon}}{|\epsilon|} \quad (2.54)$$

$$= \frac{\epsilon}{0} \geq h \quad (2.55)$$

$$[\langle \mu, \nu \rangle] = \frac{\sum_{\epsilon} [\epsilon]}{|\epsilon|} \quad (2.56)$$

$$[\epsilon] = \prod_{\epsilon} (\epsilon) \geq h \quad (2.57)$$

$$= \prod_{\epsilon} \quad (2.58)$$

The vote of record will be zero if the membership grade of $\langle \mu, \nu \rangle$ in that record is less than h . However the vote of the consequent will also be zero if the vote of the antecedent is less than h .

2.11 SHARP BOUNDARY PROBLEM IN RULE-BASED EXPERT SYSTEM

The sharp boundary problem (SBP) is as a result of the quantitative attributes partitioning strategy where boundary cases are underestimated or overestimated. This consequently affects the accuracy of the expert system (Verlinde et al., 2006)

In the medical domain, the use of rule based expert system has increased greatly because of the scarcity of human experts in the domain and the availability of fast growing databases which could be used to model inferences and discover patterns in form of rules. In real live application, medical databases contain different kinds of attributes such as binary and quantitative attributes (Delgado et al., 2001). Binary takes values from 0 or 1; for instance, a patients smoking status could be 'yes' or 'no'. Quantitative attributes that are categorical, numerical, or non-fractional in nature, take values from an ordered numerical scale, often a subset of the real number (Kuok et al., 1999). Quantitative attributes are very common in medical databases. For example heart disease patients can take age values between 20-79 years, result from laboratory test for systolic blood pressure level could take values within <120 to ≥ 160 mm/Hg while cholesterol measures could be within the range of <160 to ≥ 280 mg/dL.

In building an expert system, quantitative attributes need to be partitioned into ranges because of the very wide range of values defining their domain. There are several approaches to partitioning quantitative attributes as discussed in literature (Han & Kamber, 2001). The partitioning process is referred to as binning; that is, an interval is considered as a "bin". The common binning strategies are: 1) Equiwidth binning, where the interval size of each bin is the same; 2) Equidepth binning, where each bin has approximately the same number of tuples assigned to it; and 3) Homogeneity-based binning, where bin size is determined so that the tuples in each bin are uniformly distributed. Also, there is the Distance based partitioning strategy, which seems most intuitive since it groups quantitative values that are closed together within the same interval (Han & Kamber, 2001). All of these partitioning strategies are subject to sharp boundary problem because of the classical set theory (Kuok et al., 1999). However, to

prevent this problem, in Allahverdi et al., (2007), fuzzy logic concept was introduced into a rule-based expert system to determine coronary heart disease risk. The design gives the user the risk ratio and most of the experimented test data risk ratio from the fuzzy approach was reported to give relatively the same percentage risk as Adult Treatment Panel III (ATP III) calculation. This reflects the extent to which fuzzy concept was able to prevent sharp boundary problem. In this thesis a comparative study would be undergone to investigate the effect of SBP on quantitative binary partition strategy and fuzzy partition strategy in building a medical rule-base expert system. This will be based on expert-driven approach of data partitioning (Verlinde et al., 2006).

2.12 THE CONTEXT OF THIS RESEARCH

From the foregoing, a number of gaps exist in literature which defines the context of this research. The first is the need for acquiring a knowledge-base that will emulate human perception of medical concept in order to avoid sharp boundary problem which has not been adequately addressed by existing medical fuzzy expert systems. The second is the problem of large number of rules in the knowledge-base to which literature has not been able to provide a fuzzy association rule mining with incorporation of expert's opinion (expert-driven approach) solution to the best of our knowledge. These two gaps become the premise for the central research question being investigated in this thesis, which is: How do we facilitate a complete and comprehensible knowledge-base in medical fuzzy expert system that will emulate human perception, void of sharp boundary problem and solve the problem of system unwieldiness while accuracy is still gained.

For adequate explication, the central question has been split into the following two research questions:

1. How do we acquire a knowledge-base that will emulate human perception of application domain concept in order to avoid sharp boundary problem? And
2. How can an ES developer develop a comprehensive fuzzy rule-based expert system which eliminates redundant rules in order to solve the problem of rule-base unwieldiness and provides for knowledge-base update?

This thesis aims at proposing a viable solution to these questions

2.13 SUMMARY

The Chapter presents the issues that define the research context of this thesis. It started with a discussion of the necessity for expert systems and the progress made so far in building an expert systems. Secondly, an argument for fuzzy concept in medical rule-based expert system is presented as justified by the life threatening nature of medical decision making consequences and the quantitative nature of medical data which makes possible for sharp boundary problem. This is followed by an overview of fuzzy set theory, characteristics, operations and definition of fuzzy terms. Thereafter, the chapter specifically reviewed fuzzy rule-based expert system, taking a survey of medical expert systems, the limitations of existing approaches and the gaps that this thesis attempts to fill. Next was the subject of knowledge acquisition and building of expert systems. This chapter also identified the need for automated knowledge acquisition in building a complete and comprehensible medical expert system. After this a painstaking review of the state-of-the-art in medical fuzzy expert system knowledge acquisition was carried out taking heart disease as a case study. Fuzzy association rule mining expert-driven approach was identified as a possible means of

acquiring a knowledge-base with limited number of rules that will correspond with human perception of the domain concept. The chapter closed by establishing existing fuzzy association rule mining techniques with incorporation of domain experts' opinion factors as a competent tool for knowledge extraction, especially in medical domain and by formally articulating the research context of this thesis.

CHAPTER THREE

FUZZY ASSOCIATION RULE MINING EXPERT-DRIVEN (FARME-D) APPROACH TO KNOWLEDGE ACQUISITION

3.1 INTRODUCTION

The **Fuzzy Association Rule Mining Expert-Driven (FARME-D)** approach to knowledge acquisition is the proposed solution to the two research questions posed in this thesis. This chapter presents an overview of FARME-D as an approach to knowledge acquisition where experts' opinion factors are incorporated into the existing fuzzy association rule mining. The proposed approach is committed to the extraction of interesting knowledge from domain experts' past experiences based on experts' perception of the data. Instead of using data-driven approach (where data partitions and rules consequences are generated automatically from the data table) this proposed solution uses expert-driven approach where fuzzy interval partitions, membership functions calibration and rules consequences are determined by the domain experts. This is used for modelling a comprehensive fuzzy rule-based expert system where the system rules correspond to human expert perception of decision making. The chapter provides insight to the proposed solution strategies and underlining assumptions, the structure of FARME-D integration with FES standard architecture, and its main sub-processes. Also, the modalities for the validation of the FARME-D approach are discussed. The chapter closes with a summary and discussion on expected results.

3.2 OVERVIEW OF THE PROPOSED SOLUTION: FARME-D APPROACH

FARME-D is an automated knowledge acquisition approach which incorporates domain experts' opinion factors into the existing fuzzy association rule mining process (domain expert determines fuzzy interval partitions, membership functions calibration and rules consequences). This is committed to modelling of Fuzzy Rule-Based Expert System simply called Fuzzy Expert System (FES). It is a specialized pattern discovery technique that involves domain expert's opinion, excels in extracting interesting knowledge in form of rules which correspond to the domain expert perception and void of sharp boundary problem. FARME-D enhances FES comprehensibility while accuracy is maintained; it also aims at providing a platform that enhances instant update of the knowledge-base in case new knowledge is discovered. The integration of FARME-D with FES standard architecture would give birth to new fuzzy rule-based expert system architecture called Fuzzy Association Rule Mining Expert System (FARMES). FARME-D is proposed as a solution to the two research questions that have been highlighted in this thesis. It is designed as an integrated automated knowledge acquisition approach to facilitate the modelling of FES in knowledge engineering and enhances knowledge-base frequent updates. Further details on FARME-D approach and FARMES architecture are presented next.

3.2.1 Limitation and Assumptions

The application of FARME-D in modelling fuzzy association rule mining expert systems is constrained by a set of preconditions that guarantees its practicability in knowledge acquisition. These are:

1. Data and technical description of the problem domain is used.

2. In modelling expert systems, the simplicity advantage of production rules knowledge representation is adopted.
3. Only structure historical database is accommodated in the mining process.

In addition, FARME-D is based on the following assumptions:

1. The determinant factors for solving problems are known and predetermined in advance by the domain experts.
2. Data stored in organizations are quantitative in nature and growing in an increasingly rapid way with increasing number of variables.
3. Organizations have historical data bank where the past human experts' experiences could be retrieved.
4. The historical data set are in a structured form.

FARME-D is designed for specialized automated knowledge acquisition for modelling FES. It does not address the entire structure of FES. As such, the limitation and assumptions of FARME-D are all directed from the principle that governs the practice of automated knowledge acquisition and fuzzy association rule mining processes (Pirna & Maiorescu, 2008; Delgado et al., 2001). The limitation is meant to provide a guide on how the knowledge acquired could be managed to enhance the ES knowledge-base. The set assumptions, on the other hand, are those that facilitate the most utilization of fuzzy association rule mining technique and specify the scenario when FARME-D is optimally applicable.

3.3 COMPONENTS OF FARME-D KNOWLEDGE ACQUISITION

FARME-D adopted the Cross Industry Standard Process for Data Mining model framework for the mining process explained in section 2.12.1. Personal interaction with domain experts and literature was used to capture human experts' opinion about the domain data. This component is the main contribution of this thesis to the existing fuzzy expert system architecture. The component focuses on the extraction of interesting knowledge from past examples based on domain expert perception of the data. It uses the existing fuzzy association rule mining technique based on expert-driven approach (domain expert set the interval boundaries, define the membership functions and the rules consequences) as a knowledge discovery technique in order to solve the problem of knowledge-base unwieldiness, and knowledge-base update. FARME-D is integrated into FES, to facilitate the knowledge-base instant update in case of new experiences identified and validated by the human expert. This component enhances the FES comprehensibility, makes the system knowledge-base void of sharp boundary problem and correspond to the human perception of the application domain. It comprises of five major components which are: application domain historical database, human domain expert, fuzzification engine, expert-driven data miner and rule interpretation engine, as show in Figure 3.1.

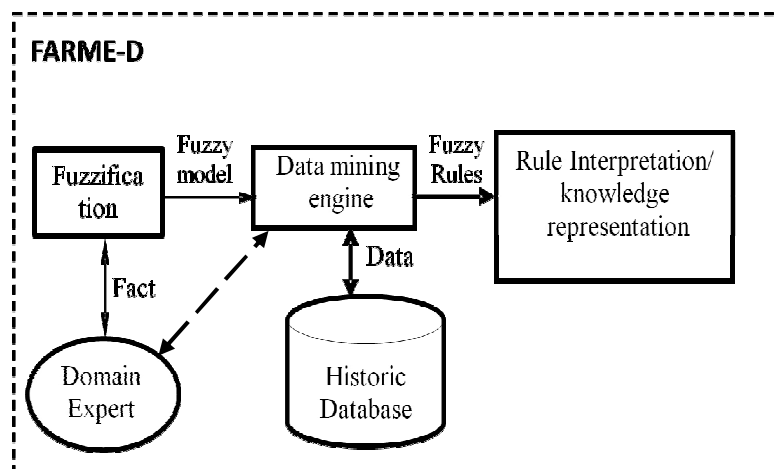


Figure 3.1 The Structure of FARME-D Knowledge Acquisition Approach

3.3.1 Historical database

The historical database is an important component of FARME-D approach, since the proposed approach is acquiring knowledge from past examples. The data-set of 389 records consisting of 8 attributes of non-smoking men with no diabetics history of Cleveland Clinic Foundation database and Hungarian database from University of California, Irvine (UCI), online machine learning repository was used for the mining process. The historical database includes the description of the data stored, input variables and the result variable. Each record contains both the input variables values and the output variable value. Another assumption of this approach is that the historical database is in a structured form. FARME-D does not have the capacity to mine the unstructured database. The database model platform supported by FARME-D is Relational Data Based (RDM) model and the choice of Database Management System (DBMS) is Structure Query Language Server Management Studio, 2005.

3.3.2 Domain expert

The expert here refers to the domain human expert who is ready to supply every piece of information (fact) necessary for mining and fuzzification process. Information collection could be achieved through an oral interview, questionnaire approach or literature. The information includes the description of each attribute in the historical database and the application domain business rules and features. This information enhances the extracted rules by the data miner and help in partitioning linguistic variables into labels that correspond to the domain expert perception in order to avoid the sharp boundary problem (over estimation or under estimation of boundary values).

3.3.3 Expert-Driven Fuzzification Process

In this section the information gathered from the domain experts is put together to determine the dimensions and the subspaces for both the input and output variables

(linguistic variables). The fuzzy set for each linguistic variable is also determined based on the information from the expert which enhances the effectiveness of the fuzzy models. Also, the membership functions for each linguistic fuzzy set are calibrated following the expert's opinion about the data interval partitioning. Using expert-driven and data-driven approach (where data intervals are generated automatically from the data table), one may expect rules obtained to be significantly different. Hence, the membership functions obtained from data-driven approach such as clustering may not correspond with the most intuitive human perception of concept. So, this thesis engages the expert-driven approach to enhance the mining capacity of the existing fuzzy association rule mining algorithm.

3.3.4 Data Mining Engine

The data mining processes start from data pre-processing and end with Fuzzy Association Rule Mining (FARM). Data pre-processing is a supporting activity. It comprises data cleaning, data integration, data transformation and data reduction or selection. Mining activities include data pre-processing and rule elicitation, rule evaluation.

A) Data Pre-processing

The real-world data tend to be dirty, incomplete, and inconsistent. Data pre-processing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data pre-processing is an important step in the data mining process because quality decisions must be based on quality data (Han & Kamber, 2001). First and foremost the medical historical dataset upon which the mining engine process is performed must be identified. The other activities are data cleaning, data integration, data transformation, and data reduction or selection.

(i) **Data Cleaning**

Real world data tend to be incomplete, noisy, and inconsistent. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. During the data cleaning in FARME-D the missing values from the historical data set are fixed by using the attribute mean (Han & Kamber, 2001). Noise in data preprocessing is a random error or variance in a measured variable. Data smoothing is achieved by fuzzification of each attribute value based on the data description by the domain experts

(ii) **Data integration**

The data analysis task in FARME-D involves combination of data from multiple sources in order to have substantial number of records for the mining process. In order to evaluate the correlation between the attributes so as to remove redundant records the known Pearson's product moment coefficient, named after Karl Pearson was adopted (Han & Kamber, 2001). This is :

$$r_{AB} = \frac{\sum (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum (a_i - \bar{a})^2} \sqrt{\sum (b_i - \bar{b})^2}} \quad (3.1)$$

where N is the number of tuples, a_i and b_i are the respective values of A and B in tuple i , \bar{a} and \bar{b} are the respective mean values of A and B , σA and σB are the respective standard deviations of A and B and $\sum (a_i b_i)$ is the sum of the AB cross-product (that is, for each tuple, the value for A is multiplied by the value for B in that tuple). Note that $-1 \leq r_{AB} \leq +1$. If r_{AB} is greater than 0, then A and B are positively correlated, meaning that the values of A increase as the values of B increase. The higher the value, the stronger the correlation (i.e., the more each attribute implies the other). Hence, a higher value may indicate that A (or B) may be removed as a redundancy. If the resulting value is equal to 0, then A and B are independent and there is no correlation

between them. If the resulting value is less than 0, then A and B are negatively correlated, where the values of one attribute increase as the values of the other attribute decrease.

This means that each attribute discourages the other. So, in FARME-D process such records with higher value of correlation are to be removed in order to avoid redundancy of record.

(iii) Data Transformation

In data transformation, the data are transformed into a form appropriate for mining. In FARME-D the smoothing approach was also used whereby the data set are fuzzified based on the constructed fuzzy model for each attribute. Fuzzy model is appropriate because of the quantitative nature of the medical data set and the fuzzy association rule mining technique adopted for mining process.

(iv) Data Reduction

The historical data under consideration might have more than the required attributes for mining process. Therefore, there is need for data reduction, where relevant attributes are selected from the entire database. This is based on the determinant factor of the problem solving.

B. Mining Process

Prior to the mining process proper, there must be data pre-processing activities to ensure that accurate relevant data set is prepared for the mining process. The activity involved in this section is termed Rule Elicitation.

To evolve the interesting knowledge-base, void of redundant records, there is need to identify the hidden relationship between the input attributes and the output attribute from application domain historical database (domain human experts past experiences).

This is very crucial in order to access different domain human experts' knowledge for enriching the knowledge-base. To achieve this Fuzzy Association Rule Mining (FARM) algorithm has proved sufficient over the years as discussed in chapter 2. FARM is a data mining technique that hybridizes the fuzzy concept and association rule mining in order to enhance the functionality of the traditional association rule mining algorithm in mining quantitative data attributes. In FARME-D, fuzzy association rule mining Apriori-like algorithm with quantity measure of significance and certainty factor by Gyenesisi, 2000 is adopted. The algorithm is shown in Figure 2.20. The input to the algorithm is the crisp data set from the application domain historical database. The intermediate output of this algorithm is a fuzzy database got as a result of data transformation. The final output from the algorithm is the set of rules in the form:

$$A, B, C, D \rightarrow E$$

The algorithm is modified so as to return only the 4th order antecedent rule because we have four determinant factors for our case study and all contribute to the output decision according to the expert and literature. Also, it is modified to avoid returning of the rules reverse such as

$$\rightarrow A, B, C, D$$

The choice of programming language for implementing the Fuzzy Association rule mining Apriori-like algorithm is C sharp (C#) programming language because of its supporting features for the algorithm and to enhance a user friendly interface.

3.3.5. Rule Interpretation/Knowledge Representation

After extracting all the relevant rules (interesting rules) within the context of the application domain, the next thing is to interpret the rules according to the domain expert perception and represent them in a standard knowledge representation format that will support the choice of the programming language and tools for building the expert system. The choice of knowledge representation is relational structure where all the rules and the consequents are represented as attribute on the relation. Each record represents a rule and every attribute represents a unique fuzzy set. This choice is intimated by the programming language chosen to validate the proposed approach.

3.4 INTEGRATION OF FARME-D APPROACH TO STANDARD FUZZY EXPERT SYSTEM ARCHITECTURE

The integration of FARME-D as knowledge acquisition component into the standard fuzzy expert system architecture as shown in Figure 2.11 resulted into a derived architecture called **Fuzzy Association Rule Mining Expert System (FARMES)** architecture. FARMES can be defined as an expert system which consists of fuzzification, expert-driven data mining engine (FARME-D), and knowledge-base and defuzzification subsystems, and uses collection of fuzzy membership functions and interesting fuzzy rules instead of Boolean logic to reason about data in the inference mechanism. The structure of the derived FARMES is shown in Figure 3.2. FARMES architecture provides insights into the activities involved in the modelling of a FES using FARME-D automated knowledge acquisition approach.

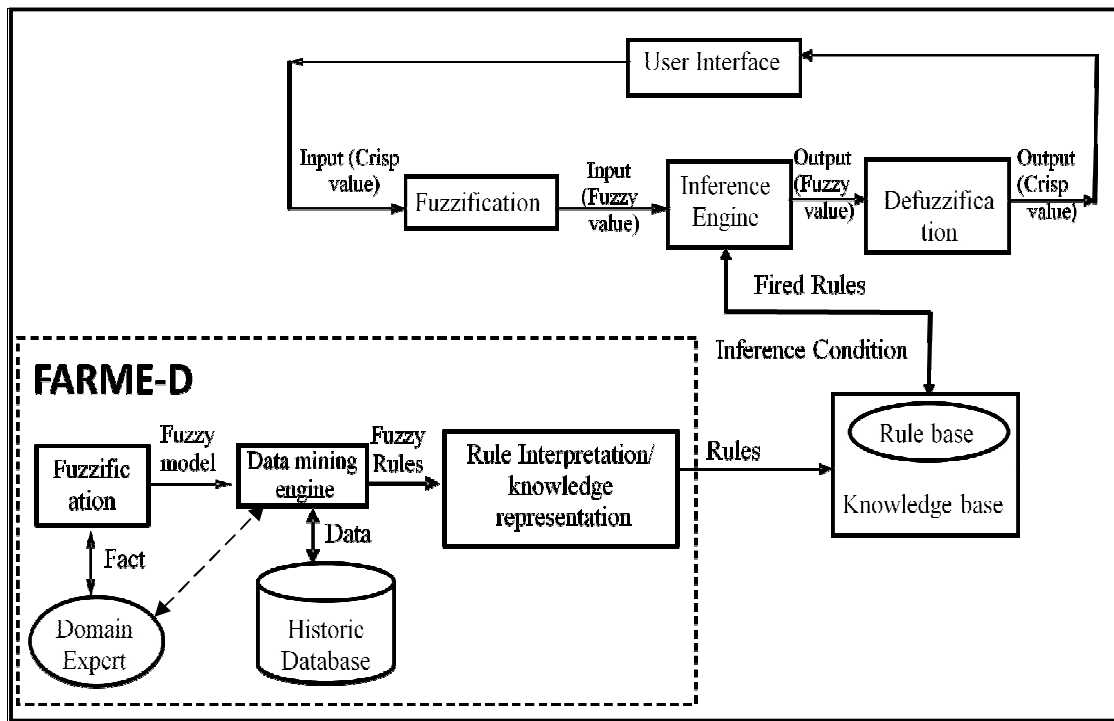


Figure 3.2 FARMES architecture

3.4.1 Summary on how FARMES works

The fuzzy association rule mining expert system being adapted from FES does the following:

- 1) Determines the fuzzy membership values activated by the inputs based on the data description by the domain expert.
- 2) Extracts interesting rules from historical database, base on experts' perception to evolve the knowledge-base
- 3) Determine which rules are fired in the rule set.
- 4) Combines the membership values for each activated rule using the AND operator.
- 5) Traces rule activation membership values back through the appropriate output fuzzy membership functions.
- 6) Utilizes defuzzification to determine the value for each output variable.
- 7) Makes decision according to the output values

3.5 TOOL SUPPORT FOR FARME-D

In standardizing FARME-D process architecture to evolve into a repeatable practice that is industrially applicable, adequate tool-support and programming language for modelling the architecture are essential. A standardized tool-support and programming language for implementing FARME-D has been identified. This is drawn mainly from the field of data mining. These tools and programming languages have been classified into functional categories as follows:

- Software Architecture Specification and Modelling: xADL (Dashofy et al., 2001), ACME (Garlan et al., 1997); ArchStudio 4.0 (<http://www.isr.uci.edu/projects/archstudio>), Ménage (Garg et al., 2003) etc.
- Software design: UML-based tools (Microsoft Visio, Rational Rose, ArgoUML etc.), MDA tools (Eclipse Modelling Framework (EMF) (<http://www.eclipse.org>), Visual Paradigm, Enterprise Architect, AndroMDA (www.modelbased.net/mda_tools.html) etc.
- Software Programming: Visual studio environment e.g C#, Integrated Development Environments (IDEs) e.g. Net Beans 5.x, Microsoft .Net , C language, C++ etc.
- Database Management System (DBMS) : Structural Query Language Server Management Studio Express, Microsoft Access, Oracle DBMS etc

Some of these tools and programming languages were utilized in the case study section (Chapter 4) of this thesis where the FARME-D is applied into the field of medicine.

3.6 APPLICATION SCENARIOS

The FARME-D approach is designed to find application in several fields where intelligent system is recommended and there is historical database upon which the past human expert experience could be referenced. It is applicable in fields such as medicine, engineering, education, agriculture, communication etc. The following are typical scenarios:

1. In the field of medicine for diagnosis and determination of risk ratio. For example in a scenario similar to Allahverdi et al. (2007) a fuzzy expert system was designed for determination of Coronary Heart Disease Risk. In the design, the standard rule-base formulation approach was used and 108 rules were evolved for the knowledge-base. According to Meesad (2000) and Aly & Vrana (2006), if all rules are returned for the knowledge-base then the system is not compact because some of the rules might not be applicable to the problem solving in the disease domain. Therefore, in order to evolve interesting rules that evolved from past experience in the domain, FARME-D approach of knowledge acquisition is appropriate.
2. Also, FARME-D knowledge acquisition approach is appropriate for knowledge elicitation in modelling a FES where the input variables are increasing exponentially. In this case the human expert may not be able to supply all relevant rules.
3. In an application domain where human experts are scarce or no more available and there is need for replication of their knowledge, then, FARME-D will be useful to extract rules from the experts stored experiences. Not only that the rules will also correspond to the domain expert perception. In other words, FARMES architecture could be adapted for modelling expert system.

3.7 VALIDATION APPROACH

In order to validate the plausibility of the proposed solution approach, a case study of fuzzy expert system modelling using FARME-D knowledge acquisition approach is discussed in chapter 4. This shows the practical real-life application scenario of FARME-D approach. The case study is chosen from the field of medicine because the medical field presents a combination of imprecise causal knowledge, very large amount of information and potentially life-threatening consequences of incorrect conclusion (Fatica et al., 1989; Aly & Vrana, 2006; Chi et al., 2001; Delgado et al., 2003). Therefore, there is a need to evolve a knowledge-base that emulates human cognitive process, corresponding with the most intuitive human perception of concept, consistent and able to give accurate result. Specifically, the case study designs a fuzzy expert system for determination of Coronary Heart Disease (CHD) risk for patient using FARME-D automated knowledge acquisition approach. The integration of FARME-D with FES gives room for knowledge-base instant update. It also, help in evolving a knowledge base void of sharp boundary problem which emulates human cognitive process, corresponding with the most intuitive human perception of concept, consistent and able to give accurate result.

3.8 FARME-D IN MEDICAL DOMAIN

To the best of our knowledge, so far, there is no research effort in medical domain that is based on FARME-D knowledge acquisition approach that has been reported in literature, most especially in modelling expert system for heart disease. This is irrespective of the fact that there are a number of approaches that have been validated for knowledge acquisition in modelling expert system for heart disease diagnosis.

In (Shi et al.,1999) an evolutionary fuzzy system was presented. A hybrid fuzzy-neural based medical diagnosis system was proposed in (Moein et al., 2008 and Christoph, 1995). Classification based data mining was used by (Harleen & Siri, 2006; Gadaras & Mikhailov, 2009). Another related work is proposed by using multi layer perceptron to build decision support system for the diagnosis of five major heart diseases (Yan et al., 2006). Also in 2009, neural network was used to generate rules for Hepatitis B intensity rate (Neshat & Yaghobi, 2009). Research work on diagnosis of Coronary Artery Disease was also carried out using rough set theory (RST). The rules were selected and fuzzified based on information from discretization of numerical attribute (Setiawan et al., 2009). Adeli & Neshat, recently designed a fuzzy expert system for heart disease diagnosis (Adeli & Neshat,2010). Eleven input linguistic variables and one output linguistic variable were used for disease diagnosis.

In some other cases, rules were generated by standard structure of rule-base formulation (Allahverdi et al., 2007; Saritas et al., 2003). In a standard structure of a fuzzy system rule-base formulation, given M dimensions and each dimension partitioned into N subspaces, there exist up to N^M rules in the fuzzy system (Meesad, 2001). Allahverdi et al., in their work considered 4 input dimensions to determine CHD patients risk ratio. Three of them were partitioned into 3 subspaces and one into 4 subspaces to form 108 rules. All the rules generated were used to build a fuzzy expert system. From their report, test case evaluation approach was used and the accuracy of the system was determined by comparing the system output with ATP III (Adult Treatment Panel III) results. ATP III results are determined by the domain expert in the field. The system was defined to be accurate to a reliable extent and has coverage, but based on quantitative measure of compactness by Meesad (2001) the

system is not compact. The similarity between this work and FARME-D approach proposed in this thesis is that they are both designed for fuzzy rule-based systems. However, the differences are as follows:

- 1) Instead of generating the rules by standard structure of rule-base formulation, the proposed approach uses a data mining engine (FARME-D) to extract interesting rules from the historical database based on the domain experts' opinion.
- (2) In our work the rule-base is not in a static mode as in Allahverdi et al., 2007; the resulted architecture (FARMES) gives room for instant update of the rule-base as new knowledge is identified.

In all, some of the proposed approaches suffer from sharp boundary problem, rule inconsistency and were based on data-driven approach. Data-driven approach does not concern itself with the membership functions corresponding with domain expert's perception of the data neither the rules. More importantly, some of them evolve a very large number of rules in the knowledge-base.

Hence, the FARME-D approach is unique, offering a more compact platform for enabling a comprehensive FES and dynamic knowledge-base which correspond with the domain expert's perception.

3.9 SUMMARY AND DISCUSSION

In this chapter, Fuzzy Association Rule Mining Expert-Driven (FARME-D) automated knowledge acquisition approach has been presented as a solution model for the two research questions posed in this thesis. FARME-D is an automated knowledge acquisition approach which incorporation application domain experts' opinion into existing fuzzy association rule mining process (domain expert determines fuzzy membership functions). This is to extract minimized number of interesting rules that

correspond intuitively with human expert's perception of decision making and void of the sharp boundary problem for modelling a fuzzy expert system. FARMES architecture is an adapted FES which incorporates FARME-D approach for knowledge acquisition. This promotes expert systems' knowledge-base instant update. The practical application of FARME-D automated knowledge acquisition will be discussed in the next chapter.

CHAPTER FOUR

PRACTICAL APPLICATION OF FARME-D IN MEDICAL DOMAIN

4.1 INTRODUCTION

This chapter presents details of a real-life knowledge engineering scenario where FARME-D has been applied in modelling Coronary Heart Disease (CHD) risk determination expert system in medical domain. The choice of the application domain came as a result of our investigation into the effect of sharp boundary problem in medical domain. The core motivation of this case study is to validate the FARME-D approach and provide a basis for its evaluation. Also, in medical domain, statistics has consistently shown that coronary heart disease is one of the leading causes of death all over the world including the African continent (Yan et al., 2006). Many people had fallen victims of such death because they lack knowledge of their heart disease risk status. However, many lives could be saved if an adequate risk determination expert system is made available for people in order to know their status.

In order to achieve this, Fuzzy Association Rule Mining Expert-Driven approach to knowledge acquisition project was undertaken within the framework of the Software Engineering and Intelligent Systems (SEIS) research cluster of Covenant University. This was aimed at developing a comprehensive Fuzzy Association Rule Mining Expert System with dynamic knowledge-base that corresponds more intuitively to human expert perception. To the best of our knowledge there is no one medical fuzzy expert system that evolves knowledge-base through fuzzy association rule mining expert-driven approach, in order to solve the problem of knowledge-base unwieldiness.

This chapter reports the result of our investigation on sharp boundary problem in medical domain and presents the practical application of the FARME-D process life cycle as undertaken in a case study aimed at validating the plausibility of the FARME-D approach.

4.2 ON SHARP BOUNDARY PROBLEM IN MEDICAL EXPERT SYSTEM

More recently, the application of conventional rule-based expert system for disease risk determination in medical domain has been on the increase. However, a major limitation to the effectiveness of rule-based expert system approach is the sharp boundary problem. This ultimately affects the accuracy of the expert system recommendations. Therefore in this thesis; an investigation into the effect of SBP in medical expert system was carried out to determine the viability of fuzzy expert system in medical domain. Specifically, a fuzzy expert system for determination of CHD risk was built as a case study. To achieve this, two different approaches of ES implementation were considered. The first adopted quantitative binary partition to determine determinant factors subspaces while the second adopted fuzzy partitioning. The partitioning ranges were determined based on data description by the expert doctors and literature (Allahverdi et al., 2007, Bayliss, 2001).

4.2.1 The Investigation Process

(a) Data Sets

The investigation was carried out with a pilot study on 20 non-smoking men record from literature (Allahverdi et al., 2007) in accordance with Adult Treatment Panel III (ATP III) Guidelines for CHD risk ratio determination by National Cholesterol

Education programme. According to the domain expert and literature, smoke, cholesterol, blood pressure, diabetes, sex and age are main risk factors that determine heart disease risk. For the purpose of this investigation and in accordance with literature, four of the factors were considered since our pilot test was based on smoking men with no medical history of diabetes. Therefore, the input attributes are age: year; cholesterol: mg/dL; high density lipoprotein cholesterol: mg/dL (HDL-C); and systolic blood pressure level: mm/Hg. The output attribute is CHD risk ratio.

(b) Quantitative binary partition

A binary partitioning strategy was used, whereby an element either belongs to a partition or not. For the input and output variables partitions we used distance-based partitioning method because it seems most intuitive, since it groups values that are close together within the same interval. For age, we have three partitions young, middle and old. For cholesterol, we also have three partitions of Low, Normal and High. High density Lipoprotein cholesterol (HDL-C) is partitioned into three linguistic terms of Low, Middle and High. The Blood pressure is partitioned into four linguistic terms: Low, Middle, High, VeryHigh. Lastly for the output linguistic variable, CHD risk, we have 5 linguistic terms of VeryLow, Low, Middle, High, VeryHigh. These can be represented as follows:

Age{ Young, Middle Old}

Cholesterol { Low, Normal and High}

HDL{Low, Middle, High}

Blood Pressure{Low, Middle, High, VeryHigh}

CHD Risk { VeryLow, Low, Middle, High, VeryHigh}

The partition ranges and the graphical representations are shown in Figures 4.1- 4.6.

Age	Linguistics term
$x < 30$	Young
$30 = x \leq 55$	Middle
> 55	Old

(a)

Cholesterol	Linguistics term
$c < 180$	Low
$180 = c \leq 260$	Normal
$C > 260$	High

(b)

HDL	Linguistics term
$h < 33$	Low
$33 = h \leq 55$	Middle
$h > 55$	High

(c)

Blood Pressure	Linguistics term
$bp < 115$	Low
$115 = bp \leq 148$	Middle
$148 < bp \leq 200$	High
$bp > 200$	VeryHigh

(d)

CHD Risk	Linguistics term
$r < 4$	VeryLow
$4 = r \leq 10$	Low
$10 < r \leq 20$	Middle
$20 < r \leq 30$	High
$r > 30$	Very High

(e)

Figure 4. 1: Input and Output variables partitioning for (a) Age, (b) Cholesterol, (c) HDL-C, (d) Blood pressure (e) CHD % risk

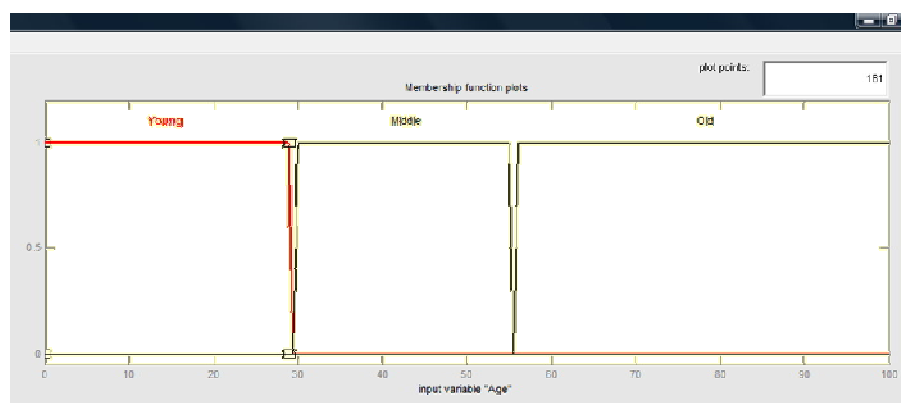


Figure 4. 2 : Binary partition for Age



Figure 4.3: Binary partition for Cholesterol

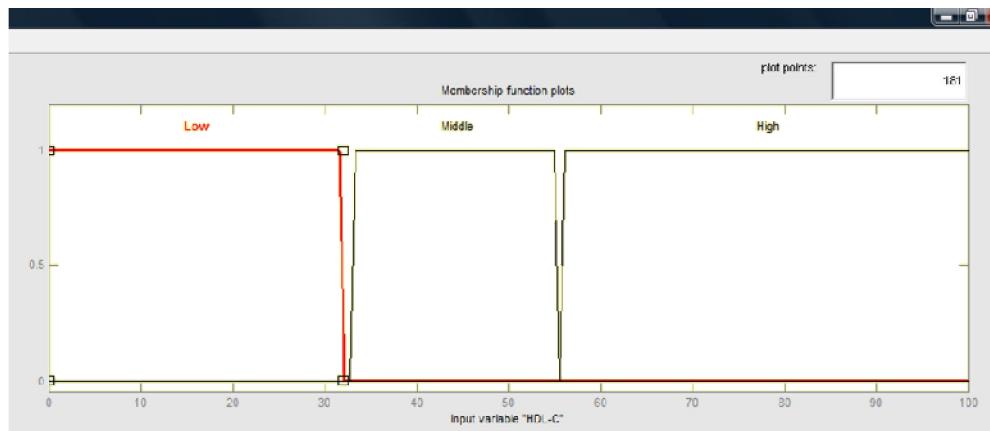


Figure 4. 4: Binary partition for HDL-C

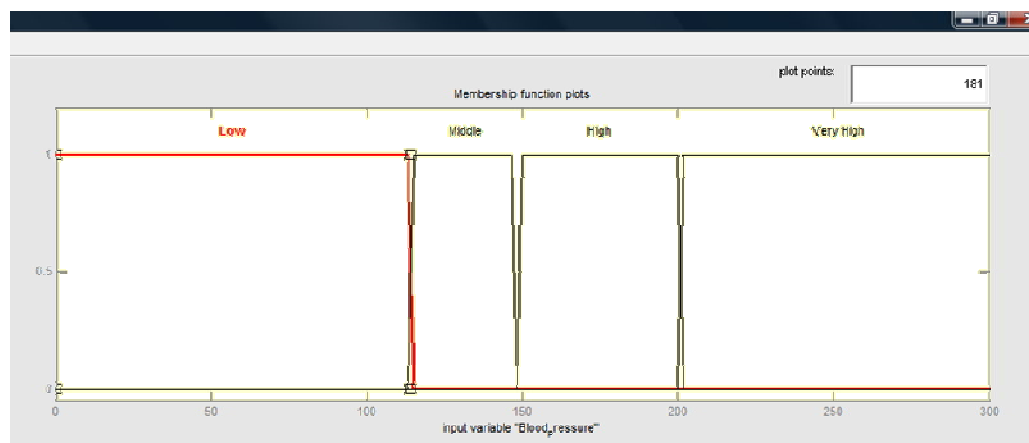


Figure 4.5: Binary partition for Blood Pressure

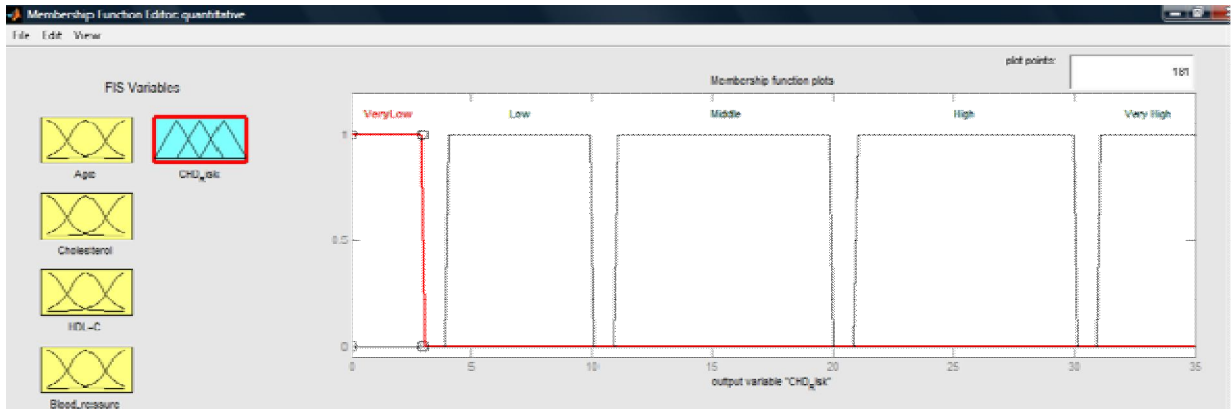


Figure 4.6: Binary partition for %CHD risk

c. Fuzzy Interval partition

Fuzzy partition is more appropriate in this domain because all the five determinant factors are quantitative in nature. The input linguistic variables: age, cholesterol, HDL-C, blood pressure, as well as output parameter: CHD risk were partitioned according to doctors' analysis and literature (Allhaverdi et al.,2007 and Baylis, 2001). Table 4.1 shows the linguistic variables and their fuzzy sets.

Table 4.1: Linguistic variables and their fuzzy sets

Linguistic variable	Domain	Fuzzy set	Membership function
Age	Input	YoungAge, Middle Age, Old Age	trapmf
Cholesterol	Input	Low, Normal, High	trapmf
HDL-C	Input	Low, Middle, High	trapmf
Blood Pressure	Input	Low, Middle, High, Very High	trapmf
CHD risk ratio	Output	VeryLow, Low, Middle, High, Veryhigh	trimf

The trapezoidal membership function (trapmf) was used to model each input linguistic label, and the membership expression. Also, for the output linguistic labels, triangular membership function (trimf) was used because of their support for the fuzzy sets data ranges. The membership functions plots are shown in Figure 4.7- 4.11. For Age value (let x) fuzzy membership expressions will be as:

$$\mu_{youngAge}(x) = \left\{ \begin{array}{ll} 1 & x \leq 20 \\ \frac{40-x}{20} & 20 \leq x \leq 40 \end{array} \right\} \quad (4.1)$$

$$\mu_{middleAge}(x) = \left\{ \begin{array}{ll} \frac{(x-20)}{20} & 20 \leq x \leq 40 \\ 1 & 40 \leq x \leq 50 \\ \frac{(60-x)}{10} & 50 \leq x \leq 60 \end{array} \right\} \quad (4.2)$$

$$\mu_{oldAge}(x) = \left\{ \begin{array}{ll} \frac{(x-50)}{10} & 50 < x \leq 60 \\ 1 & 60 \leq x \leq 70 \end{array} \right\} \quad (4.3)$$

For Cholesterol value (let c) fuzzy membership expressions will be as

$$\mu_{lowCholesterol}(c) = \left\{ \begin{array}{ll} 1 & c < 160 \\ \frac{(200-c)}{40} & 160 \leq c < 200 \end{array} \right\} \quad (4.4)$$

$$\mu_{normalCholesterol}(c) = \left\{ \begin{array}{ll} \frac{(c-160)}{40} & 160 \leq c < 200 \\ 1 & 200 \leq c \leq 240 \\ \frac{(280-c)}{40} & 240 \leq c < 280 \end{array} \right\} \quad (4.5)$$

$$\mu_{highCholesterol}(c) = \left\{ \begin{array}{ll} \frac{(c-240)}{40} & 240 \leq c < 280 \\ 1 & c \geq 280 \end{array} \right\} \quad (4.6)$$

For HDL-C value (let h) fuzzy membership expressions will be as

$$\mu_{lowHDL-C}(h) = \left\{ \begin{array}{ll} 1 & h < 25 \\ \frac{(40-h)}{15} & 25 \leq h < 40 \end{array} \right\} \quad (4.7)$$

$$\mu_{middleHDL-C}(h) = \left\{ \begin{array}{ll} \frac{(h-25)}{15} & 25 \leq h < 40 \\ 1 & 40 \leq h \leq 50 \\ \frac{(60-h)}{10} & 50 \leq h < 60 \end{array} \right\} \quad (4.8)$$

$$\mu_{\text{highHDL-C}}(h) = \begin{cases} \frac{(h-50)}{15} & 50 \leq h < 60 \\ 1 & h \geq 60 \end{cases} \quad (4.9)$$

For Blood Pressure value (let b) fuzzy membership expressions will be as:

$$\mu_{\text{lowBP}}(b) = \begin{cases} 1 & b < 100 \\ \frac{(130-b)}{30} & 100 \leq b < 130 \end{cases} \quad (4.10)$$

$$\mu_{\text{middleBP}}(b) = \begin{cases} \frac{(b-100)}{30} & 100 \leq b < 130 \\ 1 & 130 \leq b \leq 140 \\ \frac{(155-b)}{15} & 140 \leq b < 155 \end{cases} \quad (4.11)$$

$$\mu_{\text{highBP}}(b) = \begin{cases} \frac{(b-130)}{15} & 140 \leq b < 155 \\ 1 & 155 \leq b \leq 180 \\ \frac{(220-b)}{40} & 180 \leq b < 220 \end{cases} \quad (4.12)$$

$$\mu_{\text{veryHighBP}}(b) = \begin{cases} \frac{(b-220)}{40} & 180 \leq b < 220 \\ 1 & b \geq 220 \end{cases} \quad (4.13)$$

For CHD Risk the output value (let r) fuzzy membership expressions will be as:

$$\mu_{\text{veryLowRisk}}(r) = \begin{cases} 0 & r < 1 \\ \frac{(5-r)}{5} & 0 \leq r < 5 \end{cases} \quad (4.14)$$

$$\mu_{\text{lowRisk}}(r) = \begin{cases} \frac{r}{5} & 0 \leq r < 5 \\ \frac{(15-r)}{10} & 5 \leq r < 15 \end{cases} \quad (4.15)$$

$$\mu_{middleRisk}(r) = \begin{cases} \frac{(r-5)}{10} & 5 \leq r < 15 \\ \frac{(25-r)}{10} & 15 \leq r < 25 \end{cases} \quad (4.16)$$

$$\mu_{highRisk}(r) = \begin{cases} \frac{(r-15)}{10} & 15 \leq r < 25 \\ \frac{(35-r)}{10} & 25 \leq r < 35 \end{cases} \quad (4.17)$$

$$\mu_{veryHighRisk}(r) = \begin{cases} \frac{(r-25)}{10} & 25 \leq r < 35 \\ 1 & r \geq 35 \end{cases} \quad (4.18)$$

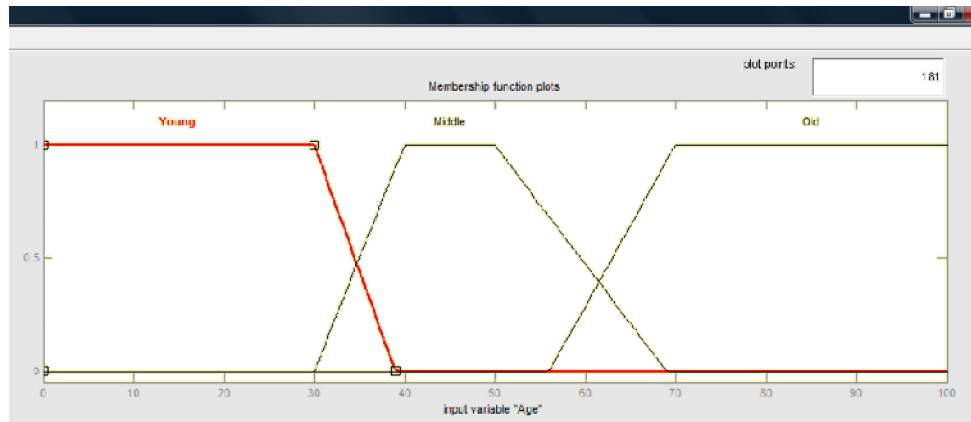


Figure 4.7 :The membership function for Age

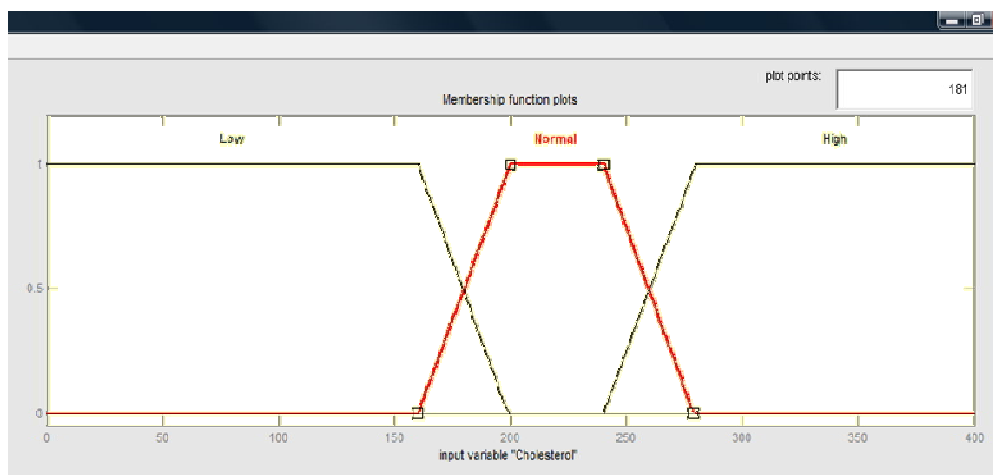


Figure 4.8: The membership function for Cholesterol

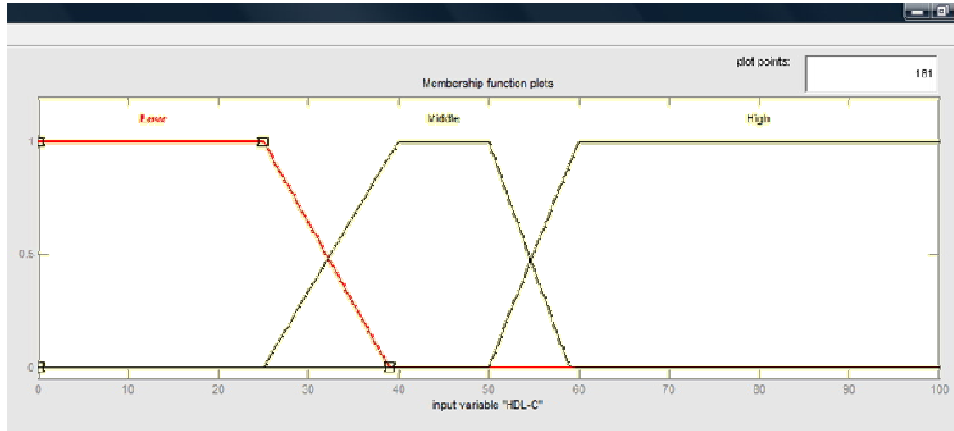


Figure 4.9: The membership function for (HDL-C)

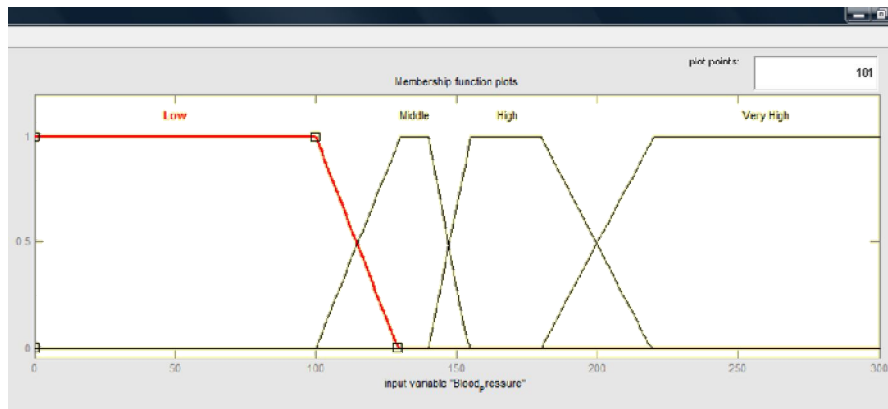


Figure 4.10: The membership function for Blood pressure

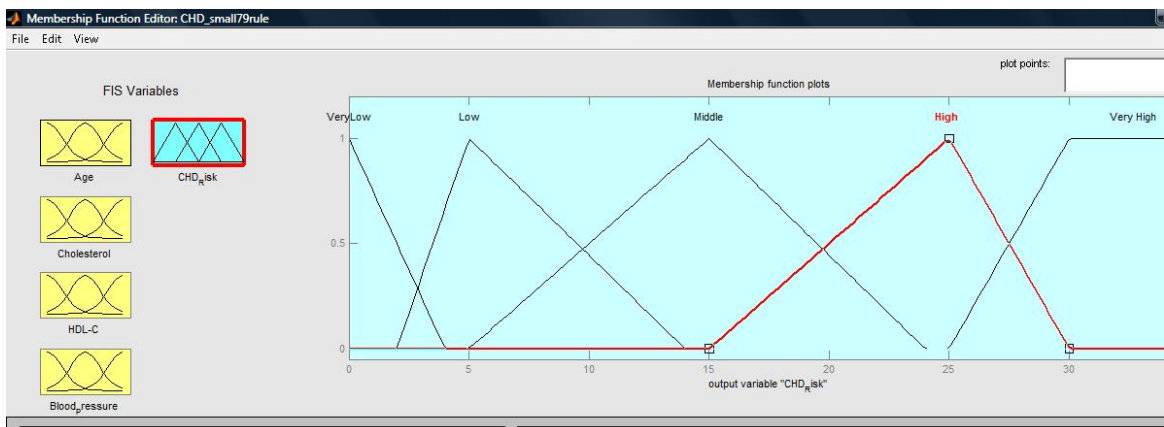


Figure 4.11: The membership function for CHD %risk

d. Rule Generation

For the purpose of this investigation the standard rule-base formulation was adopted to generate the rules, such that given M dimensions where each dimension is partitioned into N subspaces, there exist up to N^M rules in an expert system rule-base (Meesad, 2001). For this experiment we have 4 dimensions of which 3 are divided into 3 subspaces and the 4th dimension is divided into 4 subspaces as shown in Table 4.1. There exist 108 rules for the CHD risk determination expert system, based on the number of dimensions and subspace. The process is automated with C# programming language. The snapshot for the automated standard rule formulation process is shown in Figure 4.12. For each rule antecedent the consequent value is determined based on the Framingham CHD risk point score as shown in the appendix A.

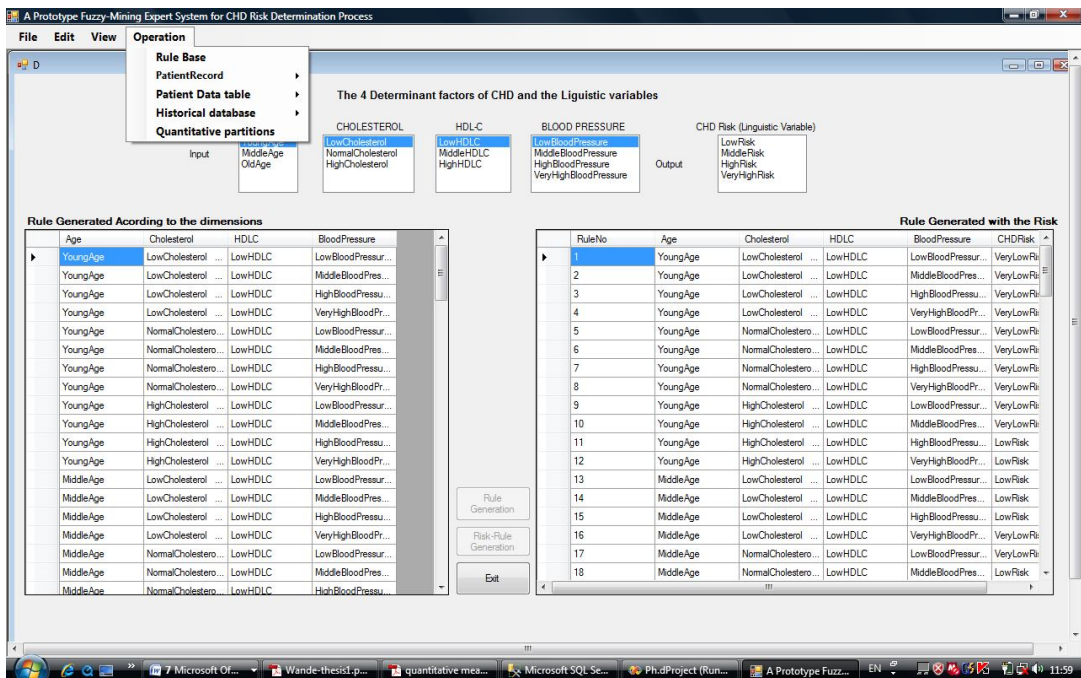


Figure 4.12: The snapshot for standard rule-base formulation process

(e) **Quantitative Binary Expert System (QBES)**

In this thesis, we modeled quantitative binary expert system based on binary partitioning strategy as discussed in section 4.2.1(B). MatLab™ fuzzy logic toolbox was used to simulate the expert system and the result is shown in table 4.2. The rule editor was generated with 108 rules. The Max-min operator of the Mandani fuzzy inference engine and centroid method of defuzzification process were adopted. For instance, a non-smoking man of age 48, with Cholesterol 260 mg/dL, HDL-C 33 mg/dL, and blood pressure 120mm/Hg, gave the 1.4 CHD risk value and fired only rule number 67 from the list of generated rules as shown in Figure 4.13.

rule 67. *If (Age is Middle) and (Cholesterol is Normal) and (HDL-C is Middle) and (Blood_Pressure is Middle) then (CHD_Risk is VeryLow) (1)*

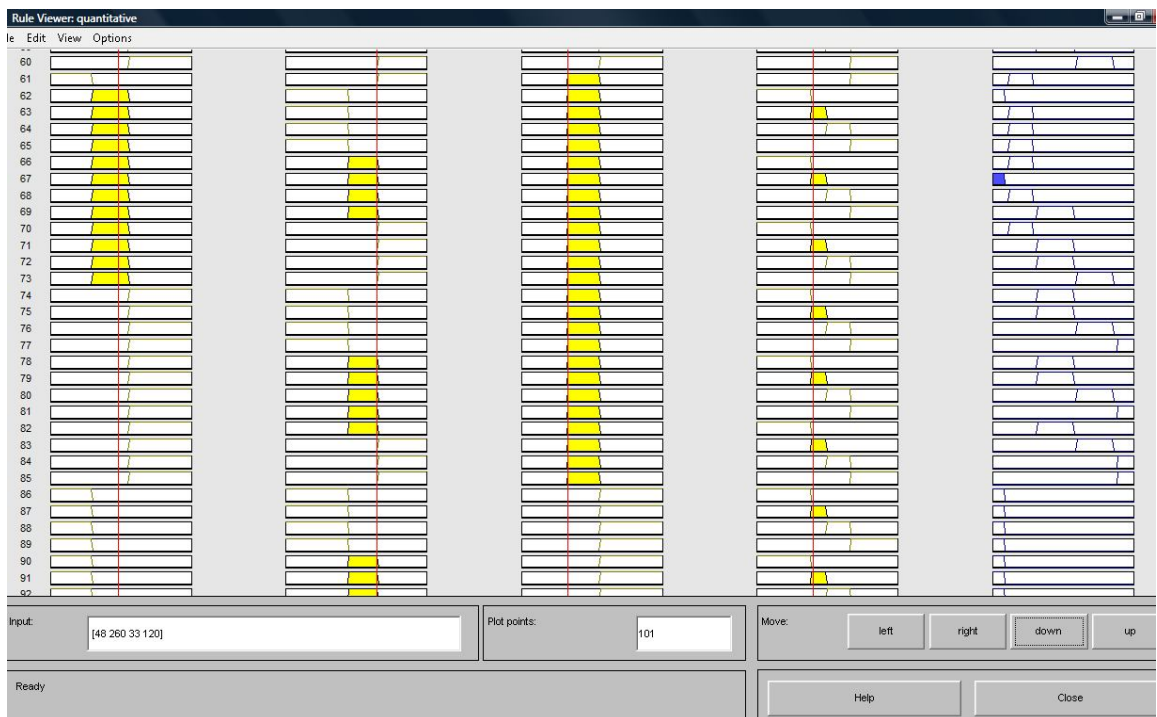


Figure 4.13 QBES CHD risk for the value Age=48, Cholesterol = 260, HDL-C=33, Bloodpressure = 120 with

(f) Fuzzy Expert System (FES)

Also, a Fuzzy Expert System (FES) was modeled based on fuzzy partitioning strategy models in subsection 4.2.1(C). Matlab fuzzy Tool box was used to simulate the expert system and the result is shown in table 4.2. The Max-min operator of the Mandani fuzzy inference engine and centroid method for defuzzification process were used. The rule editor was generated with 108 rules. For instance, a non-smoking man of age 48, with Cholesterol 260 mg/dL, HDL-C 33mg/dL, and bloodpressure 120mm/Hg, gave 10.9 CHD Risk value as shown in Figure 4.14 and fired 8 rules which include rules number 16, 17, 20, 21, 66, 67, 70, 71.

rule16. If (Age is Middle) and (Cholesterol is Normal) and (HDL-C is Low) and (Blood_Pressure is Low) then (CHD_Risk is VeryLow) (1)

rule17. If (Age is Middle) and (Cholesterol is Normal) and (HDL-C is Low) and (Blood_Pressure is Middle) then (CHD_Risk is Low) (1)

rule20. If (Age is Middle) and (Cholesterol is High) and (HDL-C is Low) and (Blood_Pressure is Low) then (CHD_Risk is Middle) (1)

rule21. If (Age is Middle) and (Cholesterol is High) and (HDL-C is Low) and (Blood_Pressure is Middle) then (CHD_Risk is Middle) (1)

rule66. If (Age is Middle) and (Cholesterol is Normal) and (HDL-C is Middle) and (Blood_Pressure is Low) then (CHD_Risk is Low) (1)

rule67. If (Age is Middle) and (Cholesterol is Normal) and (HDL-C is Middle) and (Blood_Pressure is Middle) then (CHD_Risk is VeryLow) (1)

rule70. If (Age is Middle) and (Cholesterol is High) and (HDL-C is Middle) and (Blood_Pressure is Low) then (CHD_Risk is Low) (1)

rule71. If (Age is Middle) and (Cholesterol is High) and (HDL-C is Middle) and (Blood_Pressure is Middle) then (CHD_Risk is Middle) (1)

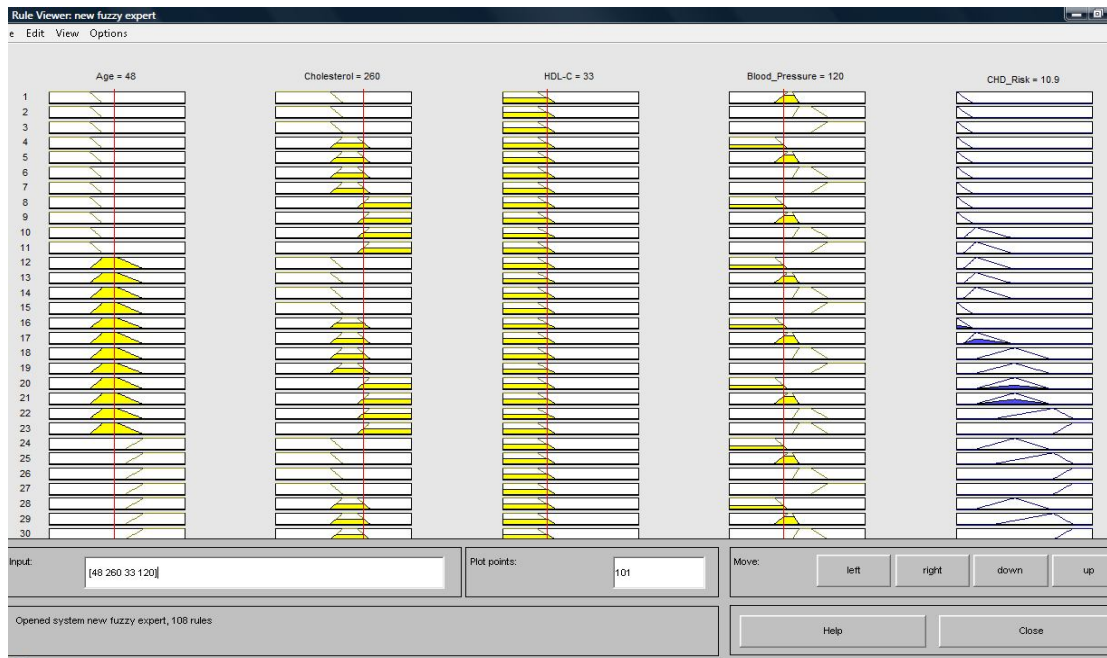


Figure 4.14: FES CHD risk for the value Age=48, Cholesterol = 260, HDL-C=33, Bloodpressure = 120 with CHD risk % = 10.9

(g) Investigation result and Recommendation

The result of the investigation is clearly expressed on Table 4.2. The table shows the 20 non-smoking men with no medical history of diabetes used as the test case. Columns 6, 7 and 8 show the result of three different approaches to determine CHD risk ratio; the ATP III result that represents the domain expert result, FES approach result and QBES approach result respectively.

Table 4.2: ATP III, FES and QBES CHD risk value according to 2+ risk factor CHD for non-smoking men

Patient no	Age	Cholesterol	HDL-C	Blood Pressure	ATP III	FES	QBES	ATP III CHD risk Linguistic value	FES CHD risk Linguistic value	QBES CHD risk Linguistic value
1	30	180	37	160	0	1.5	7	VeryLow	VeryLow	Low
2	35	190	45	145	0	4	1.4	VeryLow	VeryLow	VeryLow
3	48	260	33	120	8	10.9	1.4	Low	Low	VeryLow
4	57	300	67	110	8	9.3	7	Low	Low	Low
5	65	250	54	170	18	19.9	24	Middle	Middle	High
6	75	290	25	135	30	31.5	31	VeryHigh	VeryHigh	VeryHigh
7	30	160	49	160	0	1.5	7	VeryLow	VeryLow	Low
8	40	310	33	140	8	15.5	15.5	Low	Middle	Middle

9	55	300	26	200	30	26.9	24	High	High	High
10	60	230	39	110	11	11.2	15.5	Middle	Middle	Middle
11	70	210	45	130	16	15.5	15.5	Middle	Middle	Middle
12	30	240	50	150	0	1.5	7	VeryLow	VeryLow	Low
13	35	180	65	160	0	5	7	VeryLow	Low	Low
14	45	300	47	155	9	15.5	15.5	Low	Middle	Middle
15	55	300	49	160	16	18.9	15.5	Middle	Middle	Middle
16	65	250	41	140	18	15.6	15.5	Middle	Middle	Middle
17	70	260	38	190	30	28	24	High	High	High
18	44	210	37	180	5	9.2	7	Low	Low	Low
19	55	150	30	200	11	18	15.5	Middle	Middle	Middle
20	66	150	26	200	28	24.6	31.5	High	High	VeryHigh

From Table 4.2, it is observed that FES risk value varies as ATP III risk values based on the input values, while QBES categorises some patients with different input values under the same risk. Table 4.3 shows the extract of few instances. Categorically, this shows the effect of sharp boundary problem in the quantitative binary partitions. It is observed that in those cases expressed on Table 4.3, QBES must have overestimated record 1 values, overestimated record 5 values, underestimated record 3 values and some other not identified.

Table 4.3: Instances showing the effect of SBP on medical expert system

Record	ATP III	FES	QBES	ATP III CHD risk Linguistic value	FES CHD risk Linguistic value	QBES CHD risk Linguistic value
2	0	4	1.4	VeryLow	VeryLow	VeryLow
3	8	10	1.4	Low	Low	VeryLow**
5	18	19.9	24	Middle	Middle	High **
9	30	26.9	24	High	High	High
1	0	1.5	7	VeryLow	VeryLow	Low**
4	8	9.3	7	Low	Low	Low
18	5	9.2	7	Low	Low	Low

Also, according to the linguistic value results on column 9, 10, 11 of table 4.2, it is derived that FES has 80% result similarity with ATP III while QBES has 60 % similarity. This indicates that the FES gives more accurate result compared to QBES. In order to have a better feel of the actual picture, the charts for graphical overview of the results are shown in Figure 4.15 and 4.16. On Figure 4.16, linguistic values for CHD % risk: VeryLow, Low, Middle, High and VeryHigh were represented with values 1,2,3,4,5, respectively.

In conclusion, because of the domain under consideration this has potentially life-threatening consequences of incorrect conclusion (Nunzia et al., 1989, Aly & Vrana, 2006, Chi et al., 2001, Delgado et al.,2003). For these reasons, there is a need to generate knowledge-base that emulates human cognitive process, corresponding with the most intuitive human perception of concept, consistent and able to give accurate result. Therefore, in this research, the result of this investigation serve as a motivation for us to establish our proposed approach on medical fuzzy expert systems.

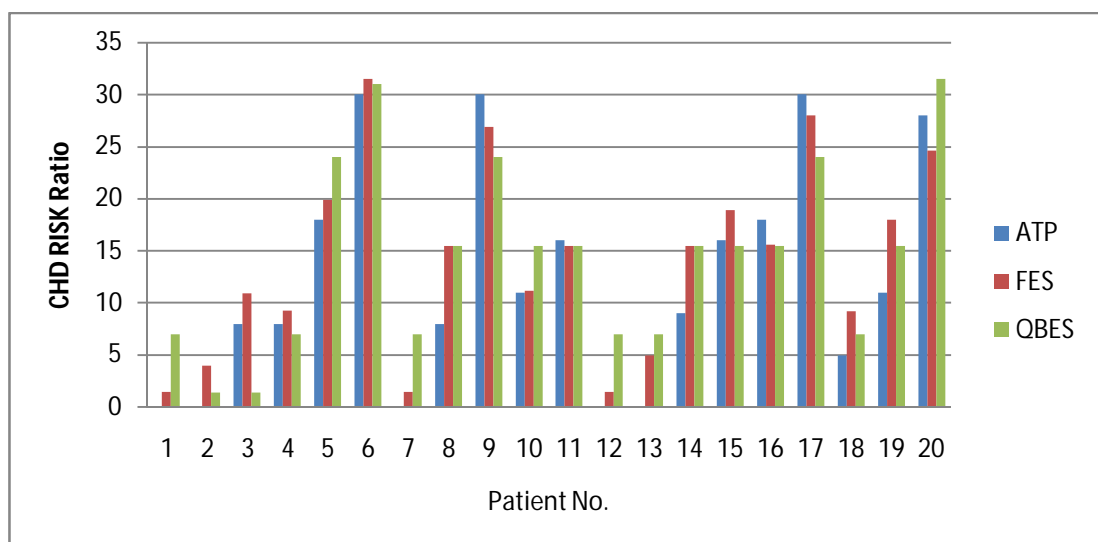


Figure 4.15: ATP III, FES and QBES CHD % risk value diagrammatic representation

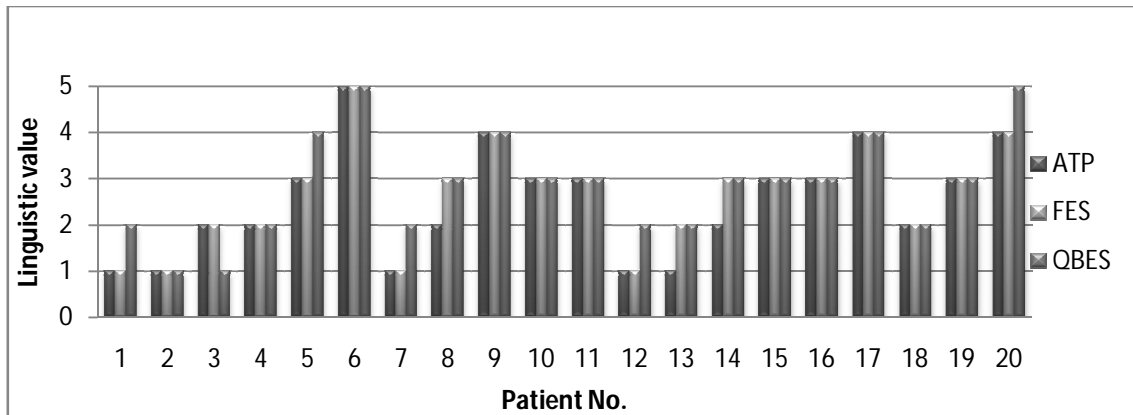


Figure 4.16: The linguistic % CHD risk diagrammatic representation for ATP III, FES and QBES

4.3 FARME-D PROCESS LIFE CYCLE IN BUILDING A FUZZY EXPERT SYSTEM IN MEDICAL DOMAIN

This section reports the practical application of FARME-D automated knowledge acquisition in medical domain. This aims to design a prototype fuzzy expert system for determining CHD risk ratio in order to validate the plausibility of proposed FARM-ED knowledge acquisition approach.

4.3.1 The Prototype User Interface

A user friendly interface was designed to communicate between the patient and the expert system with C# programming language. The interface has four main menus with some drop down submenus. The main menus are File, Edit, View and Operation. File has exit submenu at which the user gets out of the expert system environment. Edit has Patient record and Lab test. From the patient record form each user can submit their basic information and their lab test data can be submitted from Lab test form. From View, patient record can be viewed; the knowledge-base rules and patient fuzzification records can also be viewed.

The last is the Operation main menu where the whole operations take place. It has a Rule-base form, Patient record- Fuzzification, Patient Data, Table fuzzification and Historical Database. Historical Database is further subdivided into Fuzzification, Mining and Association rule mining sub-menu. Here, mining process takes place upon the historical database to acquire interesting rule for the knowledge-base. The rule generated is basically determined by the historical database. On the Rule-base form the standard rule-base formulation approach is automated which evolves 108 rules for the knowledge-base. The Patient record-fuzzification form represents the inference engine process. The submenu performs fuzzification process for each patient record, it displays the fuzzification result (based on the membership functions), fired rules, performing the implication process and aggregation; and finally determines the CHD Risk for the patient record. The snapshot for the interface is shown in Figure 4.17.

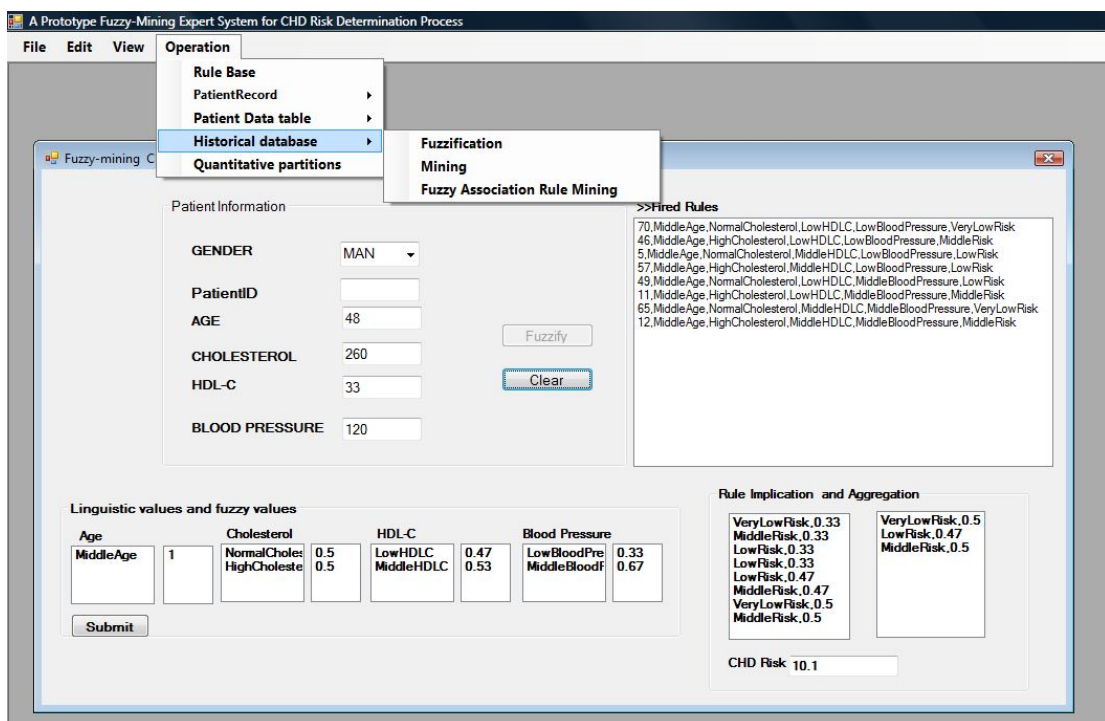


Figure 4.17: The snapshot of the main interface

4.3.2 Fuzzification process

This involves transformation of patient crisp lab test into a fuzzy input. This operation is performed basically with submenu called fuzzification under Operation menu.

i) Domain Analysis

During the cause of this research, expert medical doctors in this area were engaged in one on one oral interview about the CHD problem domain. Also, research works, in this domain were also consulted, especially the report of the research conducted by the Framingham community (Bayliss, 2001) and some others such as (Allahverdi et al., 2007; Yan, et al., 2006; Ali & Mehdi, 2010, Setiawan et al., 2009). From the analysis, seven factors were identified as major determinant factor for CHD risk which are: Age, Sex, Blood pressure, Cholesterol, Smoking status, Family History of CHD, and History of Diabetes. The data description of this attributes is shown Table 4.4. For the purpose of this research, four (4) of this attributes are considered as the determinant factors (Bayliss, 2001, Allahverdi et al., 200, Yan, et al., 2006; Adeli & Neshat, 2010, Setiawan et al., 2009). The 4 attributes are quantitative in nature. These are age: year; cholesterol: mg/dL; high density lipoprotein cholesterol: mg/dL (HDL-C); and systolic blood pressure level: mm/Hg. The final output is CHD risk and is also quantitative in nature.

Table 4.4: Description of Coronary Heart Disease determinant factors

Attribute	Description	Value description
Age	Age (year)	Numerical
Sex	Sex	Categorical
Bp	Systolic blood pressure (mmHg)	Numerical
Cholesterol	Total Serum Cholesterol (mg/dl)	Numerical
HDL-C	High density Cholesterol (mg/dl)	Numerical
Smoke	Smoke status	Categorical
FH	Family History of CHD	Categorical
HD	History of Diabetes	Categorical

(ii) Fuzzy model construction and Implementation

Fuzzy partition is more appropriate in this application domain because all the 5 attributes under consideration are quantitative. It allows for overlapping where a particular record can belong to two neighbouring linguistic labels with their membership grades. This prevents over-estimation of boundary values (Oladipupo et al., 2010). The linguistic variables (the determinant factors and output) are partitioned according to doctors' analysis. Table 4.1 shows the linguistic variables and their fuzzy sets. The trapezoidal membership function (tramf) is used to model each input fuzzy sets membership grade because of its support for the fuzzy sets data ranges. The fuzzy membership models and the graph are as stated in section 4.2.1. (c) . The fuzzification process is aimed to transform the input values into fuzzy values that are appropriate for the inference process.

The fuzzification process is implemented with C# programming language on Visual studio 8 environment. Figure 4.18 shows the snapshot of the input record before fuzzification process while Figure 4.19 shows the snapshot of the fuzzification values. The process output was evaluated by using a test case of some crisp input to validate the accuracy of the automated system.

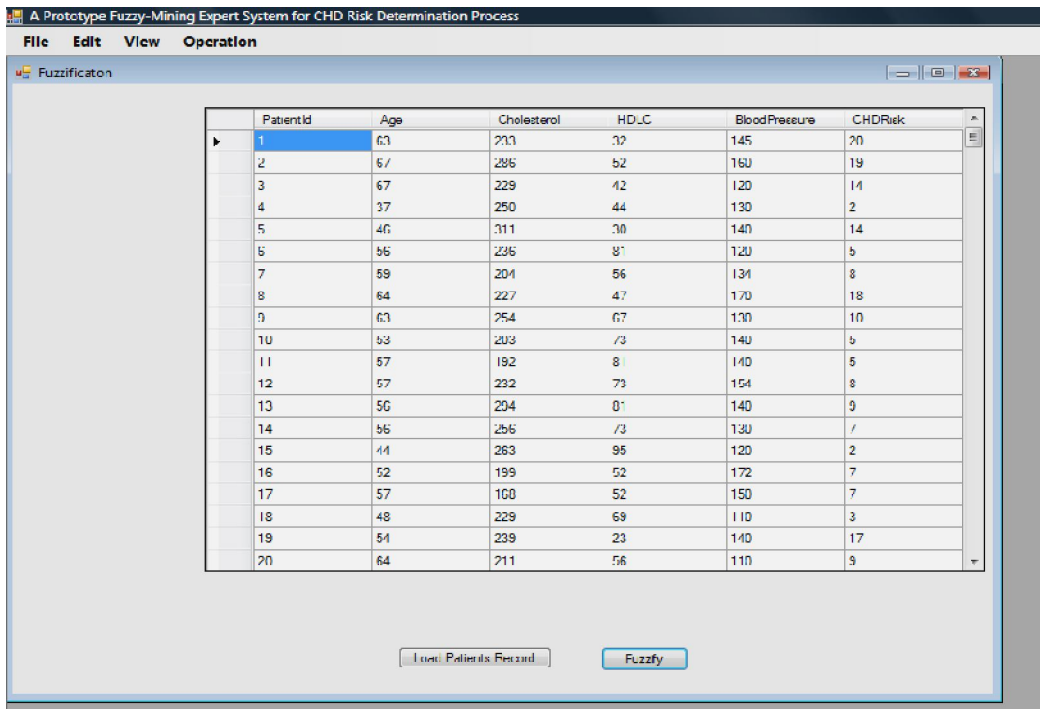


Figure 4.18: The snapshot of the fuzzification process (Crisp values)

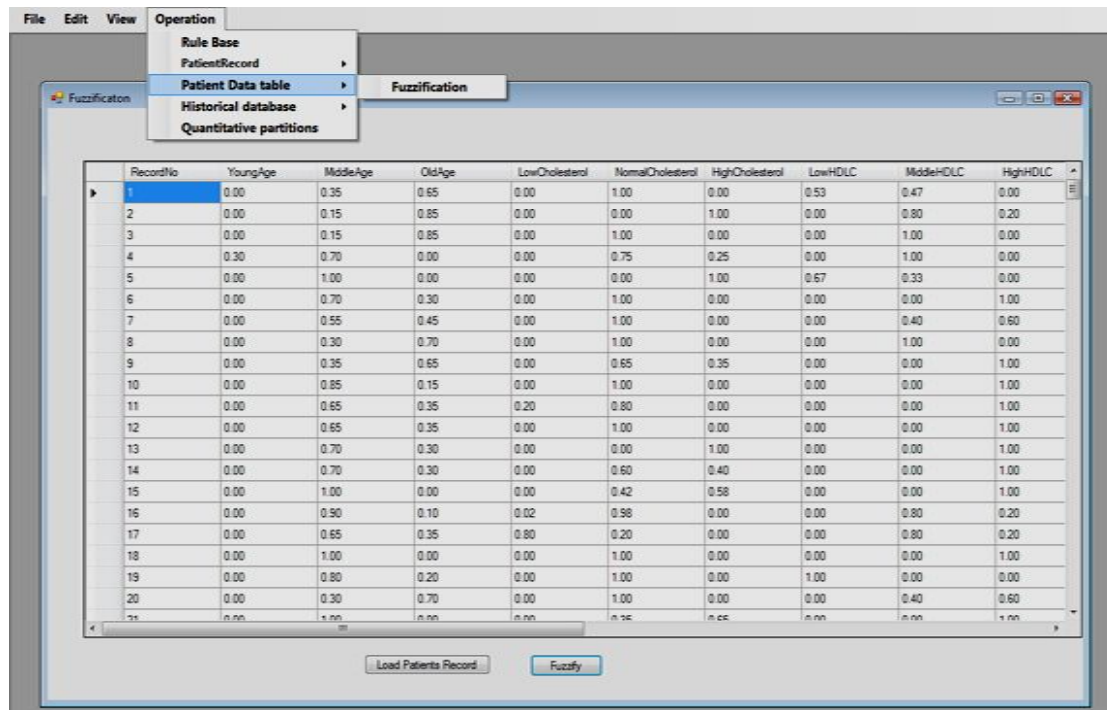


Figure 4.19: The snapshot of the fuzzification process (fuzzified values)

4.3.3 FARME-D approach

This component is the main contribution of this thesis. In modelling the prototype fuzzy expert system, Fuzzy Association Rule Mining Expert-Driven (FARME-D) approach was used to evolve rules for the knowledge-base in the place of standard rule-base formulation approach used in section 4.2.1(d). FARME-D approach aimed at generating interesting rules from mining historical database of past examples, for determining the CHD risk ratio. This is to improve the expert system comprehensibility by minimizing the redundant rules. Redundant rules mean, rules that are not applicable to problem solving in a particular domain based on the domain concept. This component is integrated with standard FES architecture to derive Fuzzy Association Rule Mining architecture that we called FARMES. By integration, we mean that rules are evolved into the expert system knowledge-base and updated directly at every instance of mining process. This enhances the dynamism mode of the knowledge-base in the prototype system.

Fuzzy concept is considered with the mining approach in order to avoid sharp boundary problem according to the investigation report and recommendation in section 4.2.1(g). Also, expert-driven approach of association rule mining is considered so as to generate rules that correspond more intuitively with human domain expert perception. To be able to achieve this, few domain experts that are available and ready to supply every piece of information necessary for mining process were interviewed. This actually buttressed our understanding of the coronary heart disease concept and generally about the cardiovascular diseases.

(i) ***Historical database***

On our visit to some hospitals in the course of this research, it was discovered that there is no local historical database for CHD presently. To this effect the mining process is based on standard Data mining repository from the Cleveland Clinic Foundation and Hungarian dataset. Three hundred and eighty-nine (389) records dataset from the repository was used in accordance with ATP III (Adult Treatment Panel) guidelines for CHD risk ratio determination by National Cholesterol Education programme. The guidelines were based on the Framingham CHD risk point scores, which were used to determine the percentage risk for each record in the sample dataset (Bayliss, 2001). This dataset is part of the collection of databases at the University of California, Irvine (UCL) collected by David Aha. The dataset contained 76 attributes. In this thesis, 4 attributes are selected for the input based on the CHD determinant factors. The total attributes for the mining process are 5, which include the CHD risk ratio attribute. The input fields are age: year; cholesterol: mg/dL; high density lipoprotein cholesterol: mg/dL (HDLC); and systolic blood pressure level: mm/Hg. The output attribute is CHD risk ratio.

(ii) ***Rule Generation***

Fuzzy Association Rule Mining (FARM) was used to discover knowledge from the imputed data set. The data set is represented in relational database format. Each record represents individual patient while attributes represent determinant factors and the CHD risk ratio. Fuzzy association rule mining apriori-like algorithm is implemented with c# programming language on visual studio 8.0 platform. The generated rules are strongly determined by the Historical data set; therefore, as new instances are discovered by the domain experts, the database is updated so also the knowledge-base is updated. The output of the mining process is the set of frequent rules, their support

and confidence value. During the mining process the existing mining algorithm was adjusted so as to return only the 4th order rule-antecedent with one attribute consequence. Also, to factor in the expert's opinion about each rule antecedent in determining the rule consequence. The input on the interface is the threshold values for support and confidence. The implementation snapshot of the process is shown in Figure 4.20.

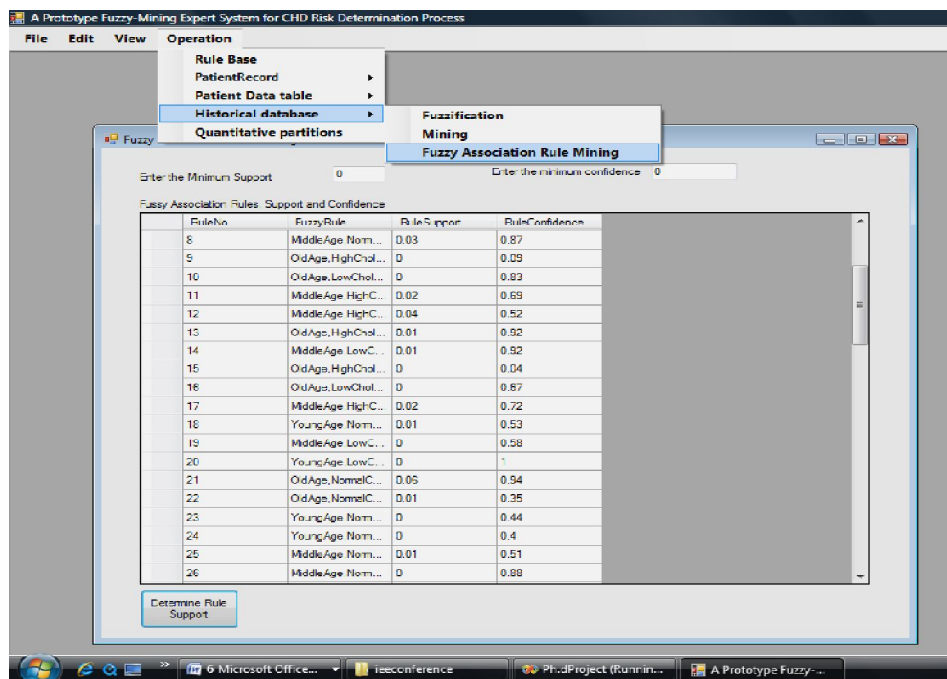


Figure 4.20: The snapshot for FARME-D output

(iii) Rule Selection

Interesting rules are in 4th order rule-antecedent. They have 4 input linguistic variables antecedent and one variable consequent. The minimum confidence of the rules is varied at a constant support value of zero. The numbers of interesting rules at different minimum confidence threshold values are shown below in Tables 4.5 and the graphical representation in Figure 4.21. It is revealed from Figure 4.21 that the higher the confidence threshold the smaller the number of generated rules. As our focus is to minimize the number of rules in the knowledge-base, the knowledge-base

completeness also cannot be traded off. In this case study, the minimum support and confidence thresholds are set to be zeros in order to ensure the expert system completeness and at the same time minimize redundant rules in the knowledge-base with a reasonable percentage. These thresholds (sup = 0.0, conf = 0.0) generate 79 interesting rules as against 108 rules by standard rule-base formulation. This signifies that, any rule which cannot have at least a zero confidence is said not to be relevant in this particular domain based on the existing cases. Table 4.6 shows the sets of frequent rules generated from the mining process with their support and confidence values. The reduced number of rules determines the compactness of the proposed fuzzy-mining expert system (Meesad,2001).

Table 4.5 Confidence value against the number of rules

Experiment. No	Confidence value ≤ 1	No.of interesting rule
1.	0.0	79
2.	0.1	74
3.	0.3	67
4.	0.4	61
5.	0.5	56
6.	0.6	48
7.	0.7	43
8.	0.8	33
9	.0.9	20
10	1.0	10

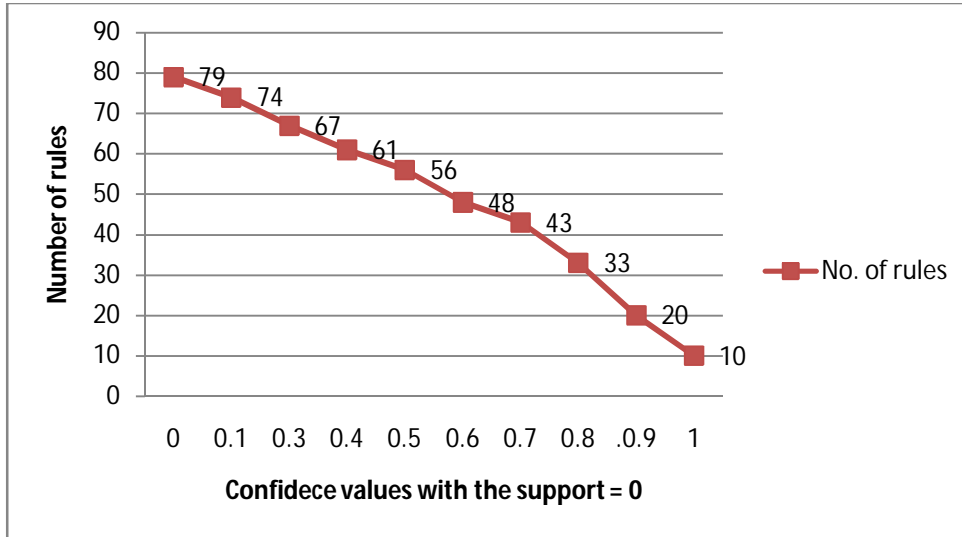


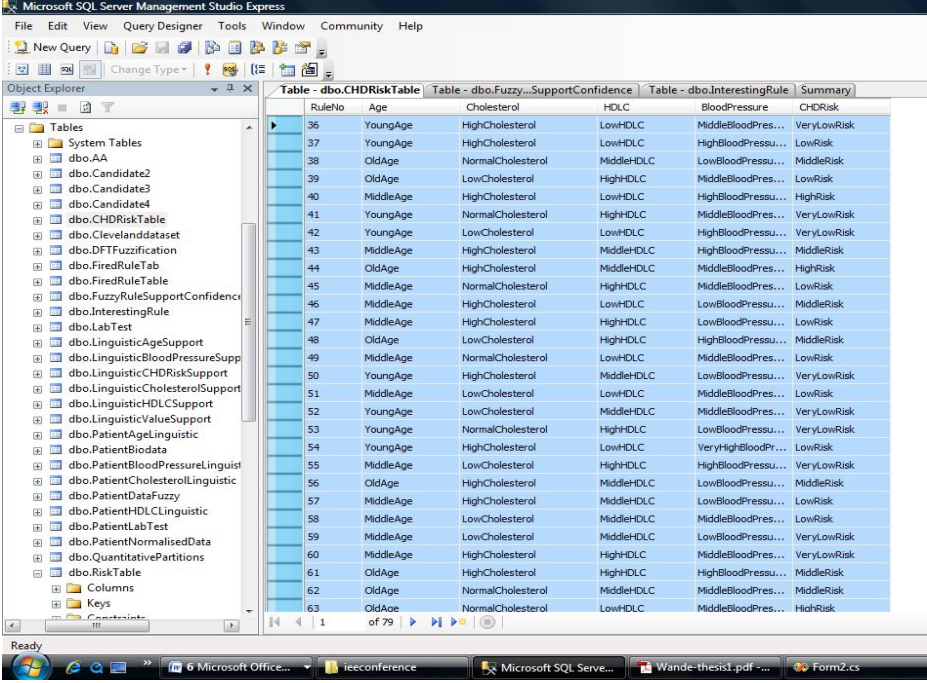
Figure 4.21: The number of rules against the confidence at a constant support zero.

Table 4.6: The extracted rules from Mining system

S/no	Rules
1	<i>OldAge,HighCholesterol,HighHDLc,MiddleBloodPressure-> MiddleRisk</i>
2	<i>OldAge,HighCholesterol,HighHDLc,HighBloodPressure-> MiddleRisk</i>
3	<i>YoungAge,LowCholesterol,LowHDLc,HighBloodPressure-> VeryLowRisk</i>
4	<i>YoungAge,NormalCholesterol,LowHDLc,HighBloodPressure-> VeryLowRisk</i>
5	<i>MiddleAge,LowCholesterol,LowHDLc,HighBloodPressure-> LowRisk</i>
6	<i>MiddleAge,NormalCholesterol,LowHDLc,MiddleBloodPressure-> LowRisk</i>
7	<i>MiddleAge,HighCholesterol,LowHDLc,LowBloodPressure-> MiddleRisk</i>
-	-----
-	-----
-	-----
77	<i>MiddleAge,HighCholesterol,HighHDLc,LowBloodPressure-> LowRisk</i>
78	<i>OldAge,NormalCholesterol,HighHDLc,HighBloodPressure-> MiddleRisk</i>
79	<i>OldAge,HighCholesterol,HighHDLc,LowBloodPressure-> LowRisk</i>

4.3.4 Knowledge-base for prototype system

The extracted rules from FARME-D process were transformed into a relational structure. The knowledge-base changes often as new facts are identified and confirmed by the domain experts. This enhances the dynamism of the knowledge-base as against the static mode of the existing systems with 108 rules (Allahverdi et al, 2007). The rules are interpreted into a relational database in form of tuples and attributes to evolve the knowledge-base. The first four (4) attributes represent the four determinant factor linguistic values. The last attribute represents the CHD risk ratio linguistic values. Each record represents a unique rule. The rule-base has four (4) antecedent linguistic values and an atomic consequence. The snapshot of the rule-base storage on SQL server platform is shown in Figure 4.22.



RuleNo	Age	Cholesterol	HDLC	BloodPressure	CHDRisk
36	YoungAge	HighCholesterol	LowHDLC	MiddleBloodPres...	VeryLowRisk
37	YoungAge	HighCholesterol	LowHDLC	HighBloodPresu...	LowRisk
38	OldAge	NormalCholesterol	MiddleHDLC	LowBloodPresu...	MiddleRisk
39	OldAge	LowCholesterol	HighHDLC	MiddleBloodPres...	LowRisk
40	MiddleAge	HighCholesterol	LowHDLC	HighBloodPresu...	HighRisk
41	YoungAge	NormalCholesterol	HighHDLC	MiddleBloodPres...	VeryLowRisk
42	YoungAge	LowCholesterol	LowHDLC	HighBloodPresu...	VeryLowRisk
43	MiddleAge	HighCholesterol	MiddleHDLC	HighBloodPresu...	MiddleRisk
44	OldAge	HighCholesterol	MiddleHDLC	MiddleBloodPres...	HighRisk
45	MiddleAge	NormalCholesterol	HighHDLC	MiddleBloodPres...	LowRisk
46	MiddleAge	HighCholesterol	LowHDLC	LowBloodPresu...	MiddleRisk
47	MiddleAge	HighCholesterol	HighHDLC	LowBloodPresu...	LowRisk
48	OldAge	LowCholesterol	HighHDLC	HighBloodPresu...	MiddleRisk
49	MiddleAge	NormalCholesterol	LowHDLC	MiddleBloodPres...	LowRisk
50	YoungAge	HighCholesterol	MiddleHDLC	LowBloodPresu...	VeryLowRisk
51	MiddleAge	LowCholesterol	LowHDLC	MiddleBloodPres...	LowRisk
52	YoungAge	LowCholesterol	MiddleHDLC	MiddleBloodPres...	VeryLowRisk
53	YoungAge	NormalCholesterol	HighHDLC	LowBloodPresu...	VeryLowRisk
54	YoungAge	HighCholesterol	LowHDLC	VeryHighBloodPr...	LowRisk
55	MiddleAge	LowCholesterol	HighHDLC	HighBloodPresu...	VeryLowRisk
56	OldAge	HighCholesterol	MiddleHDLC	LowBloodPresu...	MiddleRisk
57	MiddleAge	HighCholesterol	MiddleHDLC	LowBloodPresu...	LowRisk
58	MiddleAge	LowCholesterol	MiddleHDLC	MiddleBloodPres...	LowRisk
59	MiddleAge	LowCholesterol	MiddleHDLC	LowBloodPresu...	VeryLowRisk
60	MiddleAge	HighCholesterol	HighHDLC	MiddleBloodPres...	VeryLowRisk
61	OldAge	HighCholesterol	HighHDLC	HighBloodPresu...	MiddleRisk
62	OldAge	NormalCholesterol	MiddleHDLC	MiddleBloodPres...	MiddleRisk
63	OldAge	NormalCholesterol	LowHDLC	MiddleBloodPres...	HighRisk

Figure 4.22: The snapshot of the rule-base on SQL server platform

4.3.5 Fuzzy Inference process for prototype system

The prototype was developed with C# programming language on Visual studio, 2008 platform. Reasoning in a fuzzy expert system includes three stages: fuzzification, inference, and defuzzification. For the fuzzification process, trapezoidal membership function was used to model all the input linguistic variables and triangular membership function for the linguistic output variable based on the expert instructions and literature. The Mandani fuzzy inference engine was adopted for modelling the expert system. At the inference stage, the MIN method operator was used for the combination of rule's conditions, to determine the membership value of the conclusion, and the MAX method operator was used for rules aggregation. At the defuzzification stage, the centroid method was adopted to get the numerical output for CHD risk ratio. The detail has been critically explained in section 3.3.8. The knowledge-base was modelled with rule-base approach and populated with the interesting rules generated from FARME-D approach process (see Figure 4.22). SQL server 2005 was used as the database management system for data storage. A database titled *Research database* was created with 29 Tables and 2 store procedures. The input to the inference engine is a crisp record of the individual patient and the output is a crisp risk value for the patient. The intermediate outputs of the step by step processes of inference process are also displayed on the inference system interface. The snapshot is shown in Figure 4.23.

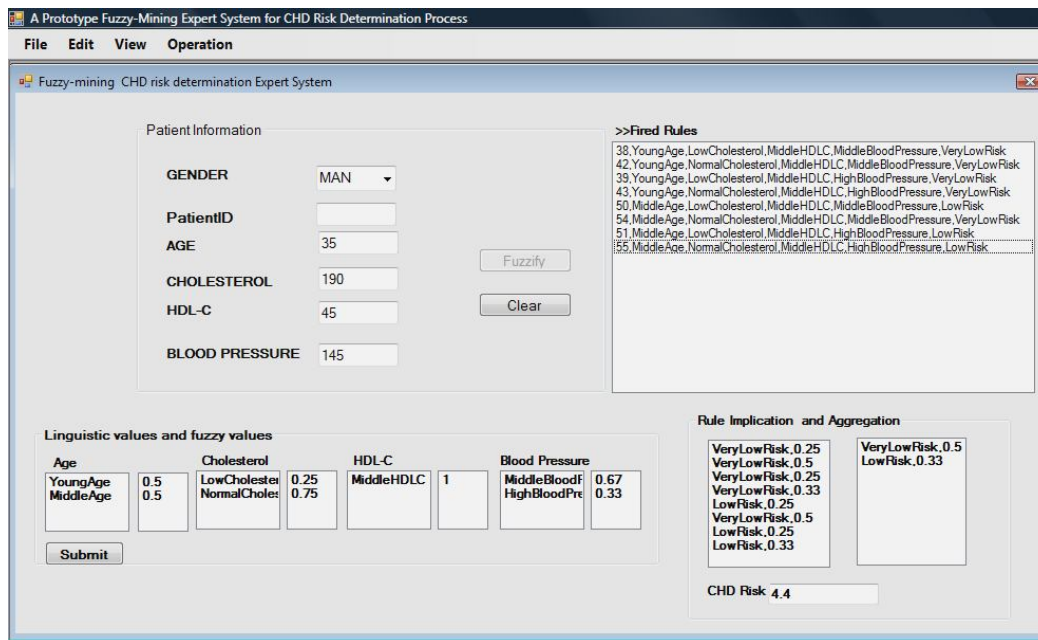


Figure 4.23: The snapshot of the inference process

4.3.6 Application Result and Discussion

The prototype system was tested using test case approach. The test cases are 20 records from non-smoking men outside the mining data set to examine the completeness of the knowledge-base. The mining process extracted 79 interesting rules as against 108 rules determined by standard structure of a fuzzy rule-base formulation used in Allahverdi et al. (2007) as discussed in section 4.2.1(c). The essence of these test cases is to determine the accuracy similarity percentage between the result we are going to have from our proposed approach with 79 rules and 108 rules knowledge-base. All other factors remain constant. The results of the test cases are shown on Table 4.7. Columns 6, 7, and 8 show the actual crisp values for each record according to the three approaches investigated (ATP III result, representing the domain expert decision, FES with 108 rules represents the standard rule-base formulation knowledge acquisition approach and the FES with 79 rules represent the proposed approach). Columns 9, 10, and 11 show the linguistic values for the three approaches. The linguistic values

represent the fuzzy values for each approach decision. The fuzzy value is the more appropriate interpretation for medical decisions.

It was observed that for the test cases, the 79 rules and 108 rules fuzzy systems gave the same risk ratio value except in one instance regardless of the number of rules. This implies that there exist 29 redundant rules among 108 rules which could make the knowledge-base unwieldy and negatively affect the system response time. The graphical interpretation of the result as well as the evaluation report is shown in chapter five.

Table 4.7: Non-Smoking men Test Case

Patient no	Age	Cholesterol	HDLC	Blood Pressure	ATP III	FES With 108 Rules	FES With 79 Rules	ATP III CHD risk Linguistic value	FES CHD risk Linguistic value with (108 rules)	FES CHD risk Linguistic value with (79 rules)
1	30	180	37	160	0	1.5	1.5	VeryLow	VeryLow	VeryLow
2	35	190	45	145	0	4.4	4.4	VeryLow	VeryLow	VeryLow
3	48	260	33	120	8	10.1	10.1	Low	Low	Low
4	57	300	67	110	8	9.3	9.3	Low	Low	Low
5	65	250	54	170	18	19.9	19.9	Middle	Middle	Middle
6	75	290	25	135	30	31.5	31.5	High	VeryHigh	VeryHigh
7	30	160	49	160	0	1.5	1.5	VeryLow	VeryLow	VeryLow
8	40	310	33	140	8	15.5	15.5	Low	Middle	Middle
9	55	300	26	200	30	26.9	25.5	High	High	High
10	60	230	39	110	11	11.2	11.2	Low	Low	Low
11	70	210	45	130	16	15.5	15.5	Middle	Middle	Middle
12	30	240	50	150	0	1.5	1.5	VeryLow	VeryLow	VeryLow
13	35	180	65	160	0	5	5	VeryLow	Low	Low
14	45	300	47	155	9	15.5	15.5	Low	Middle	Middle
15	55	300	49	160	16	18.9	18.9	Middle	Middle	Middle
16	65	250	41	140	18	15.6	15.6	Middle	Middle	Middle
17	70	260	38	190	30	28	28	High	High	High
18	44	210	37	180	5	9.2	9.2	Low	Low	Low
19	55	250	30	200	11	18	18	Middle	Middle	Middle
20	66	250	26	200	28	24.6	24.6	High	High	High

4.4 IMPLEMENTATION COMPONENTS AND TOOLS

1. Microsoft Visual Studio is an integrated development environment (IDE) from Microsoft. It can be used to develop console and graphical user interface applications along with Windows Forms applications, web sites, web applications, and web services in both native code together with managed code for all platforms supported by Microsoft Windows. Visual Studio supports different programming languages by means of language services, these languages include C/C++ (via Visual C++), VB.NET (via Visual Basic .NET), C# (via Visual C#), and F#. Support for other languages such as M, Python, and Ruby among others is available via language services installed separately. It also supports XML/XSLT, HTML/XHTML, JavaScript and CSS.
2. C# is a multi-paradigm programming language encompassing imperative, declarative, functional, generic, object-oriented (class-based), and component-oriented programming disciplines. It was developed by Microsoft within the .NET initiative and later approved as a standard by ECMA (ECMA-334) and ISO (ISO/IEC 23270). C# is one of the programming languages designed for the Common Language Infrastructure. C# is intended to be a simple, modern, general-purpose, object-oriented programming language. The language, and implementations provide support for software engineering principles such as strong type checking, array bounds checking, detection of attempts to use uninitialized variables, and automatic garbage collection. Software robustness, durability, and programmer productivity are important. The language is intended for use in developing software components suitable for deployment in distributed environments.

3. Microsoft® SQL Server™ is a database management and analysis system for e-commerce, line-of-business, and data warehousing solutions. SQL Server 2008, the latest version, includes enhanced XML support, integration of .NET Framework objects in databases, improved integration with Microsoft Visual Studio and the Microsoft Office System, as well as improved analysis, reporting, and data integration services.

4.5 SUMMARY AND DISCUSSION

Following the report of our investigation on the effect of sharp boundary problem in medical domain, this chapter discussed the full scope of the application of the FARME-D using a practical case study in medical domain. The prototype system components include; User Interface, Fuzzification, FARME-D engine, knowledge base, inference subsystem and defuzzification. The prototype was developed based on the domain requirement specification. It was modelled towards building a comprehensive fuzzy expert system. The experience and observation gained from the application of demonstrate the potential viability of the FARME-D knowledge approach.

CHAPTER FIVE

EVALUATION OF THE FARME-D APPROACH

5.1 INTRODUCTION

This chapter reports the evaluation result of FARME-D approach in enhancing the quality of Fuzzy expert system. This evaluation is mainly directed towards the fuzzy expert system knowledge-base being the back bone of the system. The chapter presents the evaluation result of quantitative measure of accuracy and comprehensibility over fuzzy expert system with FARME-D approach (FARMES) as against fuzzy expert system with standard rule-base formulation. It also gives the report of statistical analysis of the test cases result.

5.2 EVALUATION OVERVIEW

For the purpose of evaluation as reported in chapter four, test case approach was used to verify the accuracy of the new approach. The test cases consist of 20 non-smoking men record outside the mining dataset to determine the completeness of the knowledge-base. The quantitative measure of comprehensiveness is used to determine the compactness of fuzzy-mining expert system. The accuracy measure is used to determine the probability that the system can correctly make a decision. Also, t-test was carried out to determine the significant difference between FES with 79 rules (FARMES) and ATP III result, FES with 108 rules and the ATP III result. ANOVA test was also carried out to determine if there exists a significant difference between the three alternative results. All are reported in this section.

5.3 MOTIVATION FOR QUANTITATIVE MEASURE OF EVALUATION

Quantitative measures are essential and form the basis for making reliable decisions in software engineering such as fuzzy expert systems (FESs). Quantitative assessment helps us to evaluate the quality of a FES that is not accessible to our intuitive ability. Generally, in constructing a FES, an accuracy measure is a goodness measure that is usually concerned. The accuracy measure implies how good a FES can perform. Comprehensible knowledge representation is a key advantage of FESs over black box schemes such as neural networks. However, the if-then rules of a FES may not be understandable without a careful design. So, accuracy alone may not be sufficient to show the goodness of FESs (Setnes et al., 1998, Jin, 2000 and Roubos &Setnes 2001). Comprehensibility measure is an additional quantitative assessment that indicates whether a FES is understandable. Therefore, in this thesis both accuracy and

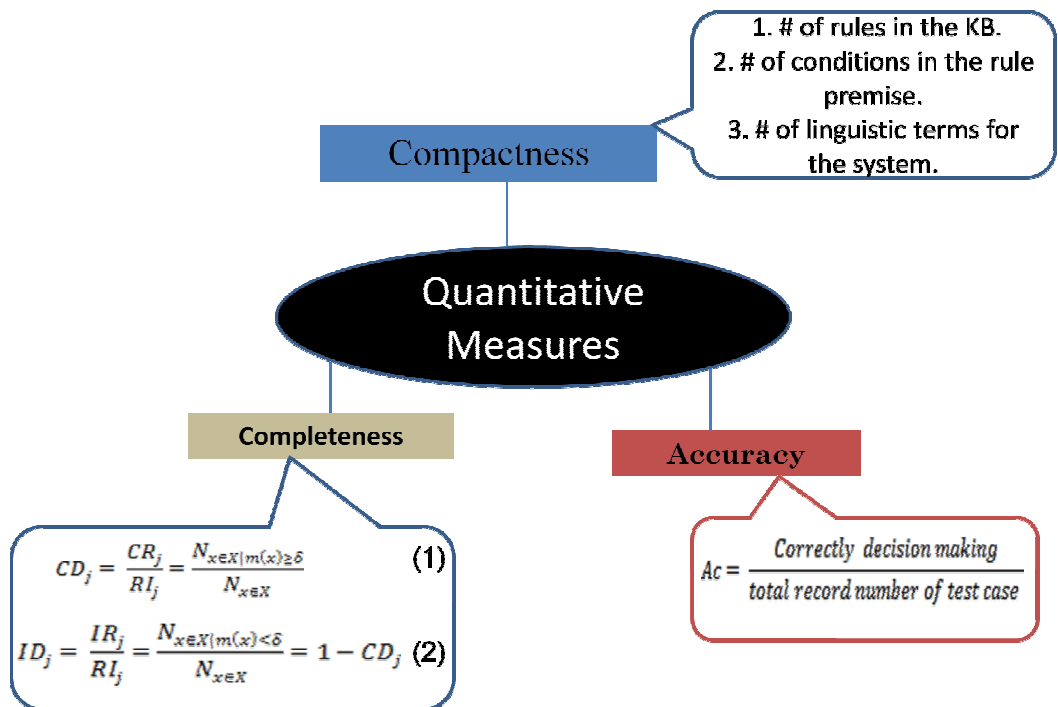


Figure 5.1: Quantitative measures and their models

comprehensibility measure are engaged in evaluating the viability of our proposed approach. Figure 5.1 shows the quantitative measure components and their models.

5.3.1 Quantitative evaluation

To determine the accuracy of the system the result obtained from the test cases are compared with the ATP III results that represent the domain expert result for the test cases. The results of these cases are reported in chapter 4. The graphical representation of the result is shown in Figure 5.2 and 5.3. Figure 5.2 shows the actual risk value for the test cases while Figure 5.3 shows the linguistic expression of the result. The linguistic expression result is very important so as to make the system's result understandable to non-experts users. On Figure 5.3 the linguistic variable are represented with values 1-5 for VeryLow, Low, Middle, High and VeryHigh respectively.

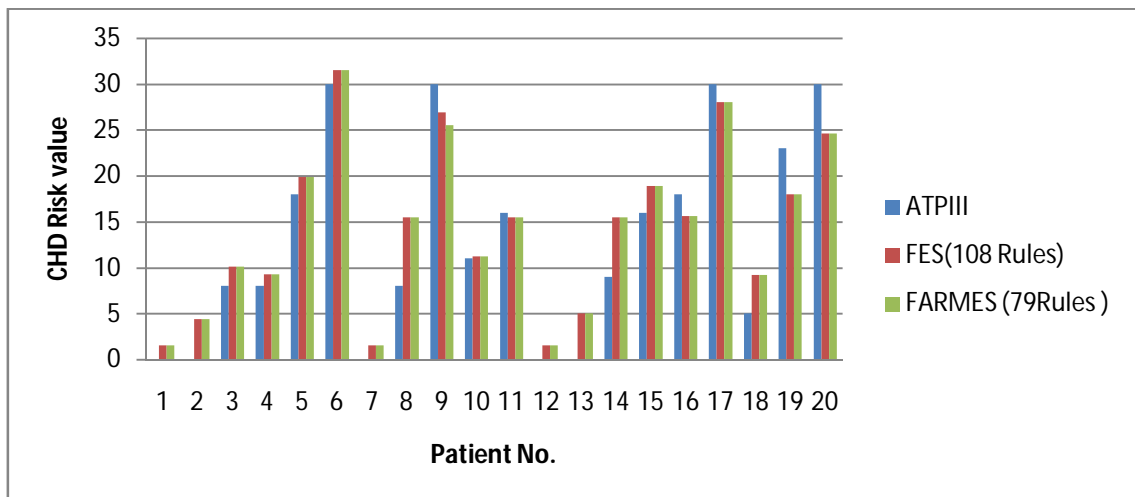


Figure 5.2: ATP III, FES with 108 rules and FES with 79 rules CHD % risk value diagrammatic representation

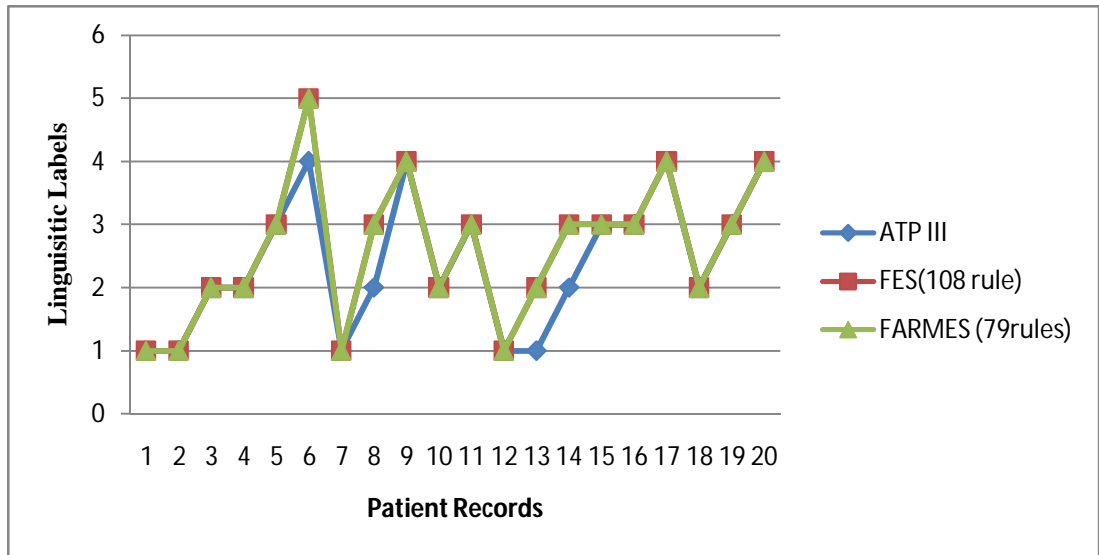


Figure 5.3: The linguistic values for CHD risk diagrammatic representation for ATP, FES with 108 rules and FES with 79 rules

To determine the comprehensibility of the prototype system, three compactness factors are considered; number of linguistic terms in each dimension, number of fuzzy rules in the rule-base and number of conditions in the rule premise. Compactness of fuzzy systems relates to three aspects: a small number of linguistic terms in each dimension, a small number of fuzzy rules in the rule-base and small number of conditions in the rule premise. Also, the completeness of the knowledge-base is determined. The report of these quantitative measures is expressed in Table 5.1.

Table 5.1: Tabular report of quantitative evaluation

Measures	FES with 79 rules (FARMES)	FES with 108 rules
Accuracy	85%	85%
Linguistic term for 5 dimensions	3,3,3,4,5	3,3,3,4,5
Number of rules	79 rules	108 rules
Size on disk	16Kb	20Kb
Conditions in rule premise	4	4
Completeness	65%	100%

5.3.2 Discussion

From Table 5.1 it is observed that the performance of both systems is similar regardless of the number of rules. This presupposes that there exist 29 % redundant rules in the FES that could make the expert system unwieldy and as a result increase the memory usage with 20%. According to Meesad (2001), if all possible rules are utilized in building an expert system, it means the comprehensibility of the system is traded-off. Therefore to enhance the comprehensibility of the prototype system it is important that 29 rules are eliminated.

Also, the completeness quality result indicates that in all test cases FES with 108 rules was able to fire all relevant rules while FES with 79 rules was not. As a result, they yielded the same result. That shows that certain rules are necessary but not important in decision making. Such rules are identified and eliminated during the mining process because they cannot satisfy the minimum threshold. So, such rules are regarded as redundant rules and deprive the system comprehensibility. For instance, Figure 5.4 and Figure 5.5 capture an instance of a test case where 16 rules are necessary but 4 rules are important in decision making.

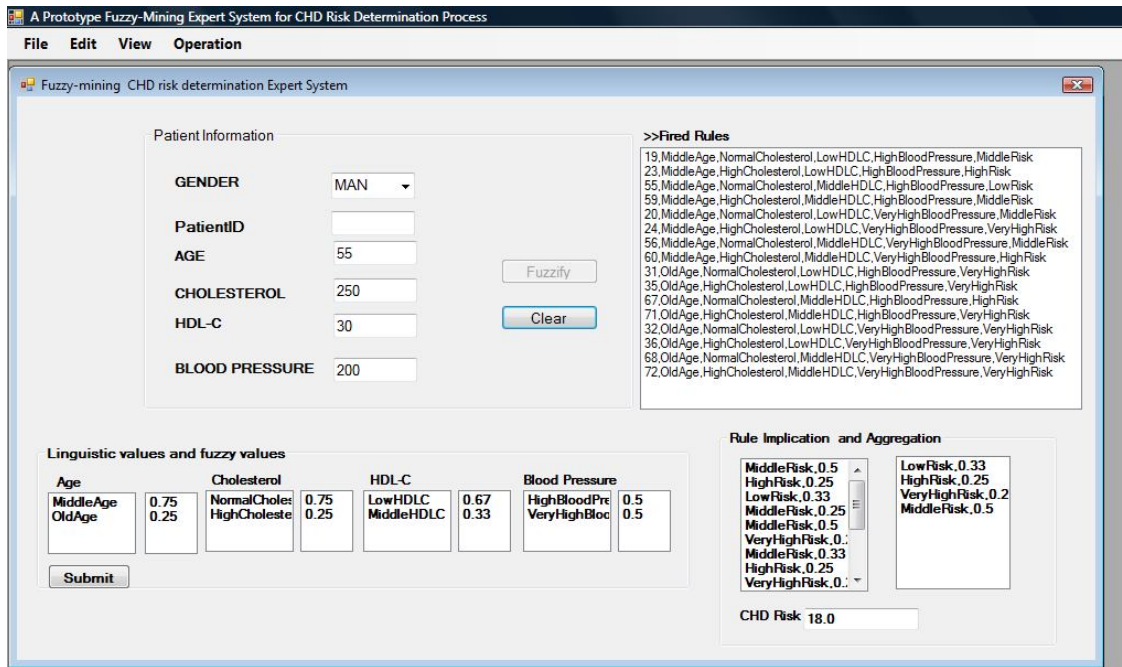


Figure 5.4: The snapshot of FES (108 rules) with 16 rules fired.

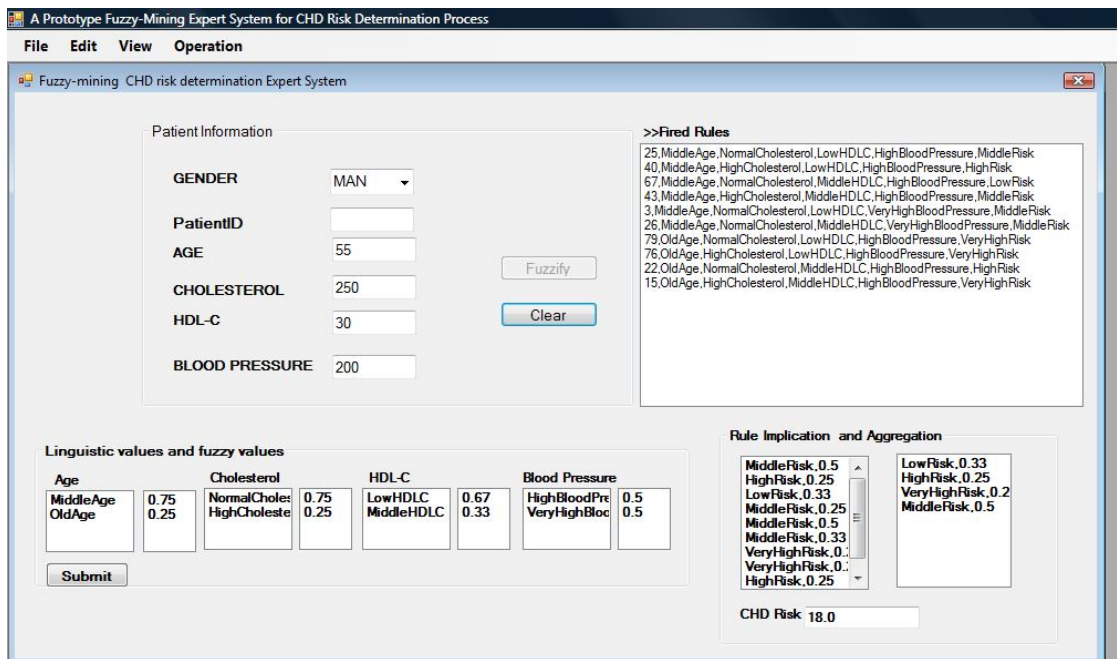


Figure 5.5: The snapshot of FES (79 rules) with 8 rules fired.

5.4 STATISTICAL EVALUATION.

(a) *t- Test*

In the constant quest to reduce variation and improve approaches, there is need to evaluate different alternatives. A t-test using two samples compares two sets of test data. It helps determine if the means (i.e., averages) are the same or different from each other.

The null and alternate Hypotheses are:

- The null hypothesis H_0 is that the mean difference $(1-x_2) = 0$ or in other words the means are the same .
- The alternative hypothesis H_a is that the mean difference $\neq 0$ or in other words the means are not the same

On performing the t-test on ATP III and proposed approach (FES with 79 rules) results with level of significance $\alpha = 0.05$, the result is shown on table 5.2

Table 5.2: t –Test result for ATP III & FES with79 rules

t-Test: Two-Sample Assuming Equal Variances		
	<i>ATP III</i>	<i>FES with 79 rules</i>
Mean	12.3	14.11
Variance	113.5894737	81.09673684
Observations	20	20
Pooled Variance	97.34310526	
Hypothesized Mean Difference	0	
Df	38	
t Stat	-0.580130871	
P(T<=t) one-tail	0.282625357	
t Critical one-tail	1.685954461	
P(T<=t) two-tail	0.565250714	
t Critical two-tail	2.024394147	

Since the null hypothesis is that the mean difference $(1-x_2) = 0$, this is a two-sided test.

Therefore, we used the two-tail values for the analysis. Since the t statistic < t critical

($0.580 < 2.024$) and $p \text{ value} > \alpha$ ($0.565 > 0.05$), we can accept the null hypothesis that the means are the same. Therefore we can say that both ATP III and our new approach give the same result at a 95% confidence level.

(b) ANOVA Test

ANOVA test is used to determine if there's a statistically significant difference between three or more alternatives. Therefore, to determine if there exists a statistically significant difference between ATP III, FES with 108 rules and FES with 79 rules, ANOVA test is appropriate. The null hypothesis is that the means are equal:

- $H_0: \text{Mean1} = \text{Mean2} = \text{Mean3}$

The alternate hypothesis is that at least one of the means is different:

- $H_a: \text{At least one of the means is different}$

The result is summarized as follows

Table 5.3: ANOVA result for ATP III, FES with 108 & FES with 79 rules

Anova: Single Factor

SUMMARY						
Groups	Count	Sum	Average	Variance		
ATP III	20	246	12.3	113.5894737		
FES (108 rules)	20	283.6	14.18	82.87326316		
FARMES (79 rules)	20	282.2	14.11	81.09673684		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	45.436	2	22.718	0.245547374	0.7831	3.1588
Within Groups	5273.63	57	92.51982			
Total	5319.066	59				

Since $f \text{ statistic} < f \text{ critical}$ ($0.246 < 3.16$) $p \text{ value} > \alpha$ ($0.78 > 0.05$) the hypothesis is accepted that their means are the same. This confirms the similar performances of the three approaches.

5.5 POSSIBILITIES FOR GENERALIZATION OF RESULT

Having showed that FARME-D approach is able to eliminate redundant rules for the case study presented in this research, we therefore postulate that FARME-D approach can indeed be applied to acquire knowledge-base in the medical domain and in other application domain, especially where there are a large number of dimensions of determinant factors for decision making.

In all, FARME-D has shown a plausible effort in identifying redundant rules that are necessary but not important for decision making through its mining capability and reduces the knowledge-base storage spaces. Hence, it enhances fuzzy expert system comprehensibility and results into a new architecture called Fuzzy Association Rule Mining Expert System (FARMES).

5.6 SUMMARY AND CONCLUSION

In this chapter a report of the evaluation measures adopted for FARME-D approach and its performance was presented. It has been shown that the elimination of redundant rules which resulted in incompleteness did not affect the accuracy of the prototype system but enhancing its comprehensibility and reduces the storage usage. Furthermore, the case study scenario has demonstrated the applicability of FARMES architecture in a real-life context and proved the viability of FARME-D knowledge-base acquisition approach. FARME-D was able to achieve 27% reduction in the number of rules evolved for the prototype system at a reduced storage usage of 20% while the system accuracy is maintained. This enhances the system comprehensibility. The case study, therefore, successfully validates FARME-D knowledge-acquisition

approach as a platform for modelling a comprehensible fuzzy rule-based expert system in medical domain.

CHAPTER SIX

SUMMARY OF FINDINGS, CONCLUSION AND FUTURE WORK

6.1 SUMMARY

This thesis has shown that knowledge acquisition is an important process in modelling accurate and comprehensible fuzzy expert system. The knowledge-base is referred to as the backbone of the expert systems; therefore, the lesser the redundant rules within the knowledge-base the more compact the expert system and less memory space utilized. Also, ability to update the knowledge-base with newly identified and confirmed knowledge enhances the quality and comprehensibility of the expert system.

However, the issues that determine comprehensibility and instant update of the knowledge-base are less attended to by most knowledge engineers, and this has resulted into non-compact, large memory usage and less understandable fuzzy expert systems.

This thesis intervened by introducing a unified solution approach called FARME-D to attend to issues concerning comprehensibility and knowledge-base instant update. FARME-D incorporates the expert's opinion factors into the existing fuzzy association rule mining process for knowledge acquisition in modelling comprehensible fuzzy expert systems. The integration of FARME-D with standard FES architecture gave birth to a new architecture called FARMES. This architecture provides a platform for elimination of redundant rules (which cause the knowledge-base to become unwieldy), a knowledge-base void of the SBP and corresponds to human perception of the application domain, and enhancement of knowledge-base instant update. FARME-D is dedicated to extracting interesting rules from existing examples (Historical database)

in an application domain based on the domain experts' opinion about the data description and analysis of the domain. FARME-D adopts fuzzy concept and expert-driven approach to avoid sharp boundary problems, ensure understandable rules to non expert and at the same time correspond to human expert perception.

FARME-D comprises five major components which are: application domain historical database, human domain expert, fuzzification engine, expert-driven data mining engine and rule interpretation engine, as show in Figure 3.1. In addition, FARME-D is based on a set of assumptions which defines the conditions for its optimal applicability.

These are:

- The determinant factors for solving problems are known and predetermined in advance by the domain experts.
- Data stored in organizations are quantitative in nature and growing in an increasingly rapid way with increasing number of variables.
- Organizations have historical data bank where the past human experts' experiences could be retrieved.
- The historical data set are in a structured form.

The thesis provides a validation of the FARME-D automated knowledge acquisition by using a case study of CHD in medical domain in order to demonstrate the applicability and viability of approach in real-life context.

The thesis made some significant contributions. Firstly, it has opened up a new perspective on how to tackle the problem of unwieldiness in rule-base expert system by offering a clear demonstration of the viability of fuzzy association rule mining expert-driven approach as the solution to this problem. Secondly, an innovative approach of knowledge acquisition was introduced to ensure instant update of

knowledge-base as new experience is acquired by the domain experts. Third, this work has introduced expert-driven approach to existing fuzzy association rule mining process which allows the extracted rules to emulate human cognitive process of decision making ability. This will also alleviate the effect of SBP in medical expert system. Lastly, the integration of FARME-D component to standard expert system architecture has resulted into a new Fuzzy Association Rule Mining Expert System (FARMES) architecture. This provides for knowledge-base instant update. This thesis also makes a first attempt to develop a prototype fuzzy association rule mining expert system for determination of Coronary Heart Disease risk ratio in medical domain.

6.2 CONCLUSION

According to Meesad (2001), if all possible rules are utilized in building an expert system, it means the comprehensibility of the system is traded-off. Therefore, to enhance the comprehensibility of medical fuzzy expert system it is important that redundant rules are eliminated.

The research has provided a theoretical and design-oriented framework that can be adopted for modelling new generation fuzzy expert systems in medical domain and others. It has also made vital contributions to three concerns in the rule-base knowledge engineering industry, these are:

- the large number of rules in the knowledge-base which causes rule-based expert system unwieldiness
- need for a knowledge-base void of the SBP and that corresponds to human perception of the application domain

- the need for instant update of the knowledge-base as new experiences are acquired by the domain experts.

Finally, if the proposed automated knowledge acquisition approach (FARME-D) which is the result of this research endeavour is adopted, it will give quality boost, needed in the rule-based expert system engineering, to modeling a comprehensible fuzzy expert system. Also, the integration of FARME-D structure with the standard FES architecture will result in a derived Fuzzy Association Rule Mining Expert System (FARMES). This will enhance the instant update of the knowledge-base and the credibility of the expert system.

6.3 FUTURE WORK

The thesis provides several opportunities for further research in the immediate future. The FARME-D approach, as modelled and implemented in this thesis, directly inherited some limitations from its parent concept of fuzzy association rule mining expert-driven approach. Notably, there exist ample of research possibilities to enhance the concept in the following areas:

- Mining process: extension of the mining process to involve text mining, image mining, voice mining and web mining in order to extend the scope of knowledge acquisition which will turn out to enrich the knowledge-base.
- Knowledge representation: extending the knowledge representation beyond production rule representation to semantic net and case bases.

REFERENCES

- .Gyenesei., A. (2001) “A fuzzy approach for mining quantitative association rules”.
Acta Cybernetica, 15:305–320.
- Abraham, A. (2001) “Neuro-Fuzzy Systems: State-of-the-Art Modeling Techniques, Connectionist Models of Neurons, Learning Processes, and Artificial Intelligence”, in Lecture Notes in Computer Science, Vol. 2084, (eds. Mira., Jose and Prieto., Alberto) Springer Verlag, Germany. pp. 269–276 .
- Abraham, A. (2005) “ Rule-based Expert Systems” Handbook of Measuring System Design, edited by Peter H. Sydenham and Richard Thorn. John Wiley & Sons, Ltd. ISBN: 0-470-02143-8. pp. 910-919.
- Adeli, A and Neshat, M. (2010) “A Fuzzy Expert System for Heart Diagnosis” Proceedings of the International MultiConference of Engineers and Computer Scientists Vol. 1, IMECS. Hong Kong. ISBN: 978-988-17012-8-2.
- Agrawal, R. and Srikant, R. (1994) “Fast algorithms for mining association rules”, Proc. Int. Conf. Very Large Data Bases (VLDB’94), Santiago, Chile, September, pp.487–499.
- Agrawal, R., Imielinski, T. and Swami, A. (1993) “Mining association rules between sets of items in large databases”, Proc. ACM-SIGMOD Int. Conf. Management of Data (SIGMOD’93), Washington, DC, July, pp.45–51.
- Allahverdi, N., Torun, S. and Saritas, I. (2007) “Design a Fuzzy Expert System for Determining of Coronary Heart Disease Risk” ACM International Conference Proceeding Series; Vol. 285, ISBN:978-954-9641-50-9 Proceedings of the 2007 international conference on Computer systems and technologies, Article No. 36.

- Aly. S. and Vrana, I. , (2006) “Toward efficient modeling of fuzzy expert systems: a survey” *AGRIC. ECON.- CZECH*, vol. 52(10). Pp. 456-460.
- Arias-Aranda, D., Castrol, J.L. , Navamo, M., Sanchez, J.M. and Zurita, J.M. (2010) “ A fuzzy Expert for business Management” *Expert System Application*. Vol.37(12) 7570-7580.
- Bayliss, J. (2001) “ Framingham risk score to predict 10 year absolute risk of CHD event west hertfordshire cardiology” from Wilson PWF, et al Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;97:1837-47.
- Buchanan BG and Shortliffe EH. (1984) “ Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project”. Addison-Wesley, 1984.
- Chan, K. C.C. and Au, W., (1998) “An Effective Algorithm for Discovering Fuzzy Rules in Relational Databases” *Fuzzy Systems Proceedings. IEEE World Congress on Computational Intelligence, 1998*
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., (1999) “CRISP-DM1.0: Step-by-step data mining guide” CRISP-DM consortium.
- Chen, G. and Wei, Q.(2002) “ Fuzzy association rules and the extended mining algorithms” *Information Sciences*, 147(1–4):201–228.
- Chi X., Haojun M, Zhen Z., Yinghong P. (2001) “Research on hybrid expert system application to blanking technology” *Journal of Material Processing Technology*, 116 (2): 95–100.
- Christoph, S. H (1995) “ A Hybrid Fuzzy-Neural Expert System for Diagnosis” *Proceedings of the international Joint Conference on Artificial Intelligence*

(IJCAI), Montreal, Canada Pages 494-500 ISBN ~ ISSN:1045-0823 , 978-1-558-60363-9

Cock, M. De, Cornelis, C and Kerre, E.E. (2003) “Fuzzy association rules: A two-sided approach,” in Proc. Int. Conf. Fuzzy Information Processing—Theories and Applications, Beijing, China, pp. 385–390.

Dashofy, E.M., van der Hoek, A. and Taylor, R.N., (2001) “A Highly-Extensible XML-Based Architecture Description Language”, Proceedings. Working IEEE/IFIP Conference on Software Architecture, Amsterdam, Netherlands: pp103-112.

Davis R and King JJ.(1984) “ The Origin of Rule-Based Systems in Artificial intelligence”. In [Buchanan and Shortliffe, 1984], pp. 20-52.

De Kock, E. (2003) “A Web-based Knowledge Based Decision Support System (KB-DSS)” University of Pretoria etd.

Delgado , M. Sanchez D., Maria J., Martin-Bautista and Vila MA. (2001) “Mining association rules with improved semantics in medical database” Artificial Intelligence in Medicine 21, pp.241-245.

Delgado, M. Marin, N. Sanchez, D. and Vila, MA.(2003) “Fuzzy Association Rules, General Model and Applications”, IEEE Transactions on Fuzzy Systems, 11(2), pp.214–225.

Duan, Y., Edwards, J.S. and Xu, M.X. (2005) “Web-based expert systems: benefits and challenges” Information and Management (42).pp.799-811.
www.elsevier.com/locate/dsw.

Durkin, J. (1994) “Expert Systems Design and Development” Prentice Hall, Englewood Cliffs, NJ.

- Fatica, N. S., Ichida, F., Engle, M. A. and Lesser, M.L. (1989) "Rug Shampoo and Kawasaki Disease" *Pediatrics*, official journal of the American Academy of Pediatrics, ; 84; 231-234.
- Feigenbaum E. A. (1982) "Knowledge Engineering in the 1980s" Department of Computer Science, Stanford University, Stanford, CA.
- Gadaras, I. and Mikhailov, L.(2009) "An Interpretable fuzzy rule-based classification methodology for medical diagnosis", *Artificial intelligence in medicine* 47(1). pp. 25-41
- Garg, A., Critchlow, M., Chen, P., Van der Westhuizen, C., van der Hoek, A.(2003) "An Environment for Managing Evolving Product Line Architectures", *Proceedings of the International Conference on Software Maintenance*, Amsterdam, The Netherlands, pp. 358-367.
- Garlan, D., Monroe, R. and Wile, D., (1997) "ACME - An Architecture Description Interchange Language", In *Proceedings of CASCON '97*, ACM Press: pp.169-183.
- Giarratano, J. and Riley, G. (1989) *Expert Systems: Principles and Programming*, PWS-Kent Publishing Co, Boston, MA.
- Gyenesi, A. (2000) "A Fuzzy Approach for Mining Quantitative Association Rules". *Turku Centre for Computer Science Technical Reports*.
- Han, J. and Kamber, M. *Data Mining (2001) "Concepts and Techniques"* ©2001 (c) Morgan Kaufmann Publishers.Simon Fraser University.
- Harleen, K, & Siri, K.W. (2006) "Empirical Study on Applications of Data mining Techniques in Healthcare" *Journal of Computer Science* 2(2). ISSN 1549-3636, pp. 194-200.

- He, Y, Tang, Y., Zhang, Y and Sunderraman, R. (2006) “ Adaptive Fuzzy Association Rule mining for effective decision support in biomedical applications” *Int. J. Data Mining and Bioinformatics*, Vol. 1, No. 1.
- Hen, Tzung-Pei, Kuo, Chan-Sheng; Chi, Sheng-Chai(1999) “ A Fuzzy Data Mining Algorithm for Quantitative Values” *Knowledge-Based Intelligent Information Engineering Systems, Third International Conference*.
- Ioannis G. and Ludmil M, (2009) “ An interpretable Fuzzy rule-based classification methodology diagnosis. *Artificial Intelligence in medicine*, Vol 47(1), pp. 25-41
- Jang, J.S.R., Sun, C.T. and Mizutani, E. (1997) “*Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*”, Prentice Hall Inc, USA.
- Jin, Y. (2000) “Fuzzy modeling of high-dimensional systems: complexity reduction and interpretability improvement,” *IEEE Trans. Fuzzy Syst.*, vol. 8, no. 2, pp. 212-221.
- Jin, Y. Seelen, W. V. and Sendhoff, B. (1999) “On generating FC3 fuzzy rule systems from data using evolution strategies,” *IEEE Trans. Syst., Man, Cybern. B*, vol. 29, no. 6, pp. 829-845,
- Kaula R., Lander L.C. (1995) “A module-based conceptual framework for large-scale expert systems” *Industrial Management & Data Systems*, 95 (2):15–23.
- Klein, M.R. and Merhlie, L.B. (1995) “*Knowledge-based Decision Support Systems with Applications in Bussiness*”. 2nd Edition. England: John Wiley & Sons.
- Kosko, B. (1992) “*Neural Networks and Fuzzy Systems. A Dynamical Systems Approach to Machine Intelligence*”, 1st ed., Prentice-Hall, Englewood Cliffs, NJ.
- Koutsojannis C. and Hatzilygeroudis I.(2006) “Fuzzy-Evolutionary Synergism in an Intelligent Medical Diagnosis System” *Lecture Notes in Computer Science*,

Knowledge-Based Intelligent Information and Engineering Systems Springer
Berlin Heidelberg, 0302-9743 (Print) 1611-3349 (Online) Volume 4252/2006.
pp. 1313-1322

Kuok, C., Fu, A. and Wong M. (1999) “ Mining fuzzy association rules in databases”.
SIGMOD Record 17(1): 41-6. (Downloaded from
<http://www.acm.org/sigs/sigmod/record/issues/9803> on 1 March 1999).

Lavrac N, Keravnou ET, Zupan B. (1996) “Intelligent data analysis in medicine and
pharmacology: an overview. Dordrecht: Kluwer Academic Publishers” p. 1±13.

Meesad, M. (2001) “Quantitative measures of a Fuzzy Expert System” Expert
Systems.pp.1-6

Moein, S., Monayjemi, S.A. and Moallem, P. (2008) “ A Novel Fuzzy-Neural Based
Medical Diagnosis System” Proceedings of World Academy of Science,
Engineering and Technology. Vol. 27. ISSN 1307-6884. pp. 157-161

Negnevitsky M. (2005) “Artificial Intelligence: A guide to Intelligent Systems” 2nd
Edition. England, Pearson Education Limited, Edinburgh Gate, Harlow. Addison
Wesley.

Neshat, M. and Yaghobi, M., (2009) “Designing a Fuzzy Expert System of Diagnosing
the Hepatitis B Intensity Rate and Comparing it with Adaptive Neural Network
Fuzzy System” Proceedings of the World Congress on Engineering and
Computer Science 2009 Vol II WCECS 2009,pp 797-802 October 20-22, 2009,
San Francisco, USA

Neves, J., Alves, V., Nelas, L., Romeu, A., and Basto, S. (1999) “An Information
System That Supports Knowledge Discovery and Data Mining in Medical
Imaging”Machine Learning and Applications: Machine Learning in Medical
Applications. Chania, Greece, pp. 37-42.

- Newell, A. and Simon, H.A. (1972) “ Human Problem Solving” Prentice Hall, Englewood Cliffs, NJ.
- Norbik, B. I. and Bharanidharan, S.(2005) “Novel Attack Detection Using Fuzzy Logic and Data Mining “Proceedings of NAFIPS 19th International Conference of the North American Fuzzy Information Processing Society, Atlanta.
- Ohsaki, M. , Abe, H., Tsumoto, S.,Yokoi, H. and Yamaguci, T., (2007) “Evaluation of rule Interestestingness measures in medical knowledge in database” Artificial Intelligence in Medicine 41, 177-196.
- Oladipupo, O.O., Uwadia, O.C and Ayo, C.K. (2010) “On sharp boundary problem in rule-based expert systems in medical domain”. International journal of Healthcare Information Systems and Informatics, 5(3) 14-26.
- Partridge, D., Abidi, S. S. R., and Goh, A. (1996) “Neural Network Applications in Medicine”. Proceedings of National Conference on Research and Development in Computer Science and Its Applications (REDECS’96), Universiti Pertanian Malaysia: Kuala Lumpur, pp. 20 - 23.
- Pirnau, M., Maiorescu, T. (2008) “Methods for Knowledge Acquisition” Metalurgia International Vol. XIII(2008), no 9, pp. 99.
- Roubos, H. and Setnes, M.(2001) “Compact and transparent fuzzy models and classifiers through iterative complexity reduction”, IEEE Trans. Fuzzy Syst., vol. 9, no. 4 (2001) 516-524.
- Rudas, I. J. and Fodor J. (2008) “ Intelligent Systems” Int. J. of Computers. Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844 Vol. m, Suppl. issue: Proceedings of ICCCC 2008, pp. 132-138
- Saritas, I. Allahverdi, N. and Sert, U. (2003) “A Fuzzy Expert System Design for Diagnosis of Prostate Cancer”, in Proc. Intern. Conference on Computer Systems

and Technologies CompSysTech'2003- CompSysTech'2003, Sofia, Bulgaria, 18-20 June.

Sasikumar, M., Ramani, S., Anjeneyulu, KSR., Chandrasekar, R., and Muthu Raman, S. (2007) "A Practical Introduction to Rule Based Expert" Published by: Narosa Publishing House, New Delhi
All rights reserved. Copyright 2007, Narosa Publishers.

Schneider, M., Lagholz, G., Kandel, A. and Chew, G. (1996) "Fuzzy Expert System Tools", 3rd ed., John Wiley & Sons, New York, NY.

Setiawan, N.A., Venkatachalam, P.A., and Hani, A. F. M (2009) "Diagnosis of Coronary Artery Disease Using Artificial Intelligence Based". Proceedings of the International Conference on Man-Machine Systems (ICoMMS) 11 – 13 October 2009, Batu Ferringhi, Penang, MALAYSIA

Setnes, M. Babuska, R. and Verbruggen, H. B.(1998) "Rule-based modeling: precision and transparency," IEEE Trans. Syst., Man, Cybern. C, vol. 28, no. 1, pp. 165-169.

Shah, S., Roy, R. and Tiwari, A. (2006) 'Development of fuzzy expert system for customer and service advisor categorization within contact center environment', Applications of soft computing: Recent Trends. ISSN print 1615-3871, ISSN Electronic 1860-0794, ISBN-10-3-540-29123-7 Springer Berlin Heidelberg, New York. pp. 197-206.

Shi, Y. Eberhart, R. and Chen, Y. (1999) "Implementation of Evolutionary Fuzzy Systems". IEEE Transactions on Fuzzy Systems, VOL. 7, NO. 2, APRIL 1999

Siti Fatimah Md Saad and Rogayah Ghazali (1999) "Data Mining for Medical Database" Proceedings of the First National Conference on Artificial Intelligence Application in Industry. Kuala Lumpur, pp. 72-79.

- Siti Nurul Huda Sheikh Abdulah and Miswan Surip (1999) "Satu Metodologi Perlombongan Data Untuk Pesakit AIDS" Proceedings of the First National Conference on Artificial Intelligence Application in Industry. Kuala Lumpur, pp. 57-71.
- Soe, S.M.M. and Zaw M.P.P (2008) "Design and Implementation of Rule-based Expert System for Fault Management" World Academy of Science, Engineering and Technology 48 2008.
- Srikant, Ramakrishnan; Agrawal, Rakesh(1996) "Mining Quantitative Association Rules in Large Relational Tables" Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data.
- Stamou, G. B and Tzafestas, S. G. (1999) "Fuzzy relation equations and fuzzy inference systems: an inside approach," IEEE Trans. Syst., Man, Cybern. B, vol. 29, no. 6, pp. 694-702.
- Turban, E. and Aronson, J.F. (2001) " Decision Support Systems and Intelligent Systems, 6th edition. Prentice Hall Englewood Cliffs, NJ.
- Turban, E., MecLean, EA. and Werherbe, J (2001) "Information Technology for Strategic Advantage". 2nd Edition. New York, Chichester, Winheim, Brisbane, Singapore, Toronto: John Wiley & Sons, Inc.
- Verlinde, H., De Cock, M. and Boute , R. (2006) "Fuzzy Versus Quantitative Association Rules: A Fair Data-Driven Comparison" IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, Vol. 36, No. 3. pp.679-684
- Walker, N. J., and Kwon, O. (1997) "ISS: An Expert System for the Diagnosis of Sexually Transmitted Diseases" 11th Annual Midwest Computer Conference (MCC'97) March 21, Springfield, Illinois.

- Wang, L.X. and Mendel, J.M. (1992) “Generating fuzzy rules by learning from examples”, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 22, No. 6, pp.1414–1427.
- Waterman, D.A. (1986) “ A Guide to Expert Systems”, Addison-Wesley, Reading, MA.
- Wilson, P.W., D’Agostino, R.B., Levy, D. Belang, A.M, Silbershatz, H and Kannel W.B. (1998) ‘Prediction of coronary heart disease using risk factor categories’, *Circulation*, Vol. 97, pp.1837–1847.
- Yan, H., Jiang, Y., Zbeng, J., Peng . C. and Li, Q. (2006) “A multilayer perceptron-based medical decision support system for heart disease diagnosis” *Expert Systems with Applications* 30 (2006) 272–281. Elsevier.
- Zadeh, L.A. (1965) “Fuzzy Sets” *Information and Control*, 8, 338–353.

Appendix

Appendix A: ATP III Guidelines At-A-Glance Quick Desk Reference

Men			Women		
Estimate of 10-Year Risk for Men			Estimate of 10-Year Risk for Women		
(Framingham Point Scores)			(Framingham Point Scores)		
Age	Points		Age	Points	
20-34	-3		20-34	-7	
35-39	4		35-39	3	
40-44	1		40-44	0	
45-49	3		45-49	3	
50-54	5		50-54	6	
55-59	8		55-59	8	
60-64	10		60-64	10	
65-69	11		65-69	12	
70-74	12		70-74	14	
75-79	13		75-79	16	
Total Cholesterol			Total Cholesterol		
	Age 20-29	Age 40-49	Age 50-59	Age 60-69	Age 70-79
<160	0	0	0	0	0
150-199	4	3	2	1	0
200-239	7	5	3	1	0
240-279	9	6	4	2	1
≥280	11	8	5	3	1
Points			Points		
	Age 20-29	Age 40-49	Age 50-59	Age 60-69	Age 70-79
Non-smoker	0	0	0	0	0
Smoker	5	5	3	1	1
HDL (mg/dL)			HDL (mg/dL)		
	Points			Points	
≥60	-1		≥60	-1	
50-59	0		50-59	0	
40-49	1		40-49	1	
<40	2		<40	2	
Systolic BP (mmHg)			Systolic BP (mmHg)		
	If Untreated	If Treated		If Untreated	If Treated
<120	0	0	<120	0	0
120-129	0	1	120-129	1	3
130-139	1	2	130-139	2	4
140-159	1	2	140-159	3	5
≥160	2	3	≥160	4	6
Point Total			Point Total		
	10-Year Risk %			10-Year Risk %	
<0	< 1		< 0	< 1	
0	1		0	1	
1	1		1	1	
2	1		1	1	
3	1		1	1	
4	1		1	1	
5	2		1	2	
6	2		2	2	
7	3		2	3	
8	4		3	4	
9	5		4	5	
10	6		5	6	
11	8		6	8	
12	10		7	10	
13	12		8	12	
14	16		9	16	
15	20		10	20	
16	25		11	25	
≥17	≥ 30		12	≥ 30	
		10-Year risk _____%			10-Year risk _____%