

FINDING PATTERNS IN STUDENT AND MEDICAL OFFICE DATA USING ROUGH SETS

by

Anwar Alenezi

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science (MSc) in Computational Science

The Faculty of Graduate Studies
Laurentian University
Sudbury, Ontario, Canada

© Anwar Alenezi, 2014

THESIS DEFENCE COMMITTEE/COMITÉ DE SOUTENANCE DE THÈSE

Laurentian Université/Université Laurentienne
Faculty of Graduate Studies/Faculté des études supérieures

Title of Thesis
Titre de la thèse FINDING PATTERNS IN STUDENT AND MEDICAL OFFICE DATA USING
ROUGH SETS

Name of Candidate
Nom du candidat Alenezi, Anwar

Degree
Diplôme Science

Department/Program
Département/Programme Computational Sciences Date of Defence
Date de la soutenance October 2, 2014

APPROVED/APPROUVÉ

Thesis Examiners/Examineurs de thèse:

Julia Johnson
(Supervisor/Directeur(trice) de thèse)

Youssou Gningue
(Committee member/Membre du comité)

Abdalla Mansur
(Committee member/Membre du comité)

Approved for the Faculty of Graduate Studies
Approuvé pour la Faculté des études supérieures
Dr. David Lesbarrères
M. David Lesbarrères
Acting Dean, Faculty of Graduate Studies
Doyen intérimaire, Faculté des études supérieures

Dr. Mehmet Resit Tolun
(External Examiner/Examineur externe)

ACCESSIBILITY CLAUSE AND PERMISSION TO USE

I, **Anwar Alenezi**, hereby grant to Laurentian University and/or its agents the non-exclusive license to archive and make accessible my thesis, dissertation, or project report in whole or in part in all forms of media, now or for the duration of my copyright ownership. I retain all other ownership rights to the copyright of the thesis, dissertation or project report. I also reserve the right to use in future works (such as articles or books) all or part of this thesis, dissertation, or project report. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that this copy is being made available in this form by the authority of the copyright owner solely for the purpose of private study and research and may not be copied or reproduced except as permitted by the copyright laws without written authority from the copyright owner.

Abstract

Data have been obtained from King Khaled General Hospital in Saudi Arabia. In this project, I am trying to discover patterns in these data by using implemented algorithms in an experimental tool, called Rough Set Graphic User Interface (RSGUI). Several algorithms are available in RSGUI, each of which is based in Rough Set theory. My objective is to find short meaningful predictive rules. First, we need to find a minimum set of attributes that fully characterize the data. Some of the rules generated from this minimum set will be obvious, and therefore uninteresting. Others will be surprising, and therefore interesting. Usual measures of strength of a rule, such as length of the rule, certainty and coverage were considered. In addition, a measure of interestingness of the rules has been developed based on questionnaires administered to human subjects. There were bugs in the RSGUI java codes and one algorithm in particular, Inductive Learning Algorithm (ILA) missed some cases that were subsequently resolved in ILA2 but not updated in RSGUI. I solved the ILA issue on RSGUI. So now ILA on RSGUI is running well and gives good results for all cases encountered in the hospital administration and student records data.

Keywords: Rough Set, Data Mining, Health Analytics, Uncertainty, Decision Making

Table of Contents

Abstract.....	III
List of Tables	VI
List of Figures	VII
1 Introduction	1
1.1 RSGUI.....	1
1.2 Healthcare Business Analytics.....	2
1.2.1 Health information management.....	2
1.2.2 Developing performance metrics for quality of healthcare delivery.....	2
1.2.3 Patient-centred health: Can analytics help?.....	2
1.2.4 Managed Health Services	2
1.2.5 Training healthcare managers.....	2
1.3 Problems addressed by the Ministry of Health.....	3
1.4 Motivation for thesis work	3
1.5 Purpose of study.....	3
1.6 Organization of the thesis	3
2 Literature Review on Rough Sets theory	5
2.1 Rough Sets Theory	5
2.1.1 Discriminant index.....	9
2.1.2 RS1 algorithm	10
2.1.3 ILA Algorithm.....	11
2.1.4 RSRPA Algorithm.....	14
2.2 RSGUI.....	14
2.3 Certainty and coverage.....	16
2.4 Previous Applications of Rough Set Theory.....	18
2.5 Methodologies of designing experiments.....	27
3 Objectives and Methodology	33
3.1 Nature of the data.....	33
3.2 Process of the forms in the hospital	35
3.3 Rough Sets Advantages	36
4 Problem Solution and Implementation	40
4.1 Web Interface for obtaining data.....	40
4.2 Discretization.....	42
4.3 Experiments to determine interesting rules.....	43
4.3.1 ILA experiments	47
4.3.2 ILA java code	48
4.4 Interestingness measures.....	49
4.5 RSRPA algorithm	56
5 Second data experiment.....	60
5.1 Student transcripts	60
5.2 NewRSGUI.....	60
5.3 NewRSGUI snippets java code.....	61
5.4 Student data analytics.....	63
5.4.1 Nature of the student records	63
5.5 Prediction results.....	65

6 Evaluation.....	68
6.1 Occupancy decision attribute evaluation rules	68
6.2 Death decision attribute rules evaluation.....	74
7 Conclusion.....	79
7.1 Major contributions and future work	80
Publication	82
References.....	83
Appendix A: Ethics Approvals	88
Appendix B: Experiment tools.....	91
Appendix C: Decision Rules and their evaluation table.....	94
Appendix D: One month data discretized	98
Appendix E: Hospital Forms Descriptions.....	100
Appendix F: Webpage Form Descriptions	108
Appendix G: NewRSGUI java code	114
Appendix H: Web interfaces design.....	119

List of Tables

Table 2.1: Example of data set and its attributes	7
Table 2.2: Example of data set and its attributes for ILA algorithm.....	11
Table 2.3: Sub-table 1 from Table 2.2.....	12
Table 2.4: Sub-table 2 from Table 2.2.....	12
Table 2.5: Modules of ethnographic study and their advantages & disadvantages	28
Table 4.1: Sample of five days of Female Medical Ward (FMW) data.....	42
Table 4.2: Discretized table for Female Medical Ward data for one month	43
Table 4.3: Summary of the FMW data in RSGUI	43
Table 4.4: Occupancy as a decision attribute for busy value table for all certain rules	52
Table 4.5: Occupancy as a decision attribute for normal value table for all certain rules.....	52
Table 4.6: Occupancy as a decision attribute for not Appendix usy value table for all certain rules	52
Table 4.7: Occupancy as a decision attribute for all decision values with certainty < 1	52
Table 4.8: Death as a decision attribute for 2 deaths decision value for all certain rules.....	53
Table 4.9: Death as a decision attribute for 1 death decision value for all certain rules.....	53
Table 4.10: Death decision attribute for no deaths decision value for all certain rules	54
Table 4.11: Death as a decision attribute for all death decision values with certainty < 1	54
Table 4.12: RSRPA for occupancy decision attribute	57
Table 4.13: RSRPA for death decision attribute	57
Table 6.1: Hospital administrator evaluations for occupancy decision attribute.....	70
Table 6.2: Hospital administrator evaluations for death decision attribute.....	76
Table B.1: An empty evaluation table.....	93
Table C.1: Evaluation by the criteria of the coverage, and the ranking of each pattern	95
Table C.2: Evaluation by the criteria of the coverage, and the ranking for each rule above..	96
Table D.1: Female Medical Ward (FMW) data for one month	98
Table D.2: Discretized table for Female Medical Ward data for one month.....	99
Table E.1: Ward acronym expansions.....	101
Table E.2: ADMISSIONS box, on left top of the form of Figure E.2	103
Table E.3: DISCHARGES box, on right top of the form of Figure E.2.....	103
Table E.4: Field of clinics hospital form of Figure E.3	106
Table F.1: New patient page Description	109

List of Figures

Figure 2.1 The lower, upper approximations, and the boundary region in rough set	6
Figure 2.2: RS1 flowchart.....	10
Figure 2.3: Female Medical Ward data in text file	14
Figure 2.4: Decision table in RSGUI software	15
Figure 2.5: RS1 rules in RSGUI	15
Figure 2.6: RS1 rules and algorithm trace	16
Figure 3.1: Daily Floor Census.....	33
Figure 3.2: Daily Flour Census.....	34
Figure 3.3: Daily Flour Census.....	35
Figure 3.4: Flow diagram of the procedure of the forms in the hospital	36
Figure 4.1: ER Diagram for a hospital website.....	41
Figure 4.2: Interestingness measures in play for rule generation	50
Figure 4.3: RS1 prediction rules	55
Figure 4.4: RSRPA result.....	57
Figure 5.1: Flowchart of NewRSGUI design	61
Figure 5.2: NewRSGUI user interface	61
Figure 5.3: Select table and number of decision attributes interface	63
Figure 5.4: Sample of a first year student transcript.....	64
Figure 5.5: MySQL query for student records	64
Figure 5.6: Student data on NewRSGUI	65
Figure 5.7: RS1 prediction of MATH 2056	66
Figure 5.8: RS1 prediction of COSC 2006	66
Figure 5.9: RS1 prediction for COSC 3407	67
Figure 5.10: RS1 second prediction for COSC 3407	67
Figure 6.1: Degree of departure from the hospital evaluation	68
Figure 6.2: The difference between evaluations for positive and negative sides	71
Figure 6.3: The difference between subject evaluation and hospital evaluation	71
Figure 6.4: The difference of death decision attribute between evaluations for positive and negative sides.....	77
Figure E.1: Daily Flour Census form.....	100
Figure E.2: Daily Ward Census Form	102
Figure E.3: Clinics Form is for one day of statistics and the day was Saturday	105
Figure F.1: Webpage Used by Ward Receptionist.....	108
Figure F.2: Web interface for gathering New Patient Information	109
Figure F.3: Results of adding a new patient to the database	110
Figure F.4: Interestingness measures in play for rule generation	110
Figure F.5: Web page for search of patient information in the database	111
Figure F.6: Webpage for review and update of patient information.....	111

Figure F. 7: New admitted patient diagnosis	112
Figure F.8: Admitted patient's status and diagnosis.....	112
Figure F.9: Knowledge acquisition by means of webpage/MySQL prototype.....	113
Figure F.10: Daily ward report.....	113

Chapter 1

1 Introduction

In fact, the human being does not have enough ability to deal with a large amount of numbers or data and get a perfect result of it. However, he or she has ingenuity in making a good decision. Actually, unlike humans, electronic devices can calculate complicated operations in seconds. However, electronic devices cannot do what the human being does, without being instructed on how to arithmetically transform data to knowledge. With daily growing of the size of medical information in the database, the need becomes greater to use an effective technique to deal with and process those data for decision making, because the ability to use and find these data is a difficult challenge. Indeed, medical data are playing a very important role these days, especially with regard to extraction of clinical knowledge.

While decision making systems have been employed in many sectors in business [1] [2], it has only been in the last few years that they are being used in healthcare management, to a significant extent.

A variety of rough set based tools have been previously developed by a Computer Science student at Laurentian University, Canada and applied to a variety of data sets. The tools also show promise for discovering patterns in medical ward data.

1.1 RSGUI

Rough Set Graphic User Interface (RSGUI) [3] is a software system providing different algorithms appropriate for decision making [4]. The different algorithms that RSGUI features are RS1 that were originally developed by Wong and Ziarko and its origins are described in [5] and improved upon in the following years (LEM, LEM2 [6]), an Inductive Learning Algorithm (ILA) [7], and Rough Set Reverse Prediction Algorithm (RSRPA) [8]. The system has been developed by Laurentian University Computer Science students, beginning in 2005, and continues to evolve as students make enhancements. RSGUI serves as the framework for this thesis.

1.2 Healthcare Business Analytics

A great deal of analytic approaches has been applied to help doctors care for their patients. Less so is work in the area of applying analytic approaches to the healthcare system itself, the motivation for which is discussed in this subsection.

1.2.1 Health information management

Information technology has brought challenges to health information management due to the huge amount of medical data to be efficiently stored, retrieved, and distributed. Security threats are increasing these days. Therefore, it is necessary to address these issues, in order to better protect patients and medical data.

1.2.2 Developing performance metrics for quality of healthcare delivery

Healthcare business analytics are seeking to reduce medical errors and to improve clinical quality of care and patient safety.

1.2.3 Patient-centred health: Can analytics help?

Healthcare analytics are trying to provide more patient-centered service as well as evidence-based practice and data analytics could help with that.

1.2.4 Managed Health Services

Healthcare business analytics could manage the health services by controlling health care costs and managing quality of care. Actually, healthcare business analytics could play an important and significant role in managing health services and improving them.

1.2.5 Training healthcare managers

Healthcare managers' training needs managerial strategy. A study in [9] found that there is a disconnect between what healthcare managers should know and what they actually know about the tasks of strategic management. More resources need to be devoted to strategic management training and the development of managers, at all levels of healthcare organizations. Therefore, training for healthcare managers needs to get more knowledge about tasks of healthcare strategic management.

1.3 Problems addressed by the Ministry of Health

Some of Ministry of Health responsibilities are to support the health service delivery system, preventative health, public health planning, assistance to the health officer to help solve the area hospital issues, performance management of administrators, staff, nurses and doctors, and health information systems and e-health.

1.4 Motivation for thesis work

It is an open research area on how to apply knowledge based and analytic techniques to healthcare information, for the purpose of managing the health care system. Research findings show that there is a wide scope for improving outdated reporting methods, data analysis aimed specifically at the healthcare environment and delivery of results to health care professionals in a user-friendly format. Regarding mining in such data sets, a recent article [10], describes in general terms the methods to detect fraud in health insurance claims. The results indicate that claim anomalies can be detected to help insurance companies recover hidden cost, which are not detectable using transaction processing systems. This is the only relevant article that meets the search criteria “Health Business Analytics” in Scholars Portal Journals in the Laurentian digital library. The scarcity of literature about Health Information Technology (HIT) illustrates the importance of introduction of business analytics in general and knowledge discovery, in particular, into the healthcare industry.

1.5 Purpose of study

The aim of this master’s thesis is to build a Web based medical ward data acquisition, management, and analysis prototype to show that knowledge discovered from data can aid the Ministry of Health in making informed decisions regarding hospital ward management. A secondary data set, student records from the registrar’s office, is considered to help reveal patterns useful for faculty and students.

1.6 Organization of the thesis

The remainder of the thesis is structured as follows. Chapter 2 presents the related work on using rough sets for improving healthcare and also reviews the framework in which our medical ward

decision making system will be developed. Chapter 3 gives an overview of the method of implementation which consists of hospital forms descriptions, processing of the forms in the hospital, rough sets tools in RSGUI, and the methodology of the thesis work to be done. Chapter 4 gives details of an experiment conducted to choose optimal decision rules among many that were generated as possibilities. In Chapter 5, the student records data set was analyzed by the same methods and another experiment was conducted to evaluate the rules generated. In Chapter 6, a questionnaire was developed for users to compare the decision rules according to a variety of quality measures. The results are also discussed in the same chapter. In Chapter 7, the conclusions and future research work are presented. In the Appendices, we provided the two ethics approvals, King Khaled General Hospital agreement to use the medical data and Laurentian University Registrar's agreement to use the students' data. As well the appendices provide the code that was used for developing a web interface, the questionnaire that was used for the experiments with human subjects, ILA code modifications to bring it into line with ILA2 and many more details regarding the forms that were used to develop the health informatics datasets.

Chapter 2

2 Literature Review on Rough Sets theory

Rough sets theory was developed in the early 1980's by Zdzislaw Pawlak [5]. It is a good technique to manage imperfect or uncertain data, and acquire knowledge from it. Rough sets make precise the processes to deal with roughly or approximately described concepts that are difficult to describe by existing information systems. While there are a variety of existing methods for learning from imprecise data, the rough sets method has an advantage for decision making in large volumes of data since it focuses on reducing the number of attributes required to characterize a concept without losing essential information required for decision making.

2.1 Rough Sets Theory

Rough sets theory is a mathematical tool for the processing of fuzzy and uncertain knowledge. It can analyze the incomplete data without any prior knowledge and reveal the internal laws. A doctoral thesis [11] used rough set-based reasoning (term-based) and pattern mining approach as a unified framework for information filtering. Rough set decision theory was used in user profiles and in theoretical modeling of the threshold settings (in the first topic-filtering stage). There are three regions in rough set-based information filtering model, which are the upper approximation, the boundary and the lower approximation. Positive documents are presented in the lower approximation region and are treated as objects. Then r-patterns are used to determine the documents relevance. The “support” based on the r-patterns of documents correspond to their decision value in the rough set theory. When the decision values for all the objects are obtained, rough set elements are sorted as per their decision values. All the unlikely relevant documents are filtered out in this stage. Further, pattern mining was used to find the information mismatches and to improve the method of precision. Computational cost in this methodology is less as compared to other information filtering systems. The use of pattern discoveries in this framework yields accurate results and overall performance of the system is significantly increased.

The rough sets theory is concerned with uncertainty and imprecise data. It can be used to extract meaningful rules from datasets. It attempts to approximate a given concept based on a minimum amount of decision rules. The rough sets theory include very important concepts such as lower approximation and upper approximation [12]. Lower approximation consists of all the elements that surely belong to the concept described in the dataset. On the other hand, upper approximation consists of all elements that possibly belong to the concept. Therefore, the difference between both approximations (upper - lower) is called the boundary region. This region contains cases on the boundary that cannot be precisely classified. Let $[x]_p$ denote, the rows of a table that cannot be distinguished one from the other by means of their properties p . Thus, the lower approximation to concept X is defined as follows:

$$\underline{R}X = \{x \mid [x]_p \subseteq X\}$$

The lower approximation is also called the positive region and it is the union of all equivalence classes $[x]_p$, where the value associated with properties p vary across equivalent classes. The upper approximation is defined as follows:

$$\overline{R}X = \{x \mid [x]_p \cap X \neq \emptyset\}$$

The upper approximation is also called the negative region. The boundary region is defined as follows:

$$\text{Boundary}(X) = \overline{R}X - \underline{R}X$$

We can say the set is rough if its boundary set is nonempty. Otherwise the set is exact [12].

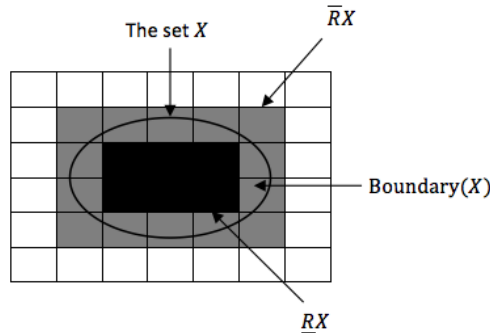


Figure 2.1 The lower, upper approximations, and the boundary region in rough set

The following example will be used to clarify the main procedures of the rough set theory. For example: consider the information table constructed below. The SSN, last_name, first_name,

birth_date, gender, and address are treated as the condition attributes. The decision attribute “duplicated example?” has possible values "y" (yes) and "n" (no).

Table 2.1: Example of data set and its attributes

Example	Condition						Decision
	SSN	LastName	FirstName	BirthDate	Gender	Address	Duplicate example?
e1	d	s	s	s	s	d	y
e2	d	d	s	d	d	d	n
e3	s	s	d	s	d	s	n
e4	s	s	d	s	d	s	y
e5	s	s	d	s	d	s	n
e6	s	s	d	s	s	d	y
e7	s	s	d	d	s	d	y
e8	s	s	d	d	s	d	n

If we define the concept that we are approximating $X_{\text{sample}} = \{e1, e4, e6, e7\}$, the duplicate examples, then \underline{X} and \overline{X} are:

$$\underline{X} = \{e1, e2, e6\}$$

$$\overline{X} = \{e1, e3, e4, e5, e6, e7, e8\}$$

The lower approximation to the concept X is $\{x \mid [x]_p \subseteq X\}$, where p is the collection of all attributes (properties) in the table for the upper approximation $\{[x]_p \cap X \neq \emptyset\}$.

Indiscernibility Classes

The indiscernibility classes for the information table with respect to condition attributes SSN, last_name, first_name, birth_date, gender, and address are:

{e1}

{e2}

{e3, e4, e5}

{e6}

{e7, e8}

Hence, a question such as: “What are the indiscernibility classes for the information table with respect to that same collection of attributes, with the attribute *address* deleted from the collection?” those indiscernibility classes would be:

{e1}

{e2}

{e3, e4, e5}

{e6}

{e7, e8}

It is demonstrated here that the indiscernibility classes without the *address* attribute, are same as the indiscernibility classes with the *address* attribute, which makes it soft to assume that the attribute *address* is redundant. This is the key idea of rough set theory. Thus, we have the possibility to reduce the number of attributes under consideration to a minimal but not necessarily unique set that expresses the same information as the original set. “Expresses the same information” means that for the purpose of defining a concept given by the decision attribute, the minimal set of condition attributes is sufficient. The concept cannot be approximated any better by considering the redundant attributes, plus those in the reduced set. The attributes in the minimal set are called “reducts”.

Certain rules

The list of certain rules that can be generated from example e1, e2, and e6 (the examples in the lower approximation) are as follows:

$$e_1 = (d, s, s, s, s) \rightarrow y$$

$$e_2 = (d, d, s, d, d) \rightarrow n$$

$$e_6 = (s, s, d, s, s) \rightarrow y$$

There are four uncertain (or possible) rules that can be generated from the table 2.1 as follows:

$$e_4 = (s, s, d, s, d) \rightarrow y$$

$$\{e_3, e_5\} = (s, s, d, s, d) \rightarrow n$$

$$e_7 = (s, s, d, d, s) \rightarrow y$$

$$e_8 = (s, s, d, d, s) \rightarrow n$$

The uncertain rules are defined from examples in the boundary region ($\overline{X} - \underline{X}$). Note that there is no reference to the redundant attribute *address* in any of the rules. For ease of presentation, all five attributes have been implicitly shown in the antecedent of the rules. However, not all of the attributes shown in the antecedents are needed. There is a difference between number of attributes in the reduct set and number of attributes on the left hand side of a predictive rule. Introduction of more concepts is required in order to understand that difference.

2.1.1 Discriminant index

In order to measure the degree of certainty, in determining whether or not elements in the universe are members of X , the notion of a discriminant index of X has to be introduced.

It is defined as follows:

$$\alpha(X) = 1 - |\bar{X} - \underline{X}| / |U|$$

If $\alpha(X) = 1$, which means $\bar{X} = \underline{X}$, then the concept X is precise. If $\alpha(X) = 0$, which means $\bar{X} = U$ and $\underline{X} = \emptyset$, then the concept X is completely uncertain [13].

2.1.2 RS1 algorithm

The RS1 algorithm originally introduced by Wong and Ziarko was used in [14] and is presented in Figure 2.2 taken from [3].

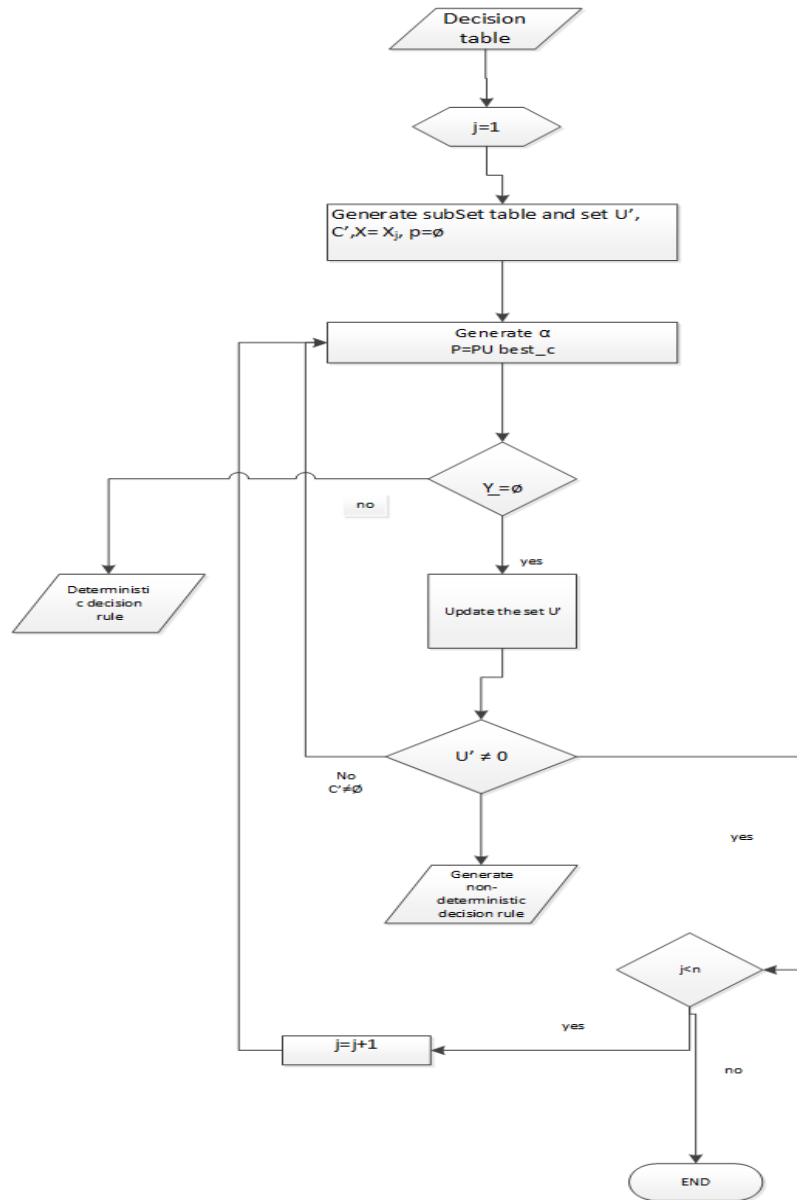


Figure 2.2: RS1 flowchart

The input for the RS1 algorithm is the information table (decision table) and the output is the collection of decision rules for each concept X defined by the decision attributes.

Let U be the universe of objects.

Let C be the set of condition attributes.

Let $X = \{X_1, X_2, \dots, X_n\}$: set of objects based on their decision attribute values.

Let J be set initially 1. J denotes the attribute combination count.

Set U as U' , and C as C' . X_j is the concept under consideration. P is the pivot set initially to 0.

The α being referred to in the flowchart is discriminant index for the subset of condition attributes j .

Compute the discriminate index $\alpha(X) = 1 - \frac{|\bar{X} - \underline{X}|}{|U|}$.

The attribute C with the highest value of α will be the best attribute. If $\alpha > \max$, then $\max = \alpha$, and update P the best $C = \{C\}$, then set P to be $P \cup \text{best } C$. Do a comparison with the discriminant indices to get the highest α number instead of calculating α for each remaining condition attribute in C .

2.1.3 ILA Algorithm

Consider the information table constructed as below. The SSN, last_name, first_name, birth_date, gender, address are treated as the condition attributes. The decision attribute “duplicated?” has possible values “y” (yes) and “n” (no).

Table 2.2: Example of data set and its attributes for ILA algorithm

Example	Condition						Decision
	SSN	Last Name	First Name	Birth Date	Gender	Address	Duplicate?
e1	d	s	s	s	s	d	y
e2	d	d	s	d	d	d	n
e3	s	s	d	s	d	s	n
e4	s	s	d	s	d	s	y
e5	s	s	d	s	d	s	n
e6	s	s	d	s	s	d	y
e7	s	s	d	d	s	d	y
e8	s	s	d	d	s	d	n

1) 2 sub-tables, Table 2.3 and Table 2.4, would be generated from the given above table after the first step of the ILA algorithm as follows:

Table 2.3: Sub-table 1 from Table 2.2

Example	Condition						Decision
	SSN	Last Name	First Name	Birth Date	Gender	Address	Duplicate?
e2	d	d	s	d	d	d	n
e3	s	s	d	s	d	s	n
e5	s	s	d	s	d	s	n
e8	s	s	d	d	s	d	n

Table 2.4: Sub-table 2 from Table 2.2

Example	Condition						Decision
	SSN	Last Name	First Name	Birth Date	Gender	Address	Duplicate?
e1	d	s	s	s	s	d	y
e4	s	s	d	s	d	s	y
e6	s	s	d	s	s	d	y
e7	s	s	d	d	s	d	y

2) If we process the Table 2.2 for concept (duplicate := n) first, then the first production rule extracted by the ILA algorithm is:

(Last Name := d) -----> (Duplicate := n)

3) Again, if we process the Table 2.3 for concept (duplicate := n) first, then the remaining rules extracted by the ILA algorithm are:

Process (Table 2.3)

J=1

{Last Name}= d (1)

The only attribute value that does not appear in the Table 2.4 is the Max_combination = 1 which is {Last Name}= d

and mark example e2

Generate rule#1 from (1) is (Last Name := d) -----> (Duplicate := n)

J=2 ...J=6

Max_Combination = \emptyset

The unmarked rows have attribute values that appear in the Table 2.4 under the same attribute combinations

Process (Table 2.4)

J=1 ---> Max_Combination = \emptyset

J=2

(SSN=d, Last Name=s) (1)

(SSN=d, Birth Date=s) (1)

(SSN=d, Gender =s) (1)

and so on.

pick the first one

mark example e1

Generate rule #2

(Last Name := s) AND (SSN := d) -----> (Duplicated := y)

J=2 ... J=5

(SSN=s, Last Name=s, First Name=d, Birth Date=s, Gender=s) (1)

(SSN=s, Last Name=s, First Name=d, Birth Date=s, Address=d) (1)

and so on.

pick the first one, and then mark example e6

Generate rule #3

(SSN := s) AND (Last Name := s) AND (First Name := d) AND (Birth Date :=s) AND (Gender := s) -----> (Duplicated := y)

Same applies for the Table 2.4

The unmarked rows have attribute values that appear in the Table 2.3 under the same attribute combinations so I marked (skipped) all the rows that apply to this case and end the algorithm

So the rules extracted by the ILA algorithm are:

(Last Name := d) -----> (Duplicate := n)

(Last Name := s) AND (SSN := d) -----> (Duplicated := y)

(SSN := s) AND (Last Name := s) AND (First Name := d) AND (Birth Date :=s) AND (Gender := s) -----> (Duplicated := y)

2.1.4 RSRPA Algorithm

The idea behind the RSRPA is that the decision attributes are given and that the algorithm predicts the condition attributes for those decision attributes. Ordinarily, rough sets theory such as RS1 and ILA are executed as:

Condition attribute1 [given], Condition attribute2 [given],..., Condition attribute N [given] \rightarrow Decision attribute [predict]

But RSRPA prediction is as:

Decision attribute [given] \leftarrow Condition attribute1 [predict], Condition attribute2 [predict],..., Condition attribute N [predict]

Notice that the same condition attributes appear as antecedents in both above cases. However, in the first rule, condition attributes are given whereas in the second rule, condition attributes are to be predicted. This difference is not a case of interchanging the roles of attributes as condition or decision. Consider a hockey game [8]. We know the outcome we want: we want our team to win. RSRPA will tell us the characteristics our team should have in order to win.

2.2 RSGUI

Rough Set Graphic User Interface (RSGUI) [3] is illustrated here. RSGUI runs from command-line (Terminal in OS X) and reads dataset from a text file (.txt) because RSGUI cannot read dataset from other database files such as MySQL. So the dataset should be written in text file and split its values by “~”. See Figure 2.3.

```

remain~admitted~death~discharged~occupancy
LARGE~LARGE~NONE~NORMAL~BUSY
LARGE~NORMAL~1~LARGE~BUSY
LARGE~LOW~2~NORMAL~NORMAL
NORMAL~LARGE~NONE~NORMAL~NORMAL
NORMAL~LARGE~NONE~LOW~BUSY
LARGE~NORMAL~NONE~NORMAL~BUSY
LARGE~LOW~NONE~NORMAL~NORMAL
NORMAL~NORMAL~NONE~LARGE~NORMAL
NORMAL~NORMAL~NONE~NORMAL~NOT_BUSY
LOW~LARGE~NONE~LOW~NORMAL
NORMAL~NORMAL~NONE~NORMAL~NORMAL
NORMAL~NORMAL~NONE~NORMAL~NORMAL
NORMAL~NORMAL~NONE~LOW~NORMAL
NORMAL~LOW~NONE~NORMAL~NOT_BUSY
LOW~NORMAL~NONE~LOW~NORMAL

```

Figure 2.3: Female Medical Ward data in text file

Row	SSN	lastName	FirstName	BirthDate	Gender	Adress	Duplicate
0	d	s	s	s	s	d	y
1	d	d	s	d	d	d	n
2	s	s	d	s	d	s	n
3	s	s	d	s	d	s	y
4	s	s	d	s	d	s	n
5	s	s	d	s	s	d	y
6	s	s	d	d	s	d	y
7	s	s	d	d	s	d	n

What would you like to set the width of the columns to? 10

Figure 2.4: Decision table in RSGUI software

RSGUI has six tabs; shown at the top of Figure 2.4, Table, Change, RS1, RSRPA, ILA, and History. The tab labeled Table, outputs the table under consideration. The task of the Change tab can be to “edit values” or “remove some columns or rows”. RS1, RSRPA, and ILA tabs do their tasks, which have been explained previously. The last tab, which is the History tab, can be used to show all operations that the user has done during his/her session.

RS1 OPTIONS

Execute Executes the RS1 algorithm on the curr

Trace Shows the steps of the RS1 algorithm.

Rules Shows the rules of the executed RS1 algo

Prediction Predicts decision attribute value(s) ba RS1 rules and entered condition attrit

STATUS

```
< (lastName := d) -----> (Duplicate := n)[certainty = 1.0][coverage = 1/8] >
< (lastName := s) AND (SSN := s) AND (FirstName := d) AND (BirthDate := s) AND (Gender := d) AND (Adre
< (lastName := s) AND (SSN := s) AND (FirstName := d) AND (BirthDate := d) AND (Gender := s) AND (Adre
< (lastName := s) AND (SSN := d) -----> (Duplicate := y)[certainty = 1.0][coverage = 1/8] >
< (lastName := s) AND (SSN := s) AND (FirstName := d) AND (BirthDate := s) AND (Gender := s) ----->
< (lastName := s) AND (SSN := s) AND (FirstName := d) AND (BirthDate := s) AND (Gender := d) AND (Adre
< (lastName := s) AND (SSN := s) AND (FirstName := d) AND (BirthDate := d) AND (Gender := s) AND (Adre
```

Figure 2.5: RS1 rules in RSGUI

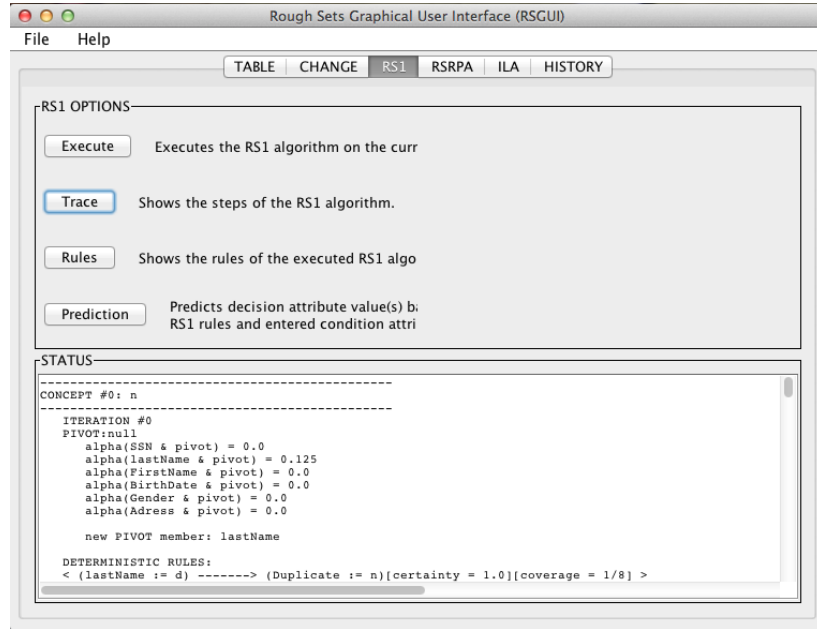


Figure 2.6: RS1 rules and algorithm trace

2.3 Certainty and coverage

Many authors [15][16][17] like to say belief B instead of certainty, and plausibility P instead of coverage. Regarding the above table, belief B and plausibility P of each of the uncertain rules are:

$$B = \frac{2}{8} \times 100 = 25\%$$

$$P = \frac{7}{8} \times 100 = 87.5\%$$

If B and P are almost equal, then our belief is high that the uncertain answer is fairly precise. The certainty of a rule can tell how strong that rule is [18]. The coverage [18] is total number covered by the rule out of all the rows in the dataset. It is computed as follows:

$$\text{Coverage} = \frac{\text{\# of rows covered by the rule}}{\text{\# of rows in the decision table}}$$

The certainty is computed as follows:

$$\text{Certainty} = \frac{\text{\# of rows covered by the rule in the concept}}{\text{\# of rows covered by the rule}}$$

The strong and deterministic rule is that the certainty equals to 1, while the nondeterministic rule is that the certainty ranges between $0.0 < \text{certainty} < 1.0$ [3].

RSGUI (Practice)

The predictive rules generated by RS1 using data in Table 2.1 are shown as follows:

1. (lastName := d) ---> (Duplicate := n)[certainty = 1.0][coverage = 1/8]
2. (lastName := s) AND (SSN := s) AND (FirstName := d) AND (BirthDate := s) AND (Gender := d) ---> (Duplicate := n)[certainty = 0.6666666666666666][coverage = 3/8]
3. (lastName := s) AND (SSN := s) AND (FirstName := d) AND (BirthDate := d) AND (Gender := s) ---> (Duplicate := n)[certainty = 0.5][coverage = 2/8]
4. (lastName := s) AND (SSN := d) ---> (Duplicate := y)[certainty = 1.0][coverage = 1/8] >
5. (lastName := s) AND (SSN := s) AND (FirstName := d) AND (BirthDate := s) AND (Gender := s) ---> (Duplicate := y)[certainty = 1.0][coverage = 1/8] >
6. (lastName := s) AND (SSN := s) AND (FirstName := d) AND (BirthDate := s) AND (Gender := d) ---> (Duplicate := y)[certainty = 0.3333333333333333][coverage = 3/8] >
7. (lastName := s) AND (SSN := s) AND (FirstName := d) AND (BirthDate := d) AND (Gender := s) ---> (Duplicate := y)[certainty = 0.5][coverage = 2/8]

The certain rules are:

1. (lastName := d) ---> (Duplicate := n)[certainty = 1.0][coverage = 1/8]
2. (lastName := s) AND (SSN := d) ---> (Duplicate := y)[certainty = 1.0][coverage = 1/8]
3. (lastName := s) AND (SSN := s) AND (FirstName := d) AND (BirthDate := s) AND (Gender := s) ---> (Duplicate := y)[certainty = 1.0][coverage = 1/8]

Example of rule of certainty 1 from the above decision rules is as follows:

$$r_1 = (\text{lastName} := d) \text{ ---> } (\text{Duplicate} := n)[\text{certainty} = 1.0][\text{coverage} = 1/8]$$

$$r_1 \text{ certainty} = \frac{\# \text{ of rows covered by the rule in the concept}}{\# \text{ of rows covered by the rule}}$$

$$= \frac{1}{1} = 1$$

Therefore, certainty of r_1 is 1, which is what we obtained from RS1 in RSGUI. Another example rule of certainty in the range between $0.0 < \text{certainty} < 1.0$, follows:

$$r_2 = (\text{lastName} := s) \text{ AND } (\text{SSN} := s) \text{ AND } (\text{FirstName} := d) \text{ AND } (\text{BirthDate} := s) \text{ AND } (\text{Gender} := d) \text{ ---> } (\text{Duplicate} := n)[\text{certainty} = 0.6666666666666666][\text{coverage} = 3/8]$$

$$r_2 \text{ certainty} = \frac{\# \text{ of rows covered by the rule in the concept}}{\# \text{ of rows covered by the rule}}$$

$$= \frac{2}{3} = 0.6666666666666666$$

The number of the rows covered by the rule in the concept, was counted from the data rows set not from the rules of the result.

Notice that the rules have been shortened meaning that not all of the attributes from the reduct set appear in the rule's left hand side. Certain rule number 1, for example, has only one antecedent condition. The certainty of each of the above rules is 100% and each rule covers 1 of 8 examples.

Table 2.1 is said to be the training set. It contains 8 examples from which rules are induced. Knowledge given by the induction rules has been discovered from the information table. The concepts being characterized are duplicated example? (Y/N). Duplicate examples are exactly like other example in the table, except that they have been removed and only the knowledge that they have had duplicates remains. This is not a process of the rough set theory but rather a nuance of this particular example.

2.4 Previous Applications of Rough Set Theory

The research in [19] discusses the development of a composite method of Wavelet packet transform (WPT) and rough sets theory (RST) for fault diagnosis of gearbox. Gearbox is a complicated rotary mechanical apparatus in which the fault vibration signal is non-linear and non-stationary. WPT is developed on the basis of wavelet transform and decomposes into low- and high frequencies simultaneously when dealing with the vibration signal. Because of this characteristic, it is possible to process non-stationary signals using WPT. However, for real time realization and online fault diagnosis, the characteristic (feature) vector should include as little elements as possible. Hence, RST is used for attribute discretization and reduction of features. Decision rules for fault diagnosis can be drawn on the basis of this WPT and RST composite algorithm.

The method was successfully applied for the diagnoses of a gearbox. The improved Naïve Scaler Algorithm in the RST reduces the complexity of discretization that further leads to attributes reduction. Overall, this method can delete the redundant features and improves the efficiency of the process. It recognizes the faults quickly and effectively; hence, it is appropriate for online diagnosis. Research on more efficient reduction algorithms may be required in future.

This research [20] presents a novel approach based on the rough sets theory for the mechanical fault diagnosis of a five-plunger pump. If C represents the condition attributes of a variety of faults in the pump and D is the decision attributes in the information system, the information table can be set to determine whether there is dependence between C and D . If the dependence exists, the decision set can be defined by the category of condition attributes set. In the present study, the condition attributes for the pump fault diagnosis were constructed by studying the vibrational signals of the pump. However, the frequency spectra of the vibration signals contained complicated information in the form of continuous variables. Discretization of these variables was done based on the maximum covariance between the classes. After the execution of the 11 steps of the algorithm, diagnostic rules were devised by calculating and reducing the data in the information table. Various technical states of the pump were efficiently diagnosed by the extracted rules.

In a similar publication [21], rough set-based fuzzy rule acquisition approach and a fault diagnosis scheme of an industrial process are discussed. In this approach, the relevant reductions from a given data table were extracted using Boolean reasoning. Reduction approximations were then extracted by means of parameter tuning. Optimum reducts were found from the large database of attributes while using inconsistency count and gain of mutual information (to reduce the abundant attributes). A heuristic reduct algorithm was proposed and successfully applied to fault diagnosis of ethylene cracking furnace. The proposed fuzzy discretization method is robust to process noise and is capable of early monitoring of the abnormal states. The limitation of such models is that they require extensive understanding of the process and it is very difficult to build a precise mathematical model of complex systems.

A composite framework of rough set with formal concept analysis has been designed for intelligent medical diagnosis of heart diseases [22]. The first step in this intelligent data mining model is ‘problem definition.’ Rough set theory was used in the ‘pre-process’ for data processing and data classification after removal of noise to mine suitable rules. A rule generation algorithm was used to generate all possible reducts by eliminating all dispensable attributes. A total of 91 rules were generated initially which were subsequently minimized to 72 candidate decision rules with the help of domain intelligence. Only 65 rules were finalized after the validation process. Formal concept analysis was used in the ‘post-process’ for better understanding of the rules (derived from RST). Though rough set has several advantages over other methods, it generates a number of rules creating difficulties in decision making [22].

The study by Grzegorz Ilczuk and Alicja Wakulicz-Deja [23] analyzed data that contain information about heart disease from an Electrocardiology clinic in Poland. Narrative medical reports were used in this case study to build a decision system. The method used was (Learning from Examples Model, version 2) LEM2 algorithm [6] to generate decision rules that are brief, and easy to understand. These authors used LEM1 and LEM2 algorithms and found that LEM2 gives better results than LEM1. Both LEM1 and LEM2 are based on LERS (Learning from Examples based on Rough Sets) [6].

A Medical expert’s knowledge was used to build attributes table from narrative medical reports. The authors used the term *shortening ratio* to mean a number that ranges between 0 and 1, where 1 means no shortening and 0 means maximum shortening [23]. Maximum shortening means fewest numbers of reducts. Reduct is a powerful tool in rough set to reduce number of attributes and get the minimal set of necessary features, which preserve the main idea of the data description [24][25]. They find greatest accuracy was achieved with shortening ratio between 0.7 and 0.8. They also examined the coverage ratio, which is a cover parameter that can determine the expected degree of coverage of the dataset [26], and found that 0.95 gave the best rules.

RSGUI does not implement LEM2, but it does measure coverage. It does not measure shortening ratio (which has proved to be useful when analyzing rough set data), but it does provide a library of tools from which the user may choose. The difference between the data that were used by Ilczuk and Wakulicz-Deja and the data that were used in this thesis is that my data were taken from tables that have only one value of each attribute for every patient, while the data used in the previous study, was taken from narrative reports.

The study in [27] supports the idea of making decisions using fewest indicators needed in order to make the final diagnosis. The research resulted in diagnosing Mitochondrial Encephalomyopathies (MEM) disease in children and was mainly based on clinical symptoms [27], but the diagnosis of clinical symptoms is not the final diagnosis of the disease. For that, another test from blood and cerebrospinal fluid was used to measure the levels of appropriate parameters. The aim of the study was to shorten the time to get the final diagnosis. The authors used LEM1 algorithm in the study to generate new rules to diagnose MEM disease and make fewer classification errors.

The rough set theory was used with medical reasoning to represent diagnostic models. Shusaku Tsumoto have introduced rough set framework in [28] to model medical diagnostic rules. The characteristics of medical reasoning and the representation of diagnostic models were discussed in this paper by the use of rough sets theory. Focusing Mechanism is an important concept in medical diagnosis used to select the final diagnosis from among many candidates. The primary ideas are rough set model and variable precision, which relate to upper approximation, and ordinal positive reasoning, which ultimately relate to focusing procedure. The study in the paper suggests that the rough set model is closely related to the medical diagnosis. Rough sets are useful in medical domains because the medical reasoning tends to reflect the concept approximation of the rough sets. Diagnostics rules were represented by the classification rules, which have covered high numbers of cases in the data and have a high accuracy too. The author concludes that because rough set theory can generate medical diagnostic rules, medical areas are

excellent for the use of rough set theory. This conclusion can be broadened to diagnostic applications in general [20][19][21][22].

The paper [25] provides a summary of the use of rough sets focusing on the following applications: (1) medical imaging, (2) discovering patterns in medical data, and (3) making a computer support decision making in the medical area. In the medical imaging area, the important tasks are image segmentation based on rough sets theory that some cases are labelled as the positive region which is clearly inside the set. However, some cases are not labelled inside the set which is called negative region in rough sets theory. The main thing required to discover patterns in medical data is data reduction. Like in [25], they discovered that the medical diagnosis can be seen as a decision-making process and computerized as a rough set process that can be accurate, objective and fast. Nowadays, a diagnostic decision support system is an important part in medical technology.

The study in [29], was used to analyse the data of Breast Cancer patients by using the rough set theory. They studied the data set of 228 breast cancer patients and described it with 16 attributes. They divided the data set into two kinds types of patients; patients who had not experienced cancer recurrence and those who have. The rough sets theory was applied to the first type of patients' data to get the important attributes. From these procedures, some helpful inductions were formulated to make decisions for breast cancer patient treatment.

For example, the rough set theory was used to evaluate the ability attributes, in order to estimate the patients' classification. The attributes were selected according to their significance and the measures of classification's quality. The rough set theory allowed us examining interrelations in the subsets of attributes. In breast cancer data there were many decision rules but there were a limited number of examples for some of the rules. A disadvantage was a weak recognition of patients from class 2, which is typical for unbalanced medical datasets. For this reason, research is ongoing in order to discover better classification strategies.

The study [30] discussed data mining in the area of childhood diabetes mellitus, where more than one hundred patients were analyzed. From the disease data set, relevant attributes and decision rules were identified. The study considered three methods for identifying relevant attributes in this domain. The first was based on reducts, the second on attribute significance and the third on attribute ranking, motivated by the wrapper approach, where the classification accuracy was used.

The rough set approach was used to discover the relevant features from the medical data set. The features discovered were found to be affecting the causes of microalbuminuria in children with diabetes mellitus. The rules discovered were consistent with clinical knowledge about type 1 diabetes. The proposed methods worked well and the author claims that they can be applied in different data sets [30]. This study is important because real life medical problems were used in this study and the data in my thesis is also drawn from a real life medical situation.

The paper presented in [31], describes the specific approach of medical reasoning, as applied by the medical expert's decision process. Rule induction was about extracting useful rules from data based on statistical significance. The author describes the two basic categories of the rule induction method, which support lower and upper approximation in rough sets. The two rules are deterministic and probabilistic rules, where deterministic rules are supported by positive examples and probabilistic are supported by large positive and small negative examples. It was postulated that medical reasoning includes two basic rules, which are positive and negative rules.

This journal article by Wakulicz-Deja and Przybyla-kasperek [32] contains an example of how to apply rough set theory in deciding the necessity of future tests and finally deciding upon a diagnosis to be prescribed by a physician, on progressive encephalopathy in a child patient. It is prerequisite to have a number of invasive tests to reach a final decision. Minimizing invasive testing is necessary to carry out the processes that result in appropriate preliminary classifications. The study was based on 3 stages. The first stage classified children into two groups:

- Children suspected of the progressive encephalopathy.
- Children having other diseases with the progressive encephalopathy.

In the second a stage sample of blood and Cerebrospinal Fluid (CSF) were taken from children suspected of having progressive encephalopathy, who did not suffer from the progressive encephalopathy disease. They then determined the locate level in the samples. In the last stage, the rules to classify eligible patients for invasive testing will be created based on these results. In the third stage, selected patients from second stage will be examined for enzyme level. Thereafter, the result of third stage will be confirmation for previous stage of preliminary diagnosis. They carried out their experiment on 114 patients (60 boys and 54 girls,) aged between 3 months and 15 years. They found increased levels of acids in 91 patients, which indicate a highly encouraging result.

They have used rough sets theory as a tool to diagnose progressive encephalopathy. Machine learning in rough set was used in this study to transfer knowledge from an expert to a knowledge base [33]. They have also used LERS to discover useful rules from data. Results of this paper tend to reduce unnecessary testing. To sum up, [32] presented a process of preliminary classification of child patients with the help of rough set theory.

Overall, medical diagnosis process is considered to be comprised of three processes, which are use of focusing mechanism, integration of additional symptoms, and defining complications from other diseases caused by unexplained symptoms. In the paper [34], Shusaku Tsumoto has attempted to develop a corresponding viewpoint of medical differential diagnosis based on rough sets. The medical differential diagnosis is a process of differentiating conditions, diagnosed from similar symptoms. A classical approach to medical diagnosis signifies that each disease carries a certain set of symptoms, which facilitates a specific degree of coverage and confidence to diagnose that disease. It was shown that utilization of differential diagnosis concept, classification on the basis of accuracy and coverage, focusing mechanism and probabilistic rules, are required to achieve an appropriate medical diagnosis.

There is a specific discussion in the paper [34] about one of the main characteristics of medical

reasoning, named focusing mechanism. It is when exclusive reasoning is used to identify the disease when the patient has no noticeable symptoms. On the other hand, inclusive reasoning is used when a patient reflects symptoms relating to specific disease. Exclusive and inclusive reasoning are modeled from upper and lower approximation as two kinds of rules. Such reasoning enables rough set model to make automated extraction or mining of rules. Thus, rule induction based on rough set model is a powerful tool for automated extraction or mining of rules following a focusing mechanism from dataset.

Research paper [35] investigates the application of rough set theory (RST) in wave height prediction. Decision rules of the rough set theory for the prediction of wave heights were derived by applying RST to Lake Superior in North America. The rough set decision making table was constructed based on the wave height data and the wave height (decision making attribute) could be correlated with wind speed (conditional attribute). The dependency between the attributes helped in the reduction of the set of attributes. Finally, the performance of the proposed RST model in wave height prediction was evaluated using statistical measures (Bias, scatter index, mean relative errors, root mean square error and correlation coefficient). In the current study, RST out-performed other methods of soft computing such as Bayesian networks, artificial neural networks, support vector machines, etc.

Interval-valued fuzzy information systems [36] do not make use of RST-based fuzzy rule extraction. However, in order to widen the application of RST, interval-valued fuzzy rough sets (IVFR) have been developed by scientists. In conventional RST, rules are obtained by attribute reduction whereas in this model rules are extracted based on the two algorithms of positive and negative approximations. This model uses a positive granulation order that defines positive approximation space. Based on this approximation, a rule extraction algorithm has been proposed (called as mine rules based on positive approximation MRBPA). Similarly, converse dynamic granulation order led to converse approximation space and subsequently, algorithm for rule extraction could be devised (called as mine rules based on converse approximation MRBCA). The comparison of the computing time and classification accuracy of MRBPA, MRBCA and fuzzy rule induction algorithm (Algorithm based on attribute reduction) revealed that running time of MRBPA was very less as compared to RIA, and it was slightly less than

RIA in case of MRBCA. Classification accuracies of both MRBPA and MRBCA outperformed RIA.

The conventional rough set algorithms cannot generate classification rules incrementally when new objects are added to large databases of information systems. The research [37] devised an algorithm capable of incremental rule extraction without re-computing the rule sets from the beginning. The algorithm is based on reduct generation and alternative rule extraction algorithms developed by the researchers in [37]. In this method [38], the original rule sets containing whole raw data were deduced from attribute reduction table. If the newly added objects are dominated by the original reducts, the rule sets are updated by partial modification of original reducts. This methodology decreases the computation time and memory space significantly. At the same time, it is capable of excluding the repetitive rules and hence, avoids the redundant rules.

ILA has been used on patients of radiology department in Hacettepe University Hospitals [39]. It has been used to discover the time that the patients spend taking a radiology exam and the demographics information. All in all, ILA has generated rules that tell that some of patients in their age groups or birth places spend more time in a specific radiology exam than the others. So ILA can give good results that can help hospital administrators to save time and better run the hospital system.

The study in [40] presented a new algorithm for the knowledge acquisition in ILA. The new algorithm, which is REX-1 aims to remove or get rid of the disadvantages of the algorithms that are in use in Inductive learning. So comparing REX-1 algorithm with some algorithms such as; ID3, C4.5, and Rules Family, the REX-1 gives general rules with respect of the priority of attributes of the set with fewer numbers of rules and higher knowledge values.

ILA2 has faster features than the basic ILA which reduces the processing time without losing much from the quality of the result. So the ILA2 is an improved version of basic ILA [41] [42]. All in all, ILA2 is better than ILA with respect of the main idea that both algorithms do. However, ILA2 is better in term of the time of processing, the size of the rules, and the accuracy of finding hidden patterns.

2.5 Methodologies of designing experiments

In this section, we review previous experimental studies involving medical data because they have guided the design of our experiments with human subjects who were asked to evaluate predictive rules induced from medical ward data. Advantages and disadvantages of different methods of experimental study are also discussed.

The presented study [43] involved both qualitative (phenomenological design) as well as quantitative (questionnaires) designs. The sample consisted of three admission wards comparable in nature, scope and dimension. The questionnaire had a 16-item rating and was based on the results of qualitative data. Quantitative data were analyzed by the software SPSS 15.0 and non-parametric correlations (Spearman's rho) between the items were also calculated.

The advantages of using questionnaires are that a large amount of information can be collected from a large number of people in a short period of time. This method is cost effective and obtained results can be easily quantified (by manual methods or by using software systems).

The major hurdle in the studies involving questionnaires is the 'disinterested' participants. In the present study only 62% of the questionnaires were returned. The sample size of the population to be analyzed needs to be large enough to arrive upon a general conclusion. The truthfulness of the respondents is also questioned in some studies. For example, in the present study nurses showed initial enthusiasm but later on they were reluctant to carry on with the interventions.

Research paper [44] presented an ethnographic study of nurses. In this type of study, the researcher, temporarily, becomes a part of the research setting. This 'involvement' makes it easy to understand the reality of research subjects. A reflexive ethnographer attempts to 'tell the story' of the subjects to the community. The basic advantage of this method is the 'snap-shot' nature of the conducted studies. These studies have a small sample size; however, they behold the 'essence' of a particular situation.

The major disadvantage of this method is the researcher's own interpretation of the issue. This interpretation may be influenced by time & place and thereby confounds the results.

The modules used in the study included participant observation, semi-structured interviews with the participants, and keeping reflective fieldwork journal. The three modules and their advantages & disadvantages are listed in Table 2.2.

Table 2.5: Modules of ethnographic study and their advantages & disadvantages

Module	Advantages	Disadvantages
Participant observation	This method is cost-effective and does not require infrastructural support.	The observer may not be acceptable by the participant. If acceptable, the observer may not be allowed to take part in the activities. Reciprocal influences may be exerted by the observer and the participant.
Semi-structured interviews	These are open-ended and are sensitive to the answers of the participant as well as the intuition of the observer or researcher. This method is cost-effective and less time-consuming.	The answers of the participants may be influenced by the researcher.
Reflective fieldwork journal	Uses the researcher's own thoughts and reflection. It helps the researcher to differentiate 'significant' information from the 'insignificant'.	Constant validity check of the journal is required

In article [45], the authors used the 'observational method' of study to assess the deficits in communication and information transfer at the hospital discharge. This method involves the identification of data sources to collect the data. These sources are the databases like MEDLINE and Cochrane Database of Systematic Reviews. Data are retrieved from the search engines using appropriate key words. More than one key word is necessary to retrieve maximum information on the subject. Once the data are extracted, they are analyzed for discrepancies and duplication. The final step in the observational studies includes data synthesis. Data synthesis involves categorization, summary and statistical analyses of the obtained data.

Articles from the peer-reviewed scientific journals are generally included in the observational studies. Therefore, the conclusions drawn from the data are accurate and true. The data retrieved

from the scientific sources are amenable to statistical analysis. The observational studies are time-consuming and the data extraction requires skilled researchers. If the topic of the research is wide, the quantitative synthesis (meta analysis) may not be appropriate to conduct. Articles of a particular language can be retrieved from the search engines (generally English); hence, the articles from the other local (national languages) are not included in the study.

The article [46] aims at development and validation of a clinical prediction rule to identify patients at high risk of developing delirium during their hospital stay. The methodology involved a prospective study having two distinct cohorts. One cohort was treated as derivation cohort and the other was validation cohort. In such studies, the predictive variables are noted and analyzed in one group of subjects (belonging to derivation cohort). Subsequently, the devised rule is validated using another group of subjects (from the validation cohort).

There are several advantages to the credit of this method. The collected data are amenable to statistical analyses (continuous variables of the derivation and validation cohorts can be compared by student's t-test, and ordinal and dichotomous variables can be compared using chi square test). The sample size is variable and may be set depending on the research study. Calculations can be made using SPSS and MedCalc software. The major disadvantage of this method is the inherent differences in the samples of two cohorts.

Study [47] involves the analysis of hospital-discharge data for the assessment of epidemiology of vertebral osteomyelitis. The authors used a hospital-discharge database in France. This database is a collection of standardized discharge summaries of hospital stays of the patients. The collected discharge summaries are then categorized into single medical or surgical diagnosis related group (DRG). This categorization is based in the 10th edition of the International Classification of Diseases (ICD-10) and makes it possible to link multiple hospital stays corresponding to one particular patient.

In the same study [47], the cases of vertebral osteomyelitis were identified on the basis of ICD-10 codes. These cases were further segmented into definite cases, probable cases and possible cases.

The advantages of this study design and methodology are as follows:

- 1) The patient databases are created by using unique identification numbers. Hence, the identity of the patient remains concealed.
- 2) DRG system makes it possible to club the patients with similar diagnosis into one category. Therefore, information retrieval is easy.
- 3) Since multiple hospital stays of a single patient are linked to the patient identification number, it is possible to analyze and review the case over a considerable period. It is also possible to estimate the incidences at the national levels.
- 4) The retrieved data are amenable to statistical analysis (SAS was used in this study [47]).

The primary disadvantage associated with this method of data collection is that the information stored in these databases is only for in-patients. So, out-patients cannot be included in the study. This method may not be appropriate to use in case of rare diseases because their diagnosis and coding are complex.

Because of the lack of available neurosurgical surgeon in the developing countries, most patients with head injuries are under the care of general surgeons. For this reason, it is especially important to prevent these injuries and while it is simple to understand the care needed for the patients, it is unfortunately not as simple to deliver this care in reality. This study [48] examines the variety of head injuries within a busy regional hospital in order to measure the quality of the care received by the patients. This study is divided in three sections, which are a prospective audit of all patients with a traumatic brain injury, over a two month period, at the Accident and Emergency (AE) department at Edendale Hospital, Pietermaritzburg. The two other audits are comprised of 25 referral letters from inpatients reviewed at random and 28 AE clerking notes to evaluate the quality of care administered. Their quality of care was evaluated by comparing the referral letters and the notes with agreed standardised markers.

Between October and November of 2007, 150 patients with head injuries (117 males and 33 females) were examined in the AE department, with a head injury warning chart. Of these 150 patients, 76 were discharged from the hospital, while 49 were admitted to the general wards, 11 were sent to the surgical intensive care unit, 10 were referred to the neurosurgical center in Durban, and 4 were pronounced dead in the AE department. Three of the 10 patients who needed

advanced neurosurgical care, required urgent burr-holes before referral. One of these patients died, but the remaining 9 patients were successfully transferred to the neurosurgery unit. The referral letters and AE clerking notes revealed major deficits.

Although traumatic brain injuries are a common reason for hospital visits, many do not require the attention of a neurosurgical surgeon and can bear the delay caused by transfer, though there are cases where there is a need for an urgent intervention. Furthermore, there is a lack of importance associated to secondary brain injuries, most likely due to insufficient comprehension, so it is imperative to impose measures to improve its awareness. Introducing formalised standard referral and management sheets could greatly enhance the quality of care of patients.

The study [49] intends to explore how deaths of patients are handled by ward staff, which include nurses and healthcare support workers. A group composed of 13 participants from two acute medical wards was interviewed regarding the death of their patients. In order to evaluate and analyse the data, the researchers used the Heideggerian phenomenological approach. Responses, influences and support were determined as the main themes but were further divided into subgroups recognized by the social psychology literature. The responses of the participants were often not noticed nor understood by their managers.

In this chapter, we have reviewed Rough Sets Theory including the concepts of Indiscernibility class, Lower and Upper approximations and Discriminant index. Rule induction algorithms RS1, ILA and RSRPA were described and the software system RSGUI into which these algorithms have been embedded was presented. Previous Applications of Rough Set Theory and previous methodologies of designing experiments around medical data were examined.

We have seen that RST has been used for fault diagnosis of gearbox [19], mechanical fault diagnosis of a five-plunger pump [20], fault diagnosis of ethylene cracking furnace [21], medical diagnosis of heart diseases [22], and diagnosing Mitochondrial Encephalomyopathies (MEM) disease [27]. Indeed, a rough set framework [28] to model medical diagnostic rules in general was developed because medical reasoning tends to reflect the concept approximation of rough

sets. Whereas rough sets have previously been used in the medical area for diagnosing medical conditions, we wish to use them now for diagnosing conditions on medical wards.

Chapter 3

3 Objectives and Methodology

Upon examination of data about the activity (admittance, discharge, death) on medical wards a number of questions come to mind. For example, what conditions on the ward tend to increase patient deaths? A variety of rough set based tools have been developed at Laurentian University that show promise for discovering patterns in the medical ward data. This project is an experiment that seeks to find out how to use these tools, in order to bring to the surface hidden relationships in the medical ward data.

3.1 Nature of the data

Data have been obtained from King Khaled General Hospital in Saudi Arabia. Mohammad H. Thany, the manager of Computer and Statistics department at the hospital, agreed to let us use the data of the summer of 2013. The statistical data provided for this research project are unidentifiable. Also the Research Ethics Board at Laurentian University has approved to use the data in this thesis. In the Daily Floor Census we have two types of forms. The first one (Figure 3.1), is the form that is compiled by the ward nurses; it has patient admission and discharge information, such as name, file numbers, diagnosis and the discharge time.

Kingdom Of Saudi Arabia
HEALTH AFFAIRS DIRECTORATE
HAJER ALKHAITHUM
KING KHALED GEN. HOSPITAL

12:00 Midnight
Date: 25/7/13
Day: 25/7/13

ADMISSIONS
(Record total on Line 2 of Summ.)

Hosp. No.	Age	Sex	Admitted	Time	Diagnosis
203875	45	F	Hsh. Hamed. Maged. Al-mahdy	11:20	Dx: H.T.A.

Received by Transfer From Other Ward (TRANS - IN)

DISCHARGES
(Record of total on Line 5 of Summ.)

Hosp. No.	Age	Sex	Discharged	Time	Diagnosis
168328	55	M	Shay. Al-had. Qasim. Al-mahdy	14:00	D.K.A.
198685	29	F	Hsh. Hamed. Rhi. Al-mahdy	14:00	Hemostomosis

Discharged By Transfer to Other Ward (TRANS - OUT)

SUMMARY FOR A DAY

1	Remaining Last Report	11
2	Admitted	1
3	Received By Transfer From Other Ward	—
4	Total (Sum Of Lines 1, 2, 3)	12
5	Discharged	2
6	Discharged By Transfer to other Ward	—
7	Transfer to other Hospitals	—
8	Died	—
9	Total No. Of Discharged And Deaths	2
10	Remaining 12:00 Midnight	10

Breakdown of Remaining patients
Male :
Female :
Pedia :
Total :
MR 022

1 Non - Saudi
Saudi
Avg. L.O.S.
Signed: [Signature]
Ward: FMW

Figure 3.1: Daily Floor Census

Figure 3.1 is a single ward form that has six tables; ADMISSIONS, DISCHARGES, TRANS-IN, TRANS-OUT, DIED, STILLBIRTH, and a SUMMARY for all tables on the form. Nurses in a ward fill the Figure 3.1 form every day, from 12:00 AM until 12:00 AM next day. The form has the date and day written on the top right corner. Dates are written in Arabic calendar style. The ward name is stamped on the top left corner. The Daily Floor Census Form (Figure 3.1) is sent from every ward to the Statistics and Computer Department after midnight, to allow a data entry person to compile them into Daily Floor Census form (Figure 3.2) the same day to be reviewed by the Statistics and Computer Department manager and then to be sent to the hospital director and signed.

INFORMATION CENTER DEPARTMENT
KING KHALED GENERAL HOSPITAL
DATE : 23 / 9 / 1433
DAY: SATURADY

1433

DAILY FLOUR CENSUS

WARD	ADMISSIONS				TRA.IN	DICHARGES				DIS.		TOT. REM. P.		NO.O. BEDS	L.O.S	BED OCC. RATE
	REM. P.		NEW ADM.			TRA.OUT	TRA.TO OH	DEATH	DIS.		TOT. REM. P.					
	S	N S	S	N S					S	N S	S	NS				
M.M.W	17	2	2	0	1	1	0	2	4	0	13	2	33	22	45%	
F.M.W	15	3	1	1	0	0	0	0	0	0	16	4	36	0	55%	
M.S.W	19	1	7	0	0	0	0	1	2	0	23	1	35	27	68%	
F.S.W	18	1	9	0	0	0	0	0	6	0	21	1	36	24	61%	
O.B.S	36	0	27	4	0	0	0	0	34	2	39	2	55	67	74%	
NEURS	31	3	0	0	0	0	0	0	4	0	27	3	35	34	57%	
PED.W	16	0	7	0	1	1	0	0	3	0	20	0	53	24	38%	
P. ICU	2	0	1	0	1	1	0	0	1	0	2	0	4	3	50%	
I.C.U	5	3	0	0		1	0	0	0	0	6	2	9	5	89%	
C.C.U	1	0	0	0	0	0	0	0	0	0	1	0	4	0	25%	
TOTAL	160	13	54	5	4	4	0	3	44	2	168	15	300	206	61%	
	173		59						46		183					

BED OCCUPANCY RATE : %61 ADULT CAPACITY %63 NERS/PEDIA: %57

Figure 3.2: Daily Flour Census

Because of the different meaning in terms, the person who created the form meant, by Flour, accurate or precise, in the top of the form. The second form of Daily Flour Census (Figure 3.2) is the summary for all wards in the hospital and has only numbers of admitted and discharged patients. It does not have names or times of discharges. It has only the numbers of

how many patients in every ward were admitted and discharged for one day. The M.M.W is the Male Medical Ward and The F.M.W is the Female Medical Ward. The remaining expansions can be found in the appendix E.

INFORMATION CENTER DEPARTMENT
KING KHALED GENERAL HOSPITAL
DATE : / 4 / 1434
DAY:

١٤٣٤

DAILY FLOOR CENSUS

WARD	ADMISSIONS					DICHARGES				DIS.	TOT. REM. P.	NO.O. BEDS	L.O.S	BED OCC. RATE
	REM. P.		NEW ADM.		TRAIN	TRA. OUT	TRA. TO OH	DEATH						
	S	N	S	N										
M.M.W	١	٢	١	١	١	١	١	١	١	١	١	٣٣	١	١
F.M.W	١	١	١	١	١	١	١	١	١	١	١	٣٦	١	١
M.S.W	١	١	١	١	١	١	١	١	١	١	١	٣٥	١	١
F.S.W	١	١	١	١	١	١	١	١	١	١	١	٣٦	١	١
O.B.S	١	١	١	١	١	١	١	١	١	١	١	٥٥	١	١
NEURS	١	١	١	١	١	١	١	١	١	١	١	٣٥	١	١
PED. W	١	١	١	١	١	١	١	١	١	١	١	٥٣	١	١
P. ICU	١	١	١	١	١	١	١	١	١	١	١	٤	١	١
I.C.U	١	١	١	١	١	١	١	١	١	١	١	٩	١	١
C.C.U	١	١	١	١	١	١	١	١	١	١	١	٤	١	١
TOTAL	١٠	١١	١٠	١١	١	١	١	١	١	١	١	٣٠٠	١	١

BED OCCUPANCY RATE: % ADULT CAPACITY % NERS/PEDIA: %

٢٤ ٢٤ ٢٤

Figure 3.3: Daily Floor Census

Usually, the Computer and Statistics department at the hospital has the same form as Figure 3.2, filled manually with Arabic numbers, demonstrated in Figure 3.3. Often, the data entry person first compiles numbers manually from Figure 3.1 to ensure and give accurate numbers since they are not fluent in English. Thus, they fill the form using Arabic numbers and then type those Arabic numbers in English on form of Figure 3.2 Dates are written in Arabic as noted on the right top corner.

3.2 Process of the forms in the hospital

The Computer and Statistics Department sends Daily Floor Census forms (Figure 3.1) to all of the wards in the hospital. Every ward must fill out the form using patient information by

midnight, as illustrated at the top right corner of the form (day of week and full date). After midnight, all forms have to be sent back to the Computer and Statistics Department and compiled into the Daily Floor Census form (Figure 3.2) by one of the data entry persons. Once it has been compiled, it should be reviewed and signed by the department manager, then sent to the hospital director for reviewing and signed once more. A diagram showing the flow of the forms through the hospital appears in Figure 3.4.

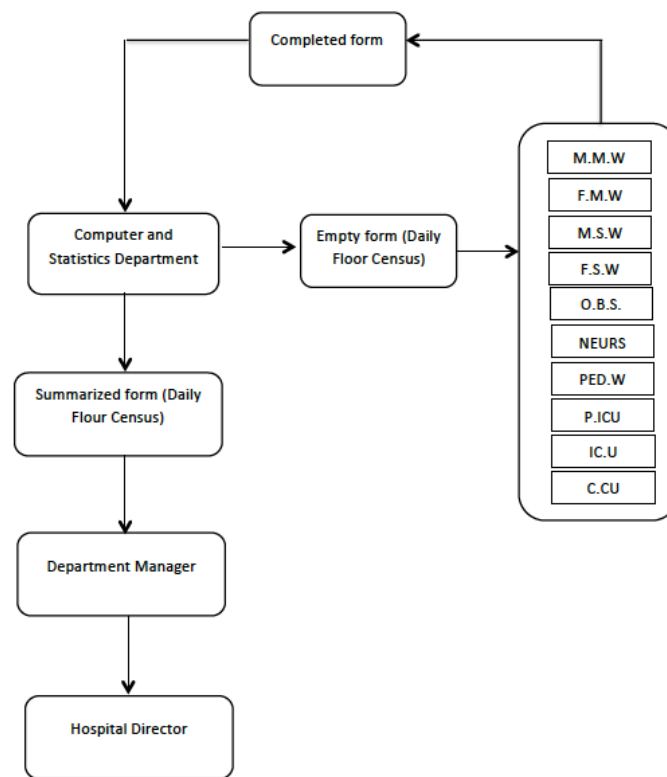


Figure 3.4: Flow diagram of the procedure of the forms in the hospital

3.3 Rough Sets Advantages

Rough set theory has attracted the attention of researchers all over the world and has been applied for a variety of purposes & situations and to solve a diversity of problems. It serves as a mathematical tool for dealing with vagueness, uncertainty & imperfect knowledge and is applicable in many branches of artificial intelligence. A few real-life applications of RST include medical data analysis, finance, banking, voice recognition, image processing, machine learning, oceanography, etc. Several problems in data analysis can be tackled using RST viz.

Characterization of set of objects in terms of attributes, finding total or partial dependencies between attributes, reducing redundant data/attributes and finding significant ones. This theory offers simple algorithms and enables straightforward interpretation of results. Rough set theory helps in finding hidden patterns in data by proving efficient algorithms. It identifies the relationships that are not found using other statistical methods as described in Section 2 pertaining to medical ward data specifically and can be utilized in the processing of both qualitative and quantitative data [50].

Since its inception, rough set theory has been compared and contrasted with other mathematical tools dealing with vagueness and uncertainty. However, the major advantage (and the difference) of RST over other tools is that it does not need any preliminary information about the data. Preliminary or additional information is needed in statistics (in the form of the probability distribution) and in Dempster-Shafer theory (in the form of the basic probability assignment). The major difference between the RST and Dempster-Shafer theory is that the later uses belief function as the main tool but the former uses a family of all sets with upper and lower approximations. Similarly, fuzzy set theory also needs preliminary information in the form of grade of membership or value of the possibility [51][52].

It has been reported that every decision algorithm based on RST reveals well-known probabilistic properties such as probability theorem and the Bayes' theorem. These algorithms satisfy total probability theorem & Bayes' theorem. Hence, the RST algorithms can draw conclusions from large data sets without referring to prior & posterior probabilities. Most of the algorithms based on RST are suited for parallel processing and at the same time, the programs implementing the methods of RST can easily run on parallel computers i.e. concurrent processing is feasible [53].

3.4 Available Rough set based tools

The Rough Set Graphic User Interface (RSGUI) [3] and its algorithms for decision making [4] have been mentioned in Subsection 1.1 and have been described in more details in Subsection 2.1. The decision table that is used in this section contains two types of attributes, which are the condition attributes and the decision attributes. For example, the condition attributes in this case are Remain_Patients, New_Admissions, Death, and Discharged_Patients, whereas the decision

attribute is Occupancy. From the medical ward data we are looking for what can make the ward busy. Therefore, the ward Occupancy is the best decision attribute in the dataset for that purpose.

Rough set data may have thousands of objects, so representing those objects in a table is the best way to have organized and understandable data. Each column represents an attribute, and each row represents attribute values for the columns. RSGUI is based on a user being able to designate any attribute or set of attributes as the decision attribute. Once the RSGUI has selected the dataset and begins execution, it will ask some questions. The first question is, “How many decision attributes are there?” If the user entered the number one for the number of decision attributes, it is going to ask another question according to the number entered to confirm the user’s choice, and will show the last right side attribute in the dataset as a decision attribute. For example, in the dataset showing in Figure 2.3 the “Duplicate” is the last right side column attribute in the data. Thus, the second question if the entry number of the first question was 1, is going to be “Is the decision attribute(s) the following: occupancy?” If it is, the user should type “yes” or “y” to display the RSGUI window Figure 2.3. If it is not, the user could type “no” or “n” to allow the user to enter the column number of the decision attribute after the message “Please enter the number of the decision attributes (starting at 0)”.

Once the information table has been displayed, the user can either change the contents of the information table, or merge columns that are desired to be treated as single columns or select from the available algorithms to generate decision rules. In the next section, the method by which the medical ward data will be analyzed using RSGUI is described.

3.5 Methodology

The medical ward data may have useful implicit rules that are not immediately obvious by viewing the table directly. To get those rules, careful selection of important or interesting decision attributes is needed. In this case, the first decision attribute that will be checked is *occupancy* to see what could make the ward busy, in order to limit the busyness if it is causing problems. For example, emergency doctors sometimes unnecessarily admit patients, when a medicine may be enough to treat their condition. The second decision might be death either to see if the occupancy caused some deaths or to know at least that the deaths cannot be controlled by the staff in the ward.

After testing the data by RSGUI, we are going to evaluate the rules that RSGUI generated. Criteria by which to evaluate the rules generated will be describe in more details in Chapter 4. The most important matter is to find the useful and interesting rules from the data and order them from the most interesting to least interesting. Many questions may come to mind for the hospital director and the manager of Computer and Statistical department. For example, what can make the ward busy? Is ward occupancy a factor contributing to staff negligence? Does staff negligence cause deaths in a ward?

The hospital sends an inpatient statistical report to the Ministry of Health Care daily. The reasons that the Ministry has asked for reports are, to analyze inpatient statistics, to verify if there is enough space, to see what can cause the busyness, and to see which of those previous reasons could have caused deaths in the ward? The Ministry could expand the ward with more rooms, beds, or employ more staff and doctors to at least limit busyness on a ward to remedy the situation. Another reason that may cause deaths or busyness is that the staff or doctors may not be qualified enough to help patients. So training may be a remedial action that will lead to higher quality patient care.

Rough set is data mining algorithm for decision making based on incomplete, inconsistent, imprecise and vague data. Most Healthcare data have these properties. Soft computing techniques of which rough sets are one along with fuzzy sets and others aim to be as precise as possible about imprecision. New knowledge in the form of predictive rules is automatically discovered and a measure of the plausibility of inferences is given. Rough set and soft computing in general are data analytics approaches because implicit, previously unknown and hidden information is brought to the surface. Knowledge discovered from the ward data will be useful for the Ministry of Health Care for policy making, and for all levels of hospital management, for improving patient safety and satisfaction and for improving hospital productivity in general.

Chapter 4

4 Problem Solution and Implementation

Our aim is to employ knowledge discovered in health information to help improve the quality of healthcare delivery and the effectiveness of healthcare managers. An innovative solution has been developed based on an analytic and knowledge management approach to decision making.

Forms that were compiled by ward nurses during the summer of 2013 have been obtained from King Khaled General Hospital in Saudi Arabia concerning events such as ADMISSIONS, DISCHARGES, TRANSFERS-IN, TRANSFERS-OUT, DEATHS, STILLBIRTHS for each of many wards such as F.M.W (Female Medical Ward). These events and their associated information were transformed to electronic tables. The field values were algorithmically discretized to yield soft values, for example, LOW, NORMAL, HIGH for attribute New-Admissions. From the numeric values that appeared in the available forms used to populate the database tables. The available tools permit attributes of discretized information tables to be changed dynamically to play the role of either condition attribute (predictor) or decision attribute (predicted). If we are looking for what can lead to a ward being busy, the Ward- Occupancy is the best decision attribute in the dataset to be used.

For accuracy of information tables and hence the accuracy of the rules induced from them, a web interface was implemented as the most natural solution. Its function are to ease the task of translating Arabic numbers to English, help populate and discretize information tables, provide access to analytic tools for decision making and permit roles of attributes to be distinguished as condition or decision for rule tuning.

4.1 Web Interface for obtaining data

Referring to Figure 3.4 that shows the flow of the forms through the hospital, it is clear that a better method of converting data from manual to electronic form is required. Additionally, for accuracy of the information tables and hence accuracy of the rules induced from them, an

interface to ease the task of translating the Arabic numbers to English would be helpful. A web interface for populating information tables appeared to be the most natural solution. An Entity Relationship diagram was developed to help design the web interface [54]. See Figure 4.1.

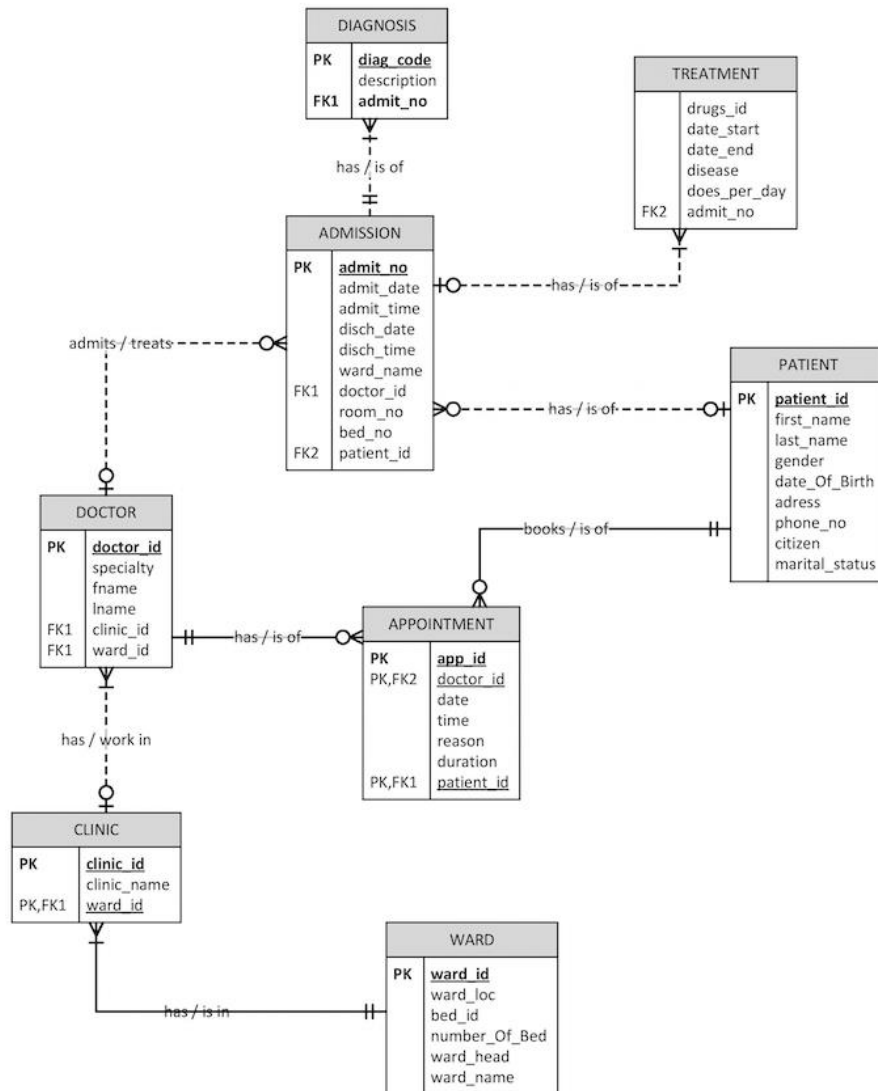


Figure 4.1: ER Diagram for a hospital website

Figure 4.1 is generated from Figure 3.1 to end up with the ward entity. The purpose for this diagram is to create a good database for all wards, as well as good web interfaces for the hospital staff and professionals to access the database. A primitive but exceedingly useful web interface was designed for acquiring data from manual forms. Web page designs with their implemented and proposed prototypes, and more detailed information flows of the hospital can be found in

Appendix F. In the remainder of this section, a data management strategy for predicting outcomes in medical ward data is advanced.

4.2 Discretization

“Discretization is the process of turning continuous values into discrete intervals” [55]. The data used in this thesis need to be discretized because the data comprise numerical and continuous values. There is an abundance of discretization algorithms[56] [55] [57] and any one of these could have been used. Discretization for obtaining discrete values from numerical data is a well understood problem. To accomplish the task, I looked at natural clustering of the data by visual inspection and also used the equal frequency binning discretization method [58]. Table 4.1 is a partial sample of one month ward data before discretization.

Table 4.1: Sample of five days of Female Medical Ward (FMW) data

Day	Remain Patient	New Admission	Death	Discharge	Occupancy
1	18	1	0	7	12
2	12	2	0	1	13
3	13	6	0	3	16
4	16	6	0	4	18
5	18	5	0	6	17

I will explain the discretization method that I am using based upon the first column (Remain Patient column) and the same discretization process is applied to all the other columns and the remainder of Table 4.1 and Table 4.2 will be shown in Appendix D in more details. There are 5 columns in total on the above table that need to be discretized.

Grab all distinct numbers in the first column. Here we have the minimum number as 9 and the maximum as 24. Suppose we want to divide the first column into 3 bins, each bin has 4 elements. 9, 11, 13, 14, | 15, 16, 17, 18, | 19, 20, 21, 24

The boundary values of the bins are as follows:

$$(14+15)/2 = 14.5 \quad , \quad (18+19)/2 = 18.5$$

Therefore, after that the value ranges will be as follows:

Bin 1: [0, 14.5] → LOW

Bin 2: [14.5, 18.5] → NORMAL

Bin 3: [18.5, +∞] → LARGE

Table 4.2 shows the sample of one month ward data from Table 4.1 discretized so that the column entries are now all soft values.

Table 4.2: Discretized table for Female Medical Ward data for one month

Day	Remain Patient	New Admission	Death	Discharge	Occupancy
1	NORMAL	LOW	NONE	LARGE	NOT_BUSY
2	LOW	LOW	NONE	LOW	NOT_BUSY
3	LOW	LARGE	NONE	LOW	NORMAL
4	NORMAL	LARGE	NONE	NORMAL	NORMAL
5	NORMAL	NORMAL	NONE	NORMAL	NORMAL

4.3 Experiments to determine interesting rules

Table 4.3 is a summary of all experiments that have been done in RSGUI using the Female Medical Ward data. RS1 and ILA algorithms have been applied upon the data for one month and also for three months. The Table 4.3 shows the attributes that have been used in the experiments, the amounts of rows and columns considered, and the total of rules obtained from every experiment.

Table 4.3: Summary of the FMW data in RSGUI

Algorithm	Conditions Attributes	Decision Attributes	Rows Considered	Columns Considered	Number of rules
RS1	remain, admitted, death, discharges	occupancy	29	5	13
ILA	remain, admitted, death, discharges	occupancy	29	5	14
RS1	remain, admitted, death, discharges	occupancy	77	5	30
ILA	remain, admitted, death, discharges	occupancy	77	5	24
RS1	remain, admitted, death, discharges	occupancy	89	5	39
RS1	remain, admitted, discharges, occupancy	death	89	5	37

Order of the result after the evaluation from the most interesting, which is the higher number, to least interesting, which is the smallest number, according to the sum of the rule weight on Table 6.1 in chapter 6 would be as follows:

1. (admitted := LOW) AND (remain := LOW) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 5/89]
2. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.9333333333333333][coverage = 15/89]
3. (remain := LARGE) AND (discharged := LOW) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
4. (remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := NONE) ---> (occupancy := BUSY)[certainty = 1.0][coverage = 2/89]

5. (remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := 1) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 1/89]
6. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LOW) -----> (occupancy := NORMAL)[certainty = 0.8333333333333334][coverage = 6/89]
7. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.5][coverage = 6/89]
8. (remain := LARGE) AND (discharged := NORMAL) AND (admitted := LARGE) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
9. (admitted := LOW) AND (remain := NORMAL) AND (discharged := LARGE) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 2/89]
10. (admitted := LOW) AND (remain := NORMAL) AND (discharged := NORMAL) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]
11. (admitted := NORMAL) AND (remain := NORMAL) AND (discharged := LARGE) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]
12. (remain := NORMAL) AND (discharged := LOW) AND (admitted := NORMAL) AND (death := NONE) -----> (occupancy := BUSY)[certainty = 0.2727272727272727][coverage = 11/89]
13. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 5/89]
14. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.7272727272727273][coverage = 11/89]
15. (discharged := LARGE) AND (death := NONE) AND (remain := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/89]
16. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 3/89]
17. (admitted := NORMAL) AND (remain := LOW) AND (discharged := LOW) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.5][coverage = 6/89]

Now ILA in RSGUI runs with more than two decision values and gives a good result by comparison with RS1 as shown following:

1. IF (remain = LARGE) AND (admitted = LARGE) THEN (occupancy = BUSY)
2. IF (remain = LOW) AND (admitted = LARGE) THEN (occupancy = NORMAL)
3. IF (remain = LOW) AND (admitted = LOW) THEN (occupancy = NOT_BUSY)
4. IF (remain = NORMAL) AND (death = 2) THEN (occupancy = NORMAL)
5. IF (remain = LARGE) AND (discharged = LOW) THEN (occupancy = BUSY)
6. IF (remain = LOW) AND (discharged = LARGE) THEN (occupancy = NOT_BUSY)
7. IF (admitted = LARGE) AND (death = 1) THEN (occupancy = BUSY)
8. IF (admitted = LOW) AND (death = 2) THEN (occupancy = NORMAL)
9. IF (admitted = NORMAL) AND (death = 2) THEN (occupancy = NOT_BUSY)
10. IF (remain = LARGE) AND (admitted = NORMAL) AND (death = 1) THEN (occupancy = BUSY)

11. IF (remain = NORMAL) AND (admitted = NORMAL) AND (death = 1) THEN (occupancy = NORMAL)
12. IF (remain = LARGE) AND (admitted = LOW) AND (death = 1) THEN (occupancy = NORMAL)
13. IF (remain = NORMAL) AND (admitted = LOW) AND (death = 1) THEN (occupancy = NOT_BUSY)
14. IF (remain = NORMAL) AND (admitted = LARGE) AND (discharged = NORMAL) THEN (occupancy = NORMAL)
15. IF (remain = LARGE) AND (admitted = LOW) AND (discharged = NORMAL) THEN (occupancy = NORMAL)
16. IF (remain = LOW) AND (admitted = NORMAL) AND (discharged = NORMAL) THEN (occupancy = NOT_BUSY)
17. IF (remain = LARGE) AND (death = NONE) AND (discharged = LARGE) THEN (occupancy = NORMAL)
18. IF (remain = NORMAL) AND (admitted = LARGE) AND (death = NONE) AND (discharged = LOW) THEN (occupancy = BUSY)

How many of the ILA rules (18 of them) made it onto the most interesting list (17 rules)?

We know that ILA generates shorter rules than RS1 to characterize the same concept and generates fewer rules overall. We will consider an ILA rule to have made it onto the most interesting list if the ILA rule or a shortened version of it is on the most interesting list. Consider the ward busy concept (occupancy = BUSY). Five rules to characterize this concept appeared in the most interesting list as determined by a heuristic applied to the 39 rules generated by RS1 (Rules 3, 4, 5, 8 and 12). Five rules with occupancy = BUSY were also generated by ILA (rules 1, 5, 7, 10 and 18).

Rule 3 from most interesting list

(remain := LARGE) AND (discharged := LOW) -----> (occupancy := BUSY)

is exactly the same as ILA rule 5 which is

IF (remain = LARGE) AND (discharged = LOW) THEN (occupancy = BUSY).

Rule 4 from most interesting list

(remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := NONE) ---> (occupancy := BUSY)

is exactly the same as ILA rule 18 which is

IF (remain = NORMAL) AND (admitted = LARGE) AND (death = NONE) AND (discharged = LOW) THEN (occupancy = BUSY).

Rule 5 from most interesting list

(remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := 1) -----> (occupancy := BUSY)

contains ILA rule 7 which is

IF (admitted = LARGE) AND (death = 1) THEN (occupancy = BUSY)

Rule 8 from most interesting list

(remain := LARGE) AND (discharged := NORMAL) AND (admitted := LARGE) -----> (occupancy := BUSY)

contains ILA rule 1 which is

IF (remain = LARGE) AND (admitted = LARGE) THEN (occupancy = BUSY).

It seems that RS1 has generated two rules both seen as interesting by the human evaluators while ILA generates only one rule instead.

Rule 12 from most interesting list

(remain := NORMAL) AND (discharged := LOW) AND (admitted := NORMAL) AND (death := NONE) -----> (occupancy := BUSY)

is more similar to ILA rule 18 which is

IF (remain = NORMAL) AND (admitted = LARGE) AND (death = NONE) AND (discharged = LOW) THEN (occupancy = BUSY)

than to ILA rule 10 which is

IF (remain = LARGE) AND (admitted = NORMAL) AND (death = 1) THEN (occupancy = BUSY)

What have we lost from most interesting rules 5 and 8 and does it show up in ILA 10?

3 & 4 same as 5 & 18, respectively.

5 contains 7

8 contains 1

ILA rule 10 does not show up on most interesting list

The ILA/RS1 comparison considered 17 ILA rules and 18 RSI rules. The 18 rules were decided upon by talking to the hospital administrator and also by certainty and coverage measures calculated by RS1. Users were asked to evaluate those 18 rules, not the original 39. ILA does not

order the rules from most interesting to least interesting but my heuristics does. However, to some extent ILA is employing an interestingness measure on the first cut when it decided which rules were most interesting and which were not interesting.

It was the certainty and coverage that seems to have contributed to deciding which rules should be presented to subjects for evaluation of interestingness as well as some semantics introduced by the hospital administrator. The analysis shows closeness between the 17 ILA rules and the 18 RS1 rules. It seems that in some sense ILA is implicitly including an interestingness measure as well as some notion of certainty and coverage. It says that using two different approaches comes up with similar results. My quality measure for rules is therefore verified as least regarding the first cut. Conversely, we learn something important about ILA that we did not know before. The ILA algorithm is more intelligent than we thought. The interesting measure (heuristic) first cut uses human intelligence to decide on a set of the most interesting rules. ILA does not use certainty or coverage to measure quality of rules. Yet ILA generates rules that are reduced as if certainty and coverage were used like RS1 explicitly used certainty and coverage to reduce the number of RS1 rules under consideration before applying the interestingness measure.

4.3.1 ILA experiments

The ILA algorithm in RSGUI does not execute with the full dataset, as it freezes for the three months of the FMW data. To understand where the problem might be, 12 specific rows were removed of the 89 rows in the dataset, which solved the issue. These 12 specific rows were identified by taking the following these steps:

One row was entered and ILA was executed. If this was successful then the first and second rows were entered into the database and so on until ILA froze at a specific row, which will be named i^{th} row for this example. The i^{th} row was removed and the previous successful rows and rows $i + 1, i + 2, \dots$ were loaded until ILA froze again. In the end, 77 rows were successfully entered and 12 rows were excluded.

The next question that was investigated was, why does ILA not function with those rows, and why does it give rules for 77 rows for the remainder of the dataset. The algorithm itself was examined with a small dataset for which it is possible to follow the algorithm step by step. This

dataset has been used in Chapter 2 to illustrate RS1 algorithm, and here it will be used to illustrate ILA algorithm. Table 4.4 is duplicated here for convenience.

It was observed that ILA in RSGUI does not freeze in my full data on binary decision attributes. It freezes if the data have 3 decision values. For example, my data have 3 decision values, which are (BUSY, NORMAL, and NOT_BUSY), so with these values it freezes. But when I changed the decision values to be only 2 decision values (BUSY and Normal) it generates rules for all 89 rows of the data set. We can conclude that the case described in Subsection 2.1.3 does not arise with the full data set and that the algorithm as previously implemented by Laurentian University students can handle only binary decision values.

ILA generates rules with the full data set only if we have binary decision values. However, it generates rules with 3 decision values if specific rows were removed. So why did ILA work on the 77 rows with multi-valued decision attribute. The rules that RS1 generated were compared with the rules that ILA generated and the rows that were removed were those that would have generated uncertain rules as measured by uncertainty metric that was printed out by RS1 for each rule. But we do not want to use the data set without the 12 rows since we are dealing with uncertain data.

4.3.2 ILA java code

After many experiments had been done on ILA with poor results with the full data set, fixing its code was needed. So by testing the ILA through the Eclipse application, some bugs and errors have been found on ILA code. Fixing the bugs was not the only ILA issue on RSGUI. In addition a for loop was missing some cases that makes the system freeze when executing ILA.

The previous programmer identified the j at the beginning of the class but he did not use it in the for loop or if statements. He identifies the j (row counter) as follows:

```
int j = 1;
```

The for loop that follows is a snippet of ILA code.

```
for (int m = 0; m < dtArray.length; m
    for (int r = 0; r < dtArray[m].rows; r++)
        if (!boolDTArray[m][r])
            done = false;
        if (done)
            break;
```


The if statement that stops the for loop above was missing and is needed if the row counter is greater or equal to $j = 1$ as below:

```
if(dtArray[0].rows >= j)
    j++;
else
    break;
trace += "\n";
```

The code that was before adding above (if statement) did not tell the for loop to stop once the row counter has reached the j limit, and it is below:

```
j++;
trace += "\n";
```

For years, students have been unable to process all uncertain data using ILA and all because of an infinite loop.

4.4 Interestingness measures

A user uses the RSGUI to discover patterns in data. To get the best result from these data, it needs to extract interesting rules. For the rule to be interesting, it needs to be useful and eventually understandable. Interestingness measures should be found by two parts: objective and subjective.

Interestingness measures are for selecting and ranking patterns, according to how interesting they are from the user's perspective. A good survey of interestingness measures can be found in [59].

In that work, criteria to determine interest of patterns are the following:

Conciseness: a rule can be concise if it has few attributes relative to the others.

Generality/ Coverage: a rule can be general or coverage if it covers a large subset of the data.

Reliability: a rule can be reliable if it is strong as measured by some metric.

Peculiarity: a rule is peculiar if it is unusual and unknown beforehand to the user.

Diversity: a rule is diverse if its attributes differ from each other.

Novelty: a rule is novel if the user is unaware of it beforehand and cannot produce it from known rules.

Surprisingness: a rule is surprising if it is unexpected and contradicts user knowledge.

Utility: a rule has utility if it is useful and helps a person to reach a goal.

Applicability: a rule is applicable if it can be applied in the future in same domain.

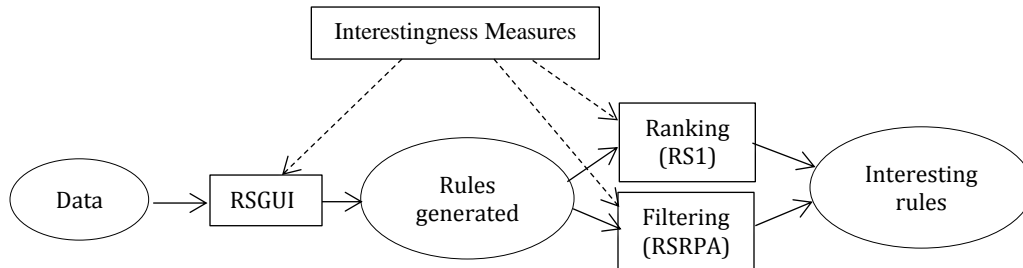


Figure 4.2: Interestingness measures in play for rule generation

In [60] alternate approaches to normalization of measures are provided to make different measures comparable by various methods. A rule generated may be good on some and bad on others, so developing a formula for taking into consideration all the evaluation criteria is needed to get a good result. The formula that I developed uses ranking from -4 to +4 for each criterion and adding up the positive and negative scores yields one weight for the rule. The rankings for each individual criterion are as follows:

+4 *Reliability*

+3 *Utility*

+2 *Conciseness*

+1 *Generality/ Coverage*

0 *Novelty*

-1 *Applicability*

-2 *Diversity*

-3 *Surprisingness*

-4 *Peculiarity*

The results of 22 rules found when death was selected as a decision attribute using RS1 with 77 rows and ordered from most interesting to least interesting with their evaluation table can be found in decision rules 1 with Table C1 in Appendix C.

According to the results of Decision Rules 1 in Appendix C, we can say that low discharges and busy wards may cause deaths. Applying ILA on the data with 77 rows and occupancy as a decision attribute, the results gave all rules a certainty of 1, which means all of the rules generated from ILA were certain rules and the 12 rows that were removed are uncertain.

The results of the 22 rules found using RS1 when selecting occupancy as a decision attribute with 77 rows and ordered from most interesting to least interesting with their evaluation table can be found in Decision Rules 2 with Table C2 in Appendix C.

Also according to the results Decision Rules 2 in Appendix C, a ward could be busy if there is a large amount of remaining patients with low discharge rates. These findings contribute toward an understanding of why ILA was not working properly and have helped me to improve RSGUI for teaching rough set concepts and for developing applications based on rough sets. Table C1 is an evaluation by the criteria for the death as a decision attribute, and Table C2 is an evaluation by the criteria for the occupancy as a decision attribute can be found in Appendix C.

Firstly, ILA algorithm was applied on the data containing the decision attribute occupancy and 12 of the 89 rows were removed from the data, in order to prevent ILA from freezing which was described in Subsection 4.3. I later decided to use another decision attribute, which is death. ILA with 77 rows and death as a decision attribute was applied on the data, but it froze. It also froze even if returning back the 12 rows to be 89 rows that were removed from the data when the occupancy was used as a decision attribute. So after many experiments I have done, I decided to fix ILA algorithm as the issue was described in Subsection 4.3.1, and ILA fixed as has been described in Subsection 4.3.2. For now we work instead with RS1. So ILA fixed results can be compared with those of RS1.

Table 4.4, Table 4.5, and Table 4.6 demonstrate the results of three different decision values, BUSY, NORMAL, and NOT_BUSY for occupancy attribute respectively. Table 4.4, Table 4.5, and Table 4.6 are all certain rules using RS1 with the occupancy as a decision attribute. The rules are separated into 3 tables; busy ward, normal ward, and not busy ward respectively. Table 4.10 contains only uncertain rules for all of the three decision values of the occupancy decision

attribute.

Table 4.4: Occupancy as a decision attribute for busy value table for all certain rules

(remain := LARGE) AND (discharged := LOW) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
(remain := LARGE) AND (discharged := NORMAL) AND (admitted := LARGE) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
(remain := LARGE) AND (discharged := LARGE) AND (admitted := NORMAL) AND (death := 1) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 1/89]
(remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := NONE) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 2/89]
(remain := LARGE) AND (discharged := NORMAL) AND (admitted := NORMAL) AND (death := 1) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 1/89]
(remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := 1) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 1/89]

Table 4.5: Occupancy as a decision attribute for normal value table for all certain rules

(discharged := NONE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 2/89]
(discharged := NORMAL) AND (death := 2) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/89]
(discharged := LOW) AND (death := 2) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/89] >
(discharged := LARGE) AND (death := NONE) AND (remain := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/89]
(discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 5/89]
(discharged := NORMAL) AND (death := NONE) AND (remain := LARGE) AND (admitted := LOW) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 2/89]
(discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 3/89]
(discharged := NORMAL) AND (death := NONE) AND (remain := LOW) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/89]
(discharged := LOW) AND (death := 1) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/89]
(discharged := NORMAL) AND (death := 1) AND (remain := LARGE) AND (admitted := LOW) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/89]
(discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NONE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/89]

Table 4.6: Occupancy as a decision attribute for not busy value table for all certain rules

(admitted := LOW) AND (remain := LOW) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 5/89]
(admitted := LOW) AND (remain := NORMAL) AND (discharged := LARGE) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 2/89]
(admitted := NORMAL) AND (remain := LOW) AND (discharged := LARGE) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 1/89]
(admitted := NORMAL) AND (remain := LOW) AND (discharged := NORMAL) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 1/89]
(admitted := NORMAL) AND (remain := LARGE) AND (discharged := LARGE) AND (death := 2) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 1/89]
(admitted := LOW) AND (remain := NORMAL) AND (discharged := NORMAL) AND (death := 1) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 1/89]

Table 4.7: Occupancy as a decision attribute for all decision values with certainty < 1

(remain := LARGE) AND (discharged := NORMAL) AND (admitted := NORMAL) AND (death := NONE) ----->
--

(occupancy := BUSY)[certainty = 0.5][coverage = 2/89]
(remain := NORMAL) AND (discharged := LOW) AND (admitted := NORMAL) AND (death := NONE) -----> (occupancy := BUSY)[certainty = 0.2727272727272727][coverage = 11/89]
(discharged := NORMAL) AND (death := NONE) AND (remain := LARGE) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.5][coverage = 2/89]
(discharged := LARGE) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.3333333333333333][coverage = 3/89]
(discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.9333333333333333][coverage = 15/89]
(discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.7272727272727273][coverage = 11/89]
(discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LOW) -----> (occupancy := NORMAL)[certainty = 0.3333333333333333][coverage = 3/89]
(discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.5][coverage = 6/89]
(discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LOW) -----> (occupancy := NORMAL)[certainty = 0.8333333333333334][coverage = 6/89]
(discharged := LOW) AND (death := 1) AND (remain := LOW) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.5][coverage = 2/89]
(admitted := NORMAL) AND (remain := NORMAL) AND (discharged := LARGE) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]
(admitted := NORMAL) AND (remain := NORMAL) AND (discharged := NORMAL) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.06666666666666667][coverage = 15/89]
(admitted := LOW) AND (remain := NORMAL) AND (discharged := NORMAL) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]
(admitted := NORMAL) AND (remain := LOW) AND (discharged := LOW) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.5][coverage = 6/89]
(admitted := LOW) AND (remain := NORMAL) AND (discharged := LOW) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.16666666666666666][coverage = 6/89]
(admitted := NORMAL) AND (remain := LOW) AND (discharged := LOW) AND (death := 1) -----> (occupancy := NOT_BUSY)[certainty = 0.5][coverage = 2/89]

Interesting rules extracted from Table 4.4, Table 4.5, Table 4.6 and Table 4.10 using RS1 and occupancy as a decision attribute, were ordered from most interesting to least interesting, using evaluation criteria described in Subsection 4.4. The result will be shown in Subsection 6.1.

Table 4.11, Table 4.12, and Table 4.13 have all certain rules using RS1 with death as a decision attribute. The rules are separated into 3 tables, 2 deaths in the ward, 1 death in the ward, and no deaths in the ward respectively. Table 4.14 contains all uncertain rules for all the three decision values of death decision attribute.

Table 4.8: Death as a decision attribute for 2 deaths decision value for all certain rules

(remain := LARGE) AND (admitted := NORMAL) AND (occupancy := NOT_BUSY) ---> (death := 2)[certainty = 1.0][coverage = 1/89]
(remain := NORMAL) AND (admitted := LARGE) AND (occupancy := NORMAL) AND (discharged := LOW) ---> (death := 2)[certainty = 1.0][coverage = 1/89]

Table 4.9: Death as a decision attribute for 1 death decision value for all certain rules

(discharged := LARGE) AND (admitted := NORMAL) AND (occupancy := BUSY) --->(death := 1)[certainty = 1.0][coverage = 1/89]

Table 4.10: Death decision attribute for no deaths decision value for all certain rules

(discharged := NONE) ---> (death := NONE)[certainty = 1.0][coverage = 2/89]
(discharged := LOW) AND (admitted := LOW) ---> (death := NONE)[certainty = 1.0][coverage = 12/89]
(discharged := LOW) AND (admitted := NONE) ---> (death := NONE)[certainty = 1.0][coverage = 1/89]
(discharged := NORMAL) AND (admitted := LARGE) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 5/89]
(discharged := LARGE) AND (admitted := NORMAL) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 3/89]
(discharged := NORMAL) AND (admitted := NORMAL) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 15/89]
(discharged := LOW) AND (admitted := LARGE) AND (remain := LOW) ---> (death := NONE)[certainty = 1.0][coverage = 3/89]
(discharged := NORMAL) AND (admitted := LARGE) AND (remain := LOW) ---> (death := NONE)[certainty = 1.0][coverage = 1/89]
(discharged := LARGE) AND (admitted := NORMAL) AND (remain := LOW) ---> (death := NONE)[certainty = 1.0][coverage = 1/89]
(discharged := LOW) AND (admitted := NORMAL) AND (remain := LARGE) ---> (death := NONE)[certainty = 1.0][coverage = 2/89]
(discharged := NORMAL) AND (admitted := NORMAL) AND (remain := LOW) ---> (death := NONE)[certainty = 1.0][coverage = 1/89]
(discharged := NORMAL) AND (admitted := LOW) AND (remain := NORMAL) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 1/89]
(discharged := LARGE) AND (admitted := NORMAL) AND (remain := LARGE) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 1/89]
(discharged := LOW) AND (admitted := NORMAL) AND (remain := NORMAL) AND (occupancy := BUSY) ---> (death := NONE)[certainty = 1.0][coverage = 3/89]
(discharged := NORMAL) AND (admitted := NORMAL) AND (remain := LARGE) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 1/89]

Table 4.11: Death as a decision attribute for all death decision values with certainty < 1

(remain := LARGE) AND (admitted := LOW) AND (occupancy := NORMAL) AND (discharged := NORMAL) ---> (death := 2)[certainty = 0.25][coverage = 4/89]
(discharged := NORMAL) AND (admitted := LARGE) AND (occupancy := BUSY) AND (remain := LARGE) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/89]
(discharged := LOW) AND (admitted := LARGE) AND (occupancy := BUSY) AND (remain := NORMAL) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/89]
(discharged := NORMAL) AND (admitted := NORMAL) AND (occupancy := BUSY) AND (remain := LARGE) ---> (death := 1)[certainty = 0.5][coverage = 2/89]
(discharged := LOW) AND (admitted := NORMAL) AND (occupancy := NORMAL) AND (remain := NORMAL) ---> (death := 1)[certainty = 0.1111111111111111][coverage = 9/89]
(discharged := NORMAL) AND (admitted := LOW) AND (occupancy := NOT_BUSY) AND (remain := NORMAL) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/89]
(discharged := LOW) AND (admitted := NORMAL) AND (occupancy := NORMAL) AND (remain := LOW) ---> (death := 1)[certainty = 0.25][coverage = 4/89]
(discharged := LARGE) AND (admitted := LOW) AND (occupancy := NOT_BUSY) AND (remain := NORMAL) ---> (death := 1)[certainty = 0.5][coverage = 2/89]
(discharged := LOW) AND (admitted := NORMAL) AND (occupancy := NOT_BUSY) AND (remain := LOW) ---> (death := 1)[certainty = 0.25][coverage = 4/89]
(discharged := NORMAL) AND (admitted := LARGE) AND (remain := LARGE) AND (occupancy := BUSY) ---> (death := NONE)[certainty = 0.6666666666666666][coverage = 3/89]
(discharged := NORMAL) AND (admitted := LOW) AND (remain := LARGE) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.5][coverage = 4/89]
(discharged := LOW) AND (admitted := LARGE) AND (remain := NORMAL) AND (occupancy := BUSY) ---> (death := NONE)[certainty = 0.6666666666666666][coverage = 3/89]
(discharged := NORMAL) AND (admitted := NORMAL) AND (remain := LARGE) AND (occupancy := BUSY) ---> (death := NONE)[certainty = 0.5][coverage = 2/89]
(discharged := LOW) AND (admitted := NORMAL) AND (remain := NORMAL) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.8888888888888888][coverage = 9/89]

(discharged := NORMAL) AND (admitted := LOW) AND (remain := NORMAL) AND (occupancy := NOT_BUSY) ---> (death := NONE)[certainty = 0.6666666666666666][coverage = 3/89]
(discharged := LOW) AND (admitted := NORMAL) AND (remain := LOW) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.75][coverage = 4/89]
(discharged := LARGE) AND (admitted := LOW) AND (remain := NORMAL) AND (occupancy := NOT_BUSY) ---> (death := NONE)[certainty = 0.5][coverage = 2/89]
(discharged := LOW) AND (admitted := NORMAL) AND (remain := LOW) AND (occupancy := NOT_BUSY) ---> (death := NONE)[certainty = 0.75][coverage = 4/89]

Interesting rules extracted from Table 4.11, Table 4.12, Table 4.13 and Table 4.14 using RS1and, this time, death as a decision attribute, are ordered from most interesting to least interesting using evaluation criteria described in Section 4.4 and, again, will be shown in Subsection 6.2.

All of the experiments have been done upon FMW data table, since the hospital has 10 wards. The process that gives the results of FMW can also be applied to the rest of the wards.

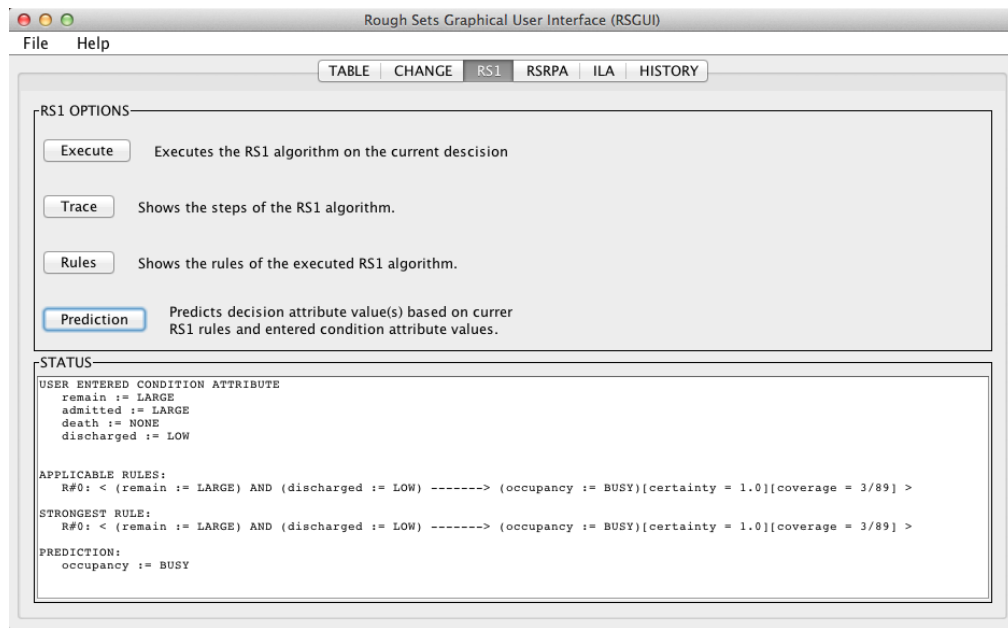


Figure 4.3: RS1 prediction rules

To explain the Figure 4.3, if the entered condition attribute values are as follows:

If remain = LARGE, AND admitted = LARGE, AND death = NONE, AND discharged = LOW

Then the applicable decision rule is:

(remain = LARGE) AND (discharged = LOW) -----> (occupancy = BUSY)[certainty = 1.0][coverage = 3/89]

Based on RS1 rules and entered condition attribute values, we get the predicted decision attribute value of occupancy is BUSY.

Also, if the entered condition attribute values for death as a decision attribute were as follows:

If remain = LARGE, AND admitted = NORMAL, AND occupancy = BUSY, AND discharged = NORMAL

Then the applicable decision rules are:

1. (discharged = NORMAL) AND (admitted = NORMAL) AND (remain = LARGE) AND (occupancy = BUSY) ----->
(death = NONE)[certainty = 0.5][coverage = 2/89]
2. (discharged = NORMAL) AND (admitted = NORMAL) AND (occupancy = BUSY) AND (remain = LARGE) ----->
(death = 1)[certainty = 0.5][coverage = 2/89]

Above rules are a decision attribute value prediction based on RS1 algorithm and user entered condition attribute values.

4.5 RSRPA algorithm

RSRPA algorithm has been described in Subsection 2.1.4 and the result and rules of the algorithm will be illustrated in this subsection.

The RSRPA algorithm as described in Subsection 2.1.4 is given the decision attribute values and the algorithm is used to predict the condition attribute values. It is illustrated in Figure 4.4 that the decision attribute value, which is BUSY in concept #0 was given and the algorithm predicted the best condition attribute values for that decision attribute value, and so on for the remaining concepts. The obtained RSRPA rules that are seen in Figure 4.4 are illustrated in Table 4.15.

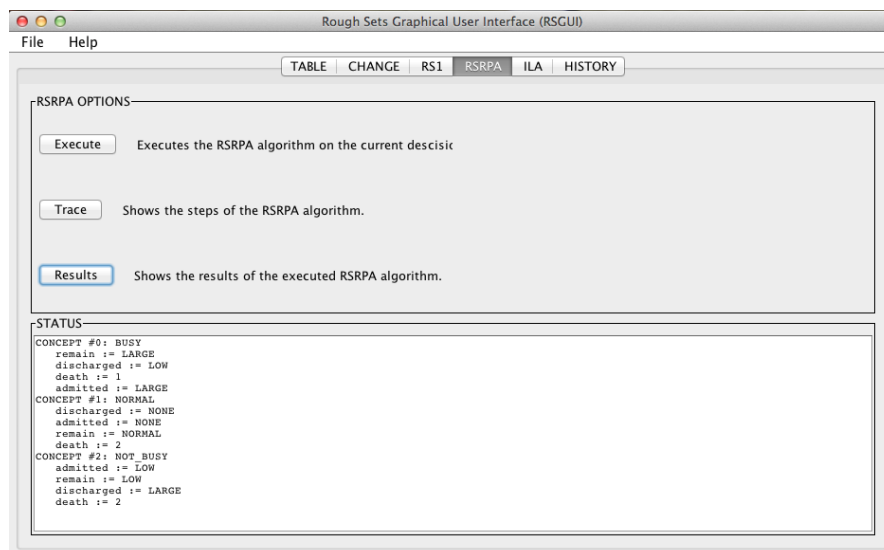


Figure 4.4: RSRPA result

The RSRPA algorithm was applied on the data with occupancy as a decision attribute by which it obtained the rules shown in Table 4.15, and also was applied on the same data with death as a decision attribute by which it obtained the rules shown in Table 4.16.

Table 4.12: RSRPA for occupancy decision attribute

Concepts		
Busy ward	Normal ward	Not busy ward
remain := LARGE discharged := LOW death := 1 admitted := LARGE	discharged := NONE admitted := NONE remain := NORMAL death := 2	admitted := LOW remain := LOW discharged := LARGE death := 2

Table 4.13: RSRPA for death decision attribute

Concepts		
2 death in the ward	1 death in the ward	No death in the ward
remain = LARGE admitted = NORMAL	remain = LARGE admitted = NORMAL	remain = NORMAL admitted = NONE

discharged = LARGE	discharged = LARGE	discharged = NONE
occupancy = NOT_BUSY	occupancy = BUSY	occupancy = NORMAL

The results of Table 4.16 are consistent with our expectations. Since death is a significant event in a ward, it is critical to understand conditions on the ward that could lead to a death.

ILA freezing issues that were described earlier in Subsection 4.3.1 have been solved. The java code of ILA had some bugs, which made the RSGUI freeze when executing ILA. Also the problem of the 12 rows that was mentioned in Subsection 4.3.1 and the two decision value restriction now are solved. The algorithm in java was written from another paper that was missing some cases but, more than that, bugs in the implementation caused the freezes and as well resulted in processing no more than two decision values. Now ILA in RSGUI runs with more than two decision values and gives a good result by comparison with RS1 as shown following:

1. IF (remain = LARGE) AND (admitted = LARGE) THEN (occupancy = BUSY)
2. IF (remain = LOW) AND (admitted = LARGE) THEN (occupancy = NORMAL)
3. IF (remain = LOW) AND (admitted = LOW) THEN (occupancy = NOT_BUSY)
4. IF (remain = NORMAL) AND (death = 2) THEN (occupancy = NORMAL)
5. IF (remain = LARGE) AND (discharged = LOW) THEN (occupancy = BUSY)
6. IF (remain = LOW) AND (discharged = LARGE) THEN (occupancy = NOT_BUSY)
7. IF (admitted = LARGE) AND (death = 1) THEN (occupancy = BUSY)
8. IF (admitted = LOW) AND (death = 2) THEN (occupancy = NORMAL)
9. IF (admitted = NORMAL) AND (death = 2) THEN (occupancy = NOT_BUSY)
10. IF (remain = LARGE) AND (admitted = NORMAL) AND (death = 1) THEN (occupancy = BUSY)
11. IF (remain = NORMAL) AND (admitted = NORMAL) AND (death = 1) THEN (occupancy = NORMAL)
12. IF (remain = LARGE) AND (admitted = LOW) AND (death = 1) THEN (occupancy = NORMAL)
13. IF (remain = NORMAL) AND (admitted = LOW) AND death = 1) THEN (occupancy = NOT_BUSY)
14. IF (remain = NORMAL) AND (admitted = LARGE) AND (discharged = NORMAL) THEN (occupancy = NORMAL)
15. IF (remain = LARGE) AND (admitted = LOW) AND (discharged = NORMAL) THEN (occupancy = NORMAL)
16. IF (remain = LOW) AND (admitted = NORMAL) AND (discharged = NORMAL) THEN (occupancy = NOT_BUSY)
17. IF (remain = LARGE) AND (death = NONE) AND (discharged = LARGE) THEN (occupancy = NORMAL)
18. IF (remain = NORMAL) AND (admitted = LARGE) AND (death = NONE) AND (discharged = LOW) THEN (occupancy = BUSY)

Above rules are all the rules that ILA has generated. They are useful and partly the same as RS1 rules. However, ILA rules are shorter than RS1 rules, which means ILA rules are more recommended rules according to:

- 1- ILA generates only interesting and shorter rules

2- By only seeing the discharge value, we can predict the ward occupancy. For example all low discharge value going to make a busy ward. However, Large discharge value most likely going to make a not busy ward.

3- ILA has generated only 18 rules but RS1 has generated 39 rules.

Therefore, reporting ILA rules to the Health Ministry instead RS1 rules is better and going to save data analyzer few times to check which term or attribute could make a busy ward. Also, what ILA has done is display the interesting rules and ignore some rules that are not interesting, but RS1 has displayed all rules whether interesting or not interesting.

In this chapter, a summary has been prepared of all of the experiments that were accomplished based on the different algorithms applied upon the medical ward data acquired over the three month period. The summary shows, for each experiment, the attributes and their designation as condition or decision, how many rows were considered, how many columns were considered, and how many rules we got. Interestingness heuristics were developed for selecting and ranking patterns according to how interesting they are from the user's perspective. Quantitative plausibility measures were applied to evaluate the quality of predictions. The interesting and plausible rules have been selected and presented in this chapter.

Chapter 5

5 Second data experiment

A student in university is interested to get a high average and complete their degrees with high GPA. In this chapter a web based RSGUI called NewRSGUI will be illustrated and prediction in student data will be done using NewRSGUI. The prediction to be made is “what courses should I take together to get a high GPA?”

5.1 Student transcripts

Course data for Math and Computer Science students have been obtained from the registrars’ office at Laurentian University. Identification of students has been removed. The Research Ethics Board at Laurentian University has approved to use the data in this thesis, (see Appendix A). The data follow each student over a four-year period and show all courses the student has taken not only the core courses they took (Math/Computer Science) but also their electives. We will discover general rules from these data to predict a collection of courses that incoming future students could take, in order to fulfill the program requirements. The predicted rules would alleviate, at least in part, the Math and Computer Sciences faculty workload for advising students on which courses to take.

5.2 NewRSGUI

On student transcript data I used NewRSGUI, which is a new version of the Rough Set Graphical User Interface and can achieve the same function as RSGUI. The differences between the two are method of startup and the input file format. RSGUI deals with word-based text databases (see Figure 2.3). It runs from the command line, in which the user has to type several commands in the command line, before the GUI is launched. Text files only can be recognized to read as an input. In order to solve these problems, it is required to convert the RSGUI to a different version, which could improve the problem of user unfriendly input method. NewRSGUI has become an executable file and can be launched by double clicking its icon. Also new in NewRSGUI, the input files no longer .txt files, and can be connected with a real database to directly extract its

data. RS1, RSRPA (Rough Set Reverse Prediction Algorithm) and ILA (Inductive Learning Algorithm) are still provided by NewRSGUI and therefore, the user does not need to relearn how to use the NewRSGUI. NewRSGUI is a more convenient tool for users to make a prediction, connect to their datasets, and it is easy to run. For these reasons, I am using NewRSGUI in my secondary data. NewRSGUI runs as follows:

1. Click the NewRSGUI icon to display the interface in Figure 5.2 and to type the information.
2. If the information typed in 1 is correct, it display another window, which allows the user to select the table of the dataset in MySql and the number of decision attributes the user is looking for.
3. Afterwards, the Decision Table collects all the data from the specified table in the database schema and passes the data as a parameter to the RSGUI and the main frame is launched.

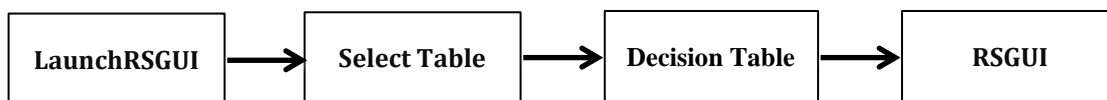


Figure 5.1: Flowchart of NewRSGUI design

The screenshot shows a user interface window with the following elements:

- Database URL:
- Database Port:
- Database Schema:
- Username:
- Password:
- Buttons: and

Figure 5.2: NewRSGUI user interface

The NewRSGUI can now be called from web medical interfaces, as seen in Figure F.1 in Appendix F. This can help hospital administrators generate rules from the same MySQL databases that are used for collecting patients' data through the web interface (Appendix F).

5.3 NewRSGUI snippets java code

This subsection shows snippets of java codes of NewRSGUI that I am using in this chapter to analyze the student records.

The following snippets of code are to create the labels and the text fields to identify the address or the Server of the database that is going to be used or analyzed using the methods that exist in the RSGUI.

```
JLabel labelURL = new JLabel("Database URL:");
...
JLabel labelPort = new JLabel("Database Port:");
...
JLabel label1 = new JLabel("Database Schema:");
...
JLabel label2 = new JLabel("Username:");
...
JLabel label3 = new JLabel("Password:");
...
```

Figure 5.2 shows how the final screen looks after being generated by the above codes. This is much better than the command line launching of RSGUI. Once the interface is displayed, there is a function that acquires data from the database identified in the interface, and the function code is as follows:

```
public static Connection getMySQLConnection(String durl,
    String port, String sch, String un, String pwd) throws Exception {
    String driver = "org.gjt.mm.mysql.Driver";
    String url = "jdbc:mysql://" + durl + ":" + port + "/" + sch;
    String username = un;
    String password = pwd;
    conn = DriverManager.getConnection(url, username, password);
    return conn
}
```

The SelectTable class is implemented as shown below, which is a Java frame class. In this class, the input is the connection. It lets the user select one table as an input for the main function. Also, there is a text field by which user is asked to input the decision attribute numbers, of columns that will be treated as decision attributes in the selected table which is seen in Figure 5.3.

```
public SelectTable(final Connection Conn)
try {
    DatabaseMetaData dmd = Conn.getMetaData();
    ResultSet rs = dmd.getTables(null, null, "%", null);
    while (rs.next()) {
        comboBoxTable.addItem(rs.getString(3));
    }
}
```



Figure 5.3: Select table and number of decision attributes interface

This subsection and Appendix G are NewRSGUI programming snippets of code that have been used to create the text fields and interface to identify the server of the database in MySQL. NewRSGUI is executing same algorithms that RSGUI is featuring but with improved startup and the input file format instead of the command lines that was used on RSGUI.

5.4 Student data analytics

These days, one of the biggest challenges is predicting the paths of students in higher education [61]. A sample of a student's records transcripts for the first year is in Figure 5.4.

5.4.1 Nature of the student records

The data were obtained from Laurentian University's registrar office from math and computer science students. We asked the registrar office for a minimum of 60 student's records and the quotation text from the email has sent to them is seen below:

"We need a minimum of 60 students' records following them for a four year period showing all courses the student has taken not only the core courses they took (Math/Computer Science) but also their electives. We will discover general rules from these data to predict a collection of courses that incoming students should take to fulfill the program requirements. This would alleviate faculty workload for advising students on which courses to enroll."

We received 30 records from the registrar's office after sending the first email, but the data we received showed that only 7 of those students had completed the degree. Afterwards, we sent a second email, asking for the remaining records. See the quotation below, from my email to them:

"We are waiting more data like what you gave us previously. The previous data was short because it contained

students who withdrew and also students who had not completed their degree. We want to follow students through a 4 year period.”

We got the second data draft with only 10 students from 36 records that completed their degree, which make the total of 17 students’ records to study.

Date	:	05/02	
Degree	:	B.A.(General)	
Majors	:	Computer Science	
Academic Programs			
Arts - 4 Year		09/97	08/98
BA(Hons) Computer Science		09/98	09/00
BA3 Computer Science		09/00	05/02
Admit Status			
UG - Sec School-Ontario Current			
Language Test Score			
COMP-E Score: 1			
Academic Standing			
2001AW	Program Complete		
Academic Record			
1997 Winter Academic Term		Credit	Mark
CLAS-1006EL	Greek Civilization	3.0	52
CLAS-1007EL	Roman Civilization	3.0	61
COSC-1046EL	Intro to Computer Science I	3.0	74 Replaced
COSC-1047EL	Intro to Computer Science II	0.0	46 Replaced
HIST-1106EL	Intro to the Twentieth Century	3.0	71
HIST-1107EL	Contemp. issues-Historical Pers	3.0	64
MATH-1036EL	Calculus I	0.0	31
MATH-1056EL	Discrete Mathematics I	0.0	25 Replaced
MATH-1912EL	Elementary Calculus	3.0	77
Session Average: 66.5			

Figure 5.4: Sample of a first year student transcript

The student’s transcript data has been inserted into MySQL database by MySQL queries. It was a long process because of the volume of records. The student transcripts data are stored in MySQL as illustrate in Figure 5.5.

```
Run SQL query/queries on database test: ⓘ
INSERT INTO `tran` (`student`, `course`, `number`, `year`, `grade`, `courseStatus`, `degreeOptions`, `sessionGPA`) VALUES
('c1', 'COSC', 1046, '1', 'B', 'YES', '3years', 'C'),
('c1', 'COSC', 1047, '1', 'F', 'YES', '3years', 'C'),
('c1', 'ARTS', 9100, '1', 'S', 'NO', '3years', 'S'),
('c1', 'FRAN', 9100, '1', 'S', 'NO', '3years', 'S'),
('c1', 'COMM', 1107, '1', 'S', 'NO', '3years', 'S'),
('c1', 'COMM', 1106, '1', 'S', 'NO', '3years', 'S'),
('c1', 'MATH', 1011, '1', 'D', 'NO', '3years', 'D')
```

Figure 5.5: MySQL query for student records

Records inserted into MySQL contain 537 rows and 8 columns for the 17 computer science students. Condition attributes are: Student number, course, course number, year, course grade, course status, student degree option, and Term GPA. Data on the NewRSGUI are seen as in Figure 5.6.

File Help		TABLE CHANGE RS1 RSRPA ILA HISTORY						
Row	degreeOptions	course	number	year	yearGPA	required?	student	grade
0	3years	ARTS	9100	1	S	NO	c1	S
1	3years	FRAN	9100	1	S	NO	c1	S
2	3years	COMM	1107	1	S	NO	c1	S
3	3years	COMM	1106	1	S	NO	c1	S
4	3years	MATH	1911	1	D	NO	c1	D
5	3years	MATH	1912	1	D	NO	c1	F
6	3years	COSC	1046	1	C	YES	c1	B
7	3years	COSC	1047	1	C	YES	c1	F
8	3years	ECON	1005	1	C	NO	c1	B
9	3years	FOLK	2005	1	C	NO	c1	C
10	3years	GEOG	1026	1	C	NO	c1	C
11	3years	GEOG	2606	1	C	NO	c1	C
12	3years	MATH	1056	1	C	YES	c1	C
13	3years	PHIL	2876	1	C	NO	c1	C
14	3years	COSC	1047	2	C	YES	c1	B
15	3years	COSC	2006	2	C	YES	c1	D
16	3years	COSC	2007	2	C	NO	c1	B
17	3years	MATH	2056	2	C	YES	c1	F
18	3years	COSC	2307	2	C	YES	c1	D
19	3years	COSC	2406	2	C	YES	c1	F
20	3years	GEOG	1027	2	C	NO	c1	B
21	3years	ECON	2015	2B	A	NO	c1	A
22	3years	MATH	2056	3	C	YES	c1	D
23	3years	COSC	2306	3	C	YES	c1	C
24	3years	COSC	2406	3	C	YES	c1	D
25	3years	COSC	3127	3	C	YES	c1	F
26	3years	COSC	3406	3	C	YES	c1	C
27	3years	COSC	3407	3	C	YES	c1	D
28	3years	COSC	3707	3	C	NO	c1	C
29	3years	COSC	4436	3	C	NO	c1	D
30	3years	COSC	4506	3	C	NO	c1	C
31	3years	COSC	4606	3	C	NO	c1	A
32	4years	CLAS	1006	1	C	NO	c2	D
33	4years	CLAS	1007	1	C	NO	c2	D
34	4years	COSC	1046	1	C	YES	c2	B
35	4years	COSC	1047	1	C	YES	c2	F
36	4years	HIST	1106	1	C	NO	c2	B
37	4years	HIST	1107	1	C	NO	c2	C

Figure 5.6: Student data on NewRSGUI

5.5 Prediction results

As mentioned in Subsection 2.2, a user can merge, change, or remove columns through the RSGUI interface without going to the database. However, merging, changing, or removing columns from the RSGUI does not affect the columns and their values in the actual database, but only changes the user's view of the database. A future student can follow a good (an exemplary) graduated student by his/her records. For example, if the user is looking for a student who received the highest mark (A) in course Discrete Mathematics II (MATH 2056), the Term GPA and the year of the degree they are enrolled in, the user can predict those conditions demonstrated below:

1. Select 3 decisions attributes, which are student number, year, and Term GPA.
2. Remove course status, and degree option columns.
3. Execute RS1 algorithm.
4. Click prediction button to predict conditions.
5. Select, which course name you are looking to predict (in this example MATH).
6. Select the number of the course (in this example 2056).
7. Select the grade course between the five grades (A, B, C, D, and F). (This example A).

In the end, the user will see the results, according to the records that we obtained as illustrate in Figure 5.7.

Prediction	
Predicts decision attribute value(s) based on current RS1 rules and entered condition attribute values.	
STATUS	
USER ENTERED CONDITION ATTRIBUTE course := MATH number := 2056 grade := A	
APPLICABLE RULES: R#316: < (number := 2056) AND (grade := A) AND (course := MATH) -----> (student AND year AND sessionGPA := c11 AND 2 AND B)[certainty = 0.5][coverage = 2/537] > R#329: < (number := 2056) AND (grade := A) AND (course := MATH) -----> (student AND year AND sessionGPA := c12 AND 2 AND A)[certainty = 0.5][coverage = 2/537] >	
STRONGEST RULE: R#316: < (number := 2056) AND (grade := A) AND (course := MATH) -----> (student AND year AND sessionGPA := c11 AND 2 AND B)[certainty = 0.5][coverage = 2/537] >	
PREDICTION: student AND year AND sessionGPA := c11 AND 2 AND B	

Figure 5.7: RS1 prediction of MATH 2056

It is obvious that there are only two students who have obtained (A) in (MATH 2056) course. However, student (c12) has received (A) in the Term GPA, while student (c11) has received (B) in the Term GPA. Therefore, following student (c12) courses is the best choice for the future student, or at least an example for the second year courses, because they have received (A) in the second year, as seen in Figure 5.7.

Same previous steps for another example of looking for a student, who has taken (A) in Data Structures course (COSC 2006) to follow. See results in Figure 5.8.

Prediction	
Predicts decision attribute value(s) based on current RS1 rules and entered condition attribute values.	
STATUS	
USER ENTERED CONDITION ATTRIBUTE course := COSC number := 2006 grade := A	
APPLICABLE RULES: R#50: < (number := 2006) AND (grade := A) AND (course := COSC) -----> (student AND year AND sessionGPA := c2 AND 3 AND C)[certainty = 0.3333333333333333][coverage = 3/537] > R#326: < (number := 2006) AND (grade := A) AND (course := COSC) -----> (student AND year AND sessionGPA := c12 AND 2 AND A)[certainty = 0.3333333333333333][coverage = 3/537] > R#493: < (number := 2006) AND (grade := A) AND (course := COSC) -----> (student AND year AND sessionGPA := c17 AND 2 AND B)[certainty = 0.3333333333333333][coverage = 3/537] >	
STRONGEST RULE: R#50: < (number := 2006) AND (grade := A) AND (course := COSC) -----> (student AND year AND sessionGPA := c2 AND 3 AND C)[certainty = 0.3333333333333333][coverage = 3/537] >	
PREDICTION: student AND year AND sessionGPA := c2 AND 3 AND C	

Figure 5.8: RS1 prediction of COSC 2006

From Figure 5.8, it is obvious that there are 3 students who have obtained (A) in (COSC2006), but only student (c12) again has obtained (A) in the Term GPA.

In this example, the future student should review (c12) student's transcript on the data, in order to know which courses student (c12) was enrolled in their second year.

Finally, we will also use the course COSC 3407 to further illustrate the predictions based on the current RS1 rules, using the same conditions and decision attributes of the previous example, as demonstrated in Figure 5.9.

Prediction	Predicts decision attribute value(s) based on current RS1 rules and entered condition attribute values.
STATUS	
USER ENTERED CONDITION ATTRIBUTE course := COSC number := 3407 grade := A	
NO APPLICABLE RULES - NO PREDICTION CAN BE MADE	

Figure 5.9: RS1 prediction for COSC 3407

By the conditions illustrated in Figure 5.9, we can see that no students have obtained (A) in the course (COSC 3407). Therefore, we can search for the grade (B), (C), or so on, which is illustrated in Figure 5.10.

Prediction	Predicts decision attribute value(s) based on current RS1 rules and entered condition attribute values.
STATUS	
USER ENTERED CONDITION ATTRIBUTE course := COSC number := 3407 grade := B	
APPLICABLE RULES: R#98: < (number := 3407) AND (grade := B) AND (course := COSC) -----> (student AND year AND termGPA := c3 AND 4 AND B)[certainty = 0.25][coverage = 4/537] > R#217: < (number := 3407) AND (grade := B) AND (course := COSC) -----> (student AND year AND termGPA := c8 AND 3 AND B)[certainty = 0.25][coverage = 4/537] > R#311: < (number := 3407) AND (grade := B) AND (course := COSC) -----> (student AND year AND termGPA := c11 AND 2 AND B)[certainty = 0.25][coverage = 4/537] > R#506: < (number := 3407) AND (grade := B) AND (course := COSC) -----> (student AND year AND termGPA := c17 AND 3 AND A)[certainty = 0.25][coverage = 4/537] >	
STRONGEST RULE: R#98: < (number := 3407) AND (grade := B) AND (course := COSC) -----> (student AND year AND termGPA := c3 AND 4 AND B)[certainty = 0.25][coverage = 4/537] >	
PREDICTION: student AND year AND termGPA := c3 AND 4 AND B	

Figure 5.10: RS1 second prediction for COSC 3407

Student (c17) had obtained (B) in the course, with a term GPA of (A) in the third year, as seen in Figure 5.10.

Depending on the student's goal, many prediction rules for different condition and decision attributes can be executed through NewRSGUI program in the same manner. This chapter has described why and how NewRSGUI has been implemented. A student could use NewRSGUI to predict what courses he/she can take and in what year to get a high grade point average GPA. An application such as NewRSGUI can give advice to students as to in what course it is easy to get a high mark and what courses taken together in a term lead to getting a high GPA based on graduated students records in the Math and Computer Science program.

Chapter 6

6 Evaluation

In this chapter, two experiments have been done, one upon the King Khaled General Hospital administrator and another on Laurentian University students. The Laurentian students in this evaluation were not the same ones as referred to in Chapter 5. In Chapter 5 students' records were examined not the actual students. I have done experiments on the hospital's administrator and Laurentian University students to see to which degree the students' opinions differed from the hospital's evaluation. Figure 6.1 shows the degree of departure from both evaluations.

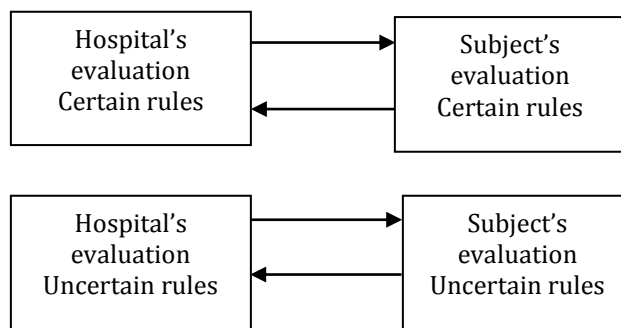


Figure 6.1: Degree of departure from the hospital evaluation

6.1 Occupancy decision attribute evaluation rules

RSGUI has generated 39 rules from the FMW data with REMAINING, ADMITTED, DISCHARGED, and DEATH as condition attributes and occupancy as a decision attribute. Some rules were removed because they were not interesting as evaluated by the hospital and because coverage was low as computed by RS1 algorithm. In the end, we obtained 17 rules to evaluate. The first five rules predict busy ward, middle seven rules predict normal ward, and last five rules predict non-busy ward. The 17 rules are as follow:

Occupancy Decision Rules:

1. (remain := LARGE) AND (discharged := LOW) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
2. (remain := LARGE) AND (discharged := NORMAL) AND (admitted := LARGE) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
3. (remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := NONE) ---> (occupancy := BUSY)[certainty = 1.0][coverage = 2/89]

4. (remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := 1) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 1/89]
5. (remain := NORMAL) AND (discharged := LOW) AND (admitted := NORMAL) AND (death := NONE) -----> (occupancy := BUSY)[certainty = 0.2727272727272727][coverage = 11/89]
6. (discharged := LARGE) AND (death := NONE) AND (remain := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/89]
7. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 5/89]
8. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 3/89]
9. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.9333333333333333][coverage = 15/89]
10. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.7272727272727273][coverage = 11/89]
11. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.5][coverage = 6/89]
12. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LOW) -----> (occupancy := NORMAL)[certainty = 0.8333333333333334][coverage = 6/89]
13. (admitted := LOW) AND (remain := LOW) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 5/89]
14. (admitted := LOW) AND (remain := NORMAL) AND (discharged := LARGE) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 2/89]
15. (admitted := NORMAL) AND (remain := NORMAL) AND (discharged := LARGE) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]
16. (admitted := LOW) AND (remain := NORMAL) AND (discharged := NORMAL) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]
17. (admitted := NORMAL) AND (remain := LOW) AND (discharged := LOW) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.5][coverage = 6/89]

In order to get the best evaluation of the outcomes, we asked one of King Khaled General Hospital administrators to evaluate the rules from most interesting to least interesting, according to the interestingness criteria.

Evaluation Tools:

The 17 rules were sent to the hospital administrator and the subjects, and they were also presented with an empty evaluation table as shown in Table B1 in Appendix B. A completed form is shown next in Table 6.1. The table gives a description of the evaluation criteria used by both hospital administrator and subjects. At the top of the rules, there is a description, in Arabic

and in English, of the ward domain and the fields named in the rules, in order to allow a better evaluation English description

“Below rules are generated from a Female Medical Ward data using the RSGUI program. The ward contains the remaining patients’ statistical data of the previous day and daily statistical data for the newly admitted, discharged, and death patients. After looking at the conditions attributes (remain, new admitted, discharged, and death), could you agree with the decision of busyness of the ward by evaluating the rule to check its criteria on Table B1 please? You are allowed to check more than one criterion on every rule. At least one criterion must be checked for every rule though”.

Hospital administrator evaluation results for above output, which is not order from most interesting to least interesting yet, for occupancy decision attribute shown on Table 6.1.

Table 6.1: Hospital administrator evaluations for occupancy decision attribute

Rule#	Reliability	Utility	Conciseness	Generality	Novelty	Applicability	Diversity	Surprisingness	Peculiarity	Sum of the rule weight
1	√	√	√	√		√	√			7
2		√		√		√				3
3	√	√				√	√			4
4	√	√				√	√			4
5		√		√				√		1
6			√		√				√	-2
7				√	√			√		-2
8					√			√		-3
9	√	√		√		√				7
10				√				√		-2
11		√		√		√				3
12		√		√		√				3
13	√	√	√	√		√				9
14		√	√			√	√			2
15		√			√	√				2
16		√				√				2
17				√	√			√	√	-6

Each criterion was associated with a weight as described in chapter 4 to reflect the importance of that particular criterion. Please recall that the weights range from -4 to +4. When the subject checked the box under a criterion the weight associated with it was accumulated in a running sum. The final sum gives the overall evaluation of the rule. For example, for the first row, the

reliability, utility, conciseness, generality, applicability, and diversity have been checked. The final sum of the first row is computed as follows: $4 + 3 + 2 + 1 - 1 - 2$ respectively. The final sum for these numbers is 7. This process was applied to all 17 rules to come up with the results on the last right column on Table 6.1.

The evaluation also has been done upon 24 Laurentian University students, 4 of them are studying in the medical school at Laurentian University Figure 6.2 is an example of the degree of departure of subject 1 evaluation from the hospital evaluation.

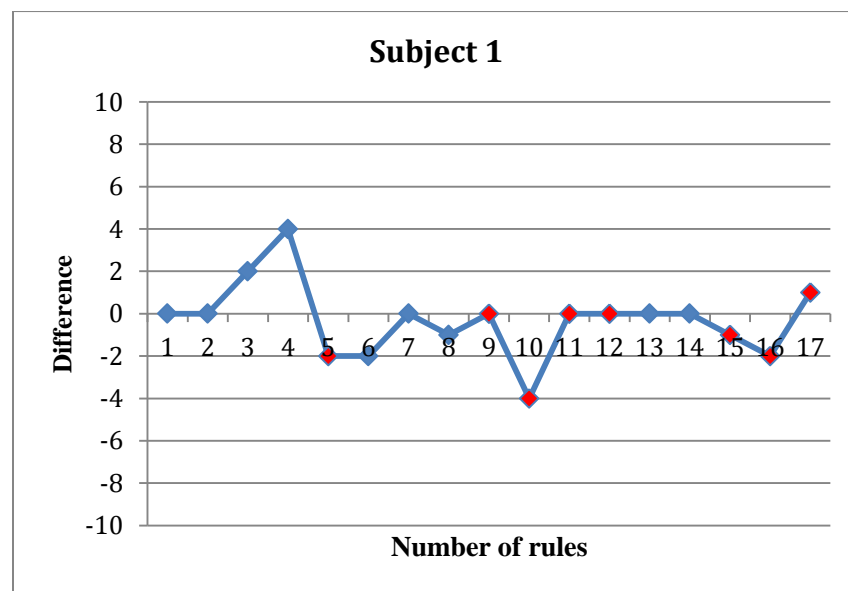


Figure 6.2: The difference between evaluations for positive and negative sides



Figure 6.3: The difference between subject evaluation and hospital evaluation

The points marked in red (5, 9, 10, 11, 12, 15, 16, and 17) are uncertain rules. Once the difference approaches to 0, it means the subject's evaluation is similar or identical to the hospital administrator's evaluation, no matter if the point on negative side or the positive side as illustrate in Figure 6.2. Therefore, Figure 6.3 is the best curve to avoid negative values.

The difference is computed as follows:

$$|a| - |b| = d \quad (\text{figure 6.2})$$

$$|a| - |b| = |d| \quad (\text{figure 6.3})$$

where **a** is the sum of hospital administrator evaluation for one rule, **b** is the sum of subject evaluation for same that rule, and **d** is the difference between both evaluations. Once it is no longer important if the difference positive or negative, we take the absolute value of the **d** = **|d|**. Only a few subject evaluations were ignored, because their differences were close to 10 and far away from 0.

Order of the result after the evaluation from the most interesting, which is the higher number, to least interesting, which is the smallest number, according to the sum of the rule weight on Table 6.1 would be as follows:

1. (admitted := LOW) AND (remain := LOW) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 5/89]
2. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.9333333333333333][coverage = 15/89]
3. (remain := LARGE) AND (discharged := LOW) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
4. (remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := NONE) ---> (occupancy := BUSY)[certainty = 1.0][coverage = 2/89]
5. (remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := 1) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 1/89]
6. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LOW) -----> (occupancy := NORMAL)[certainty = 0.8333333333333334][coverage = 6/89]
7. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.5][coverage = 6/89]
8. (remain := LARGE) AND (discharged := NORMAL) AND (admitted := LARGE) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
9. (admitted := LOW) AND (remain := NORMAL) AND (discharged := LARGE) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 2/89]
10. (admitted := LOW) AND (remain := NORMAL) AND (discharged := NORMAL) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]

11. (admitted := NORMAL) AND (remain := NORMAL) AND (discharged := LARGE) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]
12. (remain := NORMAL) AND (discharged := LOW) AND (admitted := NORMAL) AND (death := NONE) -----> (occupancy := BUSY)[certainty = 0.2727272727272727][coverage = 11/89]
13. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 5/89]
14. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.7272727272727273][coverage = 11/89]
15. (discharged := LARGE) AND (death := NONE) AND (remain := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/89]
16. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 3/89]
17. (admitted := NORMAL) AND (remain := LOW) AND (discharged := LOW) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.5][coverage = 6/89]

In regards to the results of most interesting to least interesting, we can see that a low number of discharges may lead to a busy ward. Therefore, keeping patients in the ward for a long time unnecessarily can cause problems. This is a critical finding with information obtained from King Khaled General Hospital that some private hospitals do not discharge their patients when their condition improves or they no longer need medical care. They even encourage patients to extend their stay, as it is a financial benefit for the hospitals.

The 17 rules above are useful and right, because they were given to the hospital and the Ministry to advise on what can cause the busy ward. Therefore, they can take care of the issues that the rules considered.

Rule Evaluation

We assume that the hospital administrator's evaluation of the certain rules is correct. We subtract the subject's evaluation of certain rules from hospital administrator's evaluation of certain rules (1, 2, 3, 4, 6, 7, 8, 13, and 14). So the difference between hospital evaluation and subject's evaluation of certain rules is the error. There are 9 certain rules and adding the difference values on Figure 6.3, which are (0+0+2+4+2+0+1+0+0 = 9). For subject 1 the average difference across certain rules is $9/9=1$. For uncertain rules we do not want to give a weight of 1 to the differences. Instead, the weight should be the believability of the rule < 1 . For example, if the rule is only 50% believable and the error in student evaluation of the rule is 1 then believability of the

difference should be only 0.5. If the rule is uncertain then the error in the rule will be uncertain. The average error of uncertain rules is:

$$E(X) = \frac{x_5b_5 + x_9b_9 + x_{10}b_{10} + x_{11}b_{11} + x_{12}b_{12} + x_{15}b_{15} + x_{16}b_{16} + x_{17}b_{17}}{u_t}$$

x_i is the difference of the i^{th} rule student evaluation from hospital administrator i^{th} rule evaluation and b_i is the believability of the i^{th} rule. The u_t is the total numbers of uncertain rules.

The above formula can be used as follows:

$$x_5b_5 = 2 * 0.2727 = 0.5454$$

$$x_9b_9 = 0 * 0.9333 = 0$$

$$x_{10}b_{10} = 4 * 0.7272 = 2.909$$

$$x_{11}b_{11} = 0 * 0.5 = 0$$

$$x_{12}b_{12} = 0 * 0.8333 = 0$$

$$x_{15}b_{15} = 1 * 0.6666 = 0.6666$$

$$x_{16}b_{16} = 2 * 0.6666 = 1.3333$$

$$x_{17}b_{17} = 1 * 0.5 = 0.5$$

$$E(X) = \frac{x_5b_5 + x_9b_9 + x_{10}b_{10} + x_{11}b_{11} + x_{12}b_{12} + x_{15}b_{15} + x_{16}b_{16} + x_{17}b_{17}}{u_t}$$

$$\begin{aligned} E(X) &= \frac{0.5454 + 0 + 2.909 + 0 + 0 + 0.6666 + 1.333 + 0.5}{8} \\ &= 0.744 \end{aligned}$$

So since the $E(X) = 0.744 < 1$ we conclude that there is no error in the Subject1's evaluation of the rule. Error introduced by using students is 1 for certain rules. As long as error for uncertain rules < 1 , bias introduced by using students rather than hospital administrators is insignificant.

6.2 Death decision attribute rules evaluation

This subsection illustrates the same process as the Subsection 6.1 but with a different decision attribute. The rules that RS1 generated were 37 and some were removed because they were not interesting according by the hospital and the coverage was low as computed by the RS1 algorithm. In the end, we obtained 15 rules to evaluate. The first three rules predict 2 deaths in

the ward, the middle five rules predict 1 death in the ward, and last seven rules predict no deaths in the ward. The 15 death decision rules are as follows:

1. (remain := LARGE) AND (admitted := NORMAL) AND (occupancy := NOT_BUSY) ---> (death := 2)[certainty = 1.0][coverage = 1/89]
2. (remain := NORMAL) AND (admitted := LARGE) AND (occupancy := NORMAL) AND (discharged := LOW) ---> (death := 2)[certainty = 1.0][coverage = 1/89]
3. (remain := LARGE) AND (admitted := LOW) AND (occupancy := NORMAL) AND (discharged := NORMAL) ---> (death := 2)[certainty = 0.25][coverage = 4/89]
4. (discharged := LARGE) AND (admitted := NORMAL) AND (occupancy := BUSY) --->(death := 1)[certainty = 1.0][coverage = 1/89]
5. (discharged := NORMAL) AND (admitted := LARGE) AND (occupancy := BUSY) AND (remain := LARGE) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/89]
6. (discharged := NORMAL) AND (admitted := LOW) AND (occupancy := NORMAL) AND (remain := LARGE) ---> (death := 1)[certainty = 0.25][coverage = 4/89]
7. (discharged := LOW) AND (admitted := LARGE) AND (occupancy := BUSY) AND (remain := NORMAL) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/89]
8. (discharged := LOW) AND (admitted := NORMAL) AND (occupancy := NOT_BUSY) AND (remain := LOW) ---> (death := 1)[certainty = 0.25][coverage = 4/89]
9. (discharged := LOW) AND (admitted := LOW) ---> (death := NONE)[certainty = 1.0][coverage = 12/89]
10. (discharged := NORMAL) AND (admitted := LARGE) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 5/89]
11. (discharged := NORMAL) AND (admitted := NORMAL) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 15/89]
12. (discharged := NORMAL) AND (admitted := LARGE) AND (remain := LARGE) AND (occupancy := BUSY) ---> (death := NONE)[certainty = 0.6666666666666666][coverage = 3/89]
13. (discharged := LOW) AND (admitted := NORMAL) AND (remain := NORMAL) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.8888888888888888][coverage = 9/89]
14. (discharged := LOW) AND (admitted := NORMAL) AND (remain := LOW) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.75][coverage = 4/89]
15. (discharged := LOW) AND (admitted := NORMAL) AND (remain := LOW) AND (occupancy := NOT_BUSY) ---> (death := NONE)[certainty = 0.75][coverage = 4/89]

Hospital administrator evaluation result for above output, which is not ordered from most interesting to least interesting yet, for death decision attribute is shown on Table 6.2.

Table 6.2: Hospital administrator evaluations for death decision attribute

Rule#	Reliability	Utility	Conciseness	Generality	Novelty	Applicability	Diversity	Surprisingness	Peculiarity	Sum of the rule weight
1			√				√	√	√	-7
2					√			√		-3
3		√		√			√			2
4		√	√			√				4
5	√	√				√				6
6		√				√				2
7	√	√				√	√			4
8				√	√					1
9		√	√	√	√					6
10		√	√	√						6
11	√	√	√	√		√				9
12						√		√		-4
13		√		√						4
14		√		√						4
15	√	√		√						8

Figure 6.4 provides an example of the degree of departure of subject 1 evaluation from the hospital evaluation.

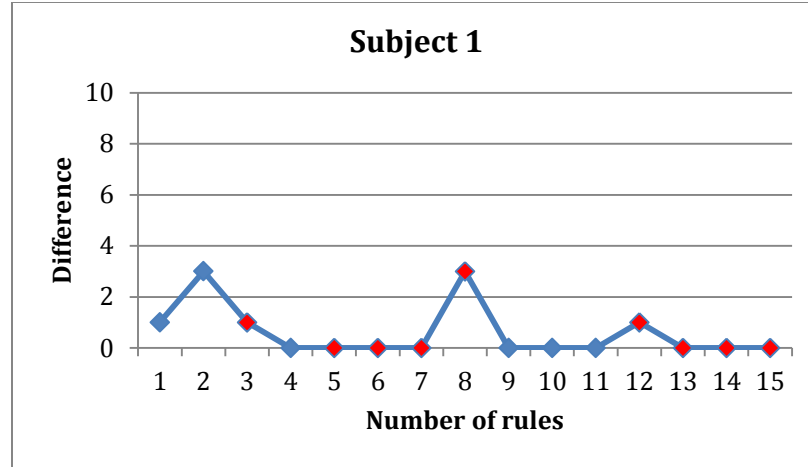


Figure 6.4: The difference of death decision attribute between evaluations for positive and negative sides

The points marked in red (3, 5, 6, 7, 8, 12, 13, 14, and 15) are uncertain rules. The death prediction rules were ordered from most interesting to least interesting, according to the weight criteria from the hospital administration evaluation in Table 6.2 is illustrated below:

Decision rules 7:

1. (discharged := NORMAL) AND (admitted := NORMAL) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 15/89]
2. (discharged := LOW) AND (admitted := NORMAL) AND (remain := LOW) AND (occupancy := NOT_BUSY) ---> (death := NONE)[certainty = 0.75][coverage = 4/89]
3. (discharged := LOW) AND (admitted := LOW) ---> (death := NONE)[certainty = 1.0][coverage = 12/89]
4. (discharged := NORMAL) AND (admitted := LARGE) AND (occupancy := BUSY) AND (remain := LARGE) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/89]
5. (discharged := NORMAL) AND (admitted := LARGE) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 5/89]
6. (discharged := LARGE) AND (admitted := NORMAL) AND (occupancy := BUSY) --->(death := 1)[certainty = 1.0][coverage = 1/89]
7. (discharged := LOW) AND (admitted := LARGE) AND (occupancy := BUSY) AND (remain := NORMAL) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/89]
8. (discharged := LOW) AND (admitted := NORMAL) AND (remain := NORMAL) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.8888888888888888][coverage = 9/89]
9. (discharged := LOW) AND (admitted := NORMAL) AND (remain := LOW) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.75][coverage = 4/89]
10. (remain := LARGE) AND (admitted := LOW) AND (occupancy := NORMAL) AND (discharged := NORMAL) ---> (death := 2)[certainty = 0.25][coverage = 4/89]
11. (discharged := NORMAL) AND (admitted := LOW) AND (occupancy := NORMAL) AND (remain := LARGE) ---> (death := 1)[certainty = 0.25][coverage = 4/89]
12. (discharged := LOW) AND (admitted := NORMAL) AND (occupancy := NOT_BUSY) AND (remain := LOW) ---> (death := 1)[certainty = 0.25][coverage = 4/89]

13. (remain := NORMAL) AND (admitted := LARGE) AND (occupancy := NORMAL) AND (discharged := LOW) ---> (death := 2)[certainty = 1.0][coverage = 1/89]

14. (discharged := NORMAL) AND (admitted := LARGE) AND (remain := LARGE) AND (occupancy := BUSY) ---> (death := NONE)[certainty = 0.6666666666666666][coverage = 3/89]

15. (remain := LARGE) AND (admitted := NORMAL) AND (occupancy := NOT_BUSY) ---> (death := 2)[certainty = 1.0][coverage = 1/89]

To summarize this chapter, the best pattern describing the situation of the busyness from the output of occupancy decision attribute is as follow:

#5 if remain admitted patients = normal, discharge = low, new admitted = large, and death = 1, then the ward will be busy.

On the other hand, the best pattern describing the situation of the death in the ward from the output of death decision attribute is as follows:

#4 if remain admitted patients = large, discharge = normal, new admitted = large, and occupancy = busy, then there is 1 death.

In this chapter, knowledge discovered in health information has been suggested to help improve the quality of healthcare delivery and the effectiveness of healthcare managers. An innovative solution has been developed based on an analytic and knowledge management approach to decision making. After analyzing the rules predicting both occupancy and death decision attributes and ordering them in descending order of interestingness, we have made the case that the top most interesting rules shown in the output are useful for the hospital to report to the Health Ministry. Therefore, the Health Ministry could solve the problem of busy ward or death on a ward by expending or hiring more staff to cover some of the inability the ward has.

Chapter 7

7 Conclusion

Upon examination of the data about the activity (admissions, discharges, remaining patients, and deaths) in medical wards, a number of questions come to mind. For example, what conditions of the ward tend to increase patient deaths? This project was an experiment to find out how to use data mining tools based on rough set theory to uncover hidden relationships in medical wards data. The objective was to approximate a given concept such as “Hospital ward busy” or “Death in a ward”, in terms of information about the ward. The power of rough set theory lies in its ability to suppress information that is not essential leaving the minimum set of attributes needed to predict interesting outcomes. The method can be applied in reverse as a mean of understanding essential factors that lead to a desired outcome.

Usually “real” data are not available for dissemination of analysis, and, therefore, the rough set theory has mainly been verified with simulated data. Methodical experimentation with RSGUI on the medical ward data obtained from King Khaled General Hospital has resulted in the selection of a set of interesting rules as measured by parameters to describe the rules. Interestingness measures were tested on a medical domain to help validate their usefulness.

By looking at the outputs of RS1 when occupancy was selected to be a decision attribute, we can see that LOW value of the discharges may cause BUSY ward, while LARGE value of discharge gives NOT BUSY ward. For that, doctors should not extend patients stay if the patients are getting better and their stays are unnecessary. By comparing the outputs of RS1 when death was selected as a decision attribute, we can see that if the value of new admitted is LARGE and the value of discharge is LOW, the ward most likely will have a death. The ward is going to be busy and staff will preoccupy with some patients and the others will not have that care. Though this rule is obvious, it is interesting due to the critical nature of the outcome. Therefore, hiring more staff and adding more beds could solve the ward busyness and result in improving patient safety and satisfaction.

Many applications use commercially available rough sets based tools that cannot be modified, RSGUI allows a programmer to modify its code, because it has been developed by Laurentian University students over the years and it is open for further development. A new version of RSGUI has been embedded in a web interface for accessing and analyzing medical ward data.

A comparison of RS1 and ILA on medical ward data has been made by counting number of rules generated and inspecting length of the rules. It has been concluded that ILA is superior for making recommendations to the Ministry of Health.

An instrument (questionnaire) was developed to collect information from subjects to help quantify the notion of interestingness of predictive rules. A limitation regarding the availability of suitable subjects for the study was mitigated by introducing a measure of error of available subject responses from suitable subject responses. The interestingness measure was applied to adjust available responses to a better correspond with suitable subject responses.

NewRSGUI features exactly the same functions as RSGUI. So the rules that RSGUI would have generated, would be the same as results that NewRSGUI is going to have, because as mentioned in Subsection 5.2 the only differences between both applications are the start up of the applications and the databases. RSGUI has been used in the medical office data not with the second data, because when it was decided to use NewRSGUI in the thesis, the results of the first data have been made and concluded. Therefore, using either of the two applications for the first or the second data are going to make same results.

7.1 Major contributions and future work

The main contributions of the thesis to ongoing research are as follows:

- Improvement of an existing experimental tool based on theories in the rough set method
- The development of a new rule evaluation technique based on the idea of interestingness of rules.
- Illustration of the fact that Inductive Learning Algorithm is superior to the RS1 algorithm in terms of number of rules generated.

- Investigation of errors introduced by using subjects that may not be a representative sample of the eventual users of the rules, and illustration by example of how such errors can be quantified.
- Determination that there is a degree of invariance in the ordering of rules across evaluators, especially for uncertain rules, by observation that the difference between hospital administrator and available subjects evaluation remained low enough so as to leave the ordering of rules undisturbed.
- Acquisition of real data that is typically unavailable due to its sensitive nature and methodical investigation of patterns implicit in that data.

Further research is expected, particularly with regard to the measurement of errors introduced by using non-representative subjects for evaluating interestingness of rules. It was observed that an almost identical set of rules was generated by ILA as was decided by the hospital administrator independent of any knowledge discovery tools. It appears that ILA is employing the same knowledge as human subjects used when making a first cut to eliminate uninteresting rules. A challenging problem for further research is to use the formalization provided by the ILA algorithm to better understand what intelligence is being employed to order rules.

Publication

Anwar Alenezi, Julia Johnson (2014) Finding Patterns in Medical Ward Data Using Rough Sets, Proceedings of International Conference on Analytics Driven Solutions (ICAS 2014), pp. 135-144, Ottawa, Canada, September, 2014.

References

- [1] R. N. Saxena and A. Srinivasan, *Business analytics: a practitioner's guide*. New York: International Series in Operations Research & Management Science, 2013.
- [2] J. N. . Gupta, G. A. Forgionne, and M. Mora T., *Intelligent Decision-Making Support Systems: Foundations, Applications, and Challenges*. London: Springer, 2006.
- [3] P. Campeau, "Predicting the Most Favorable Behavior of Artificial Objects using Rough Sets," Honors thesis, Laurentian University, Sudbury, Canada, 2005.
- [4] J. A. Johnson and G. M. Johnson, "RSGUI with Reverse Prediction Algorithm," in *Bello, R., Falcan, R., Pedrycz, W., & Kacprzyk, J. (Eds.), Studies in Fuzziness and Soft Computing, Vol. 224*, 2008, pp. 287–306.
- [5] H. Zuo and J. Johnson, "A Self-learning Audio Player That Uses a Rough Set and Neural Net Hybrid Approach," *RSKT 2013*, pp. 405–412, 2013.
- [6] J. W. Grzymala-busse, "A Local Version of the MLEM2 Algorithm for Rule Induction 2 . Blocks of Attribute-Value Pairs," *Fundam. Informaticae*, vol. 100, pp. 99–116, 2010.
- [7] M. Tolun and S. Abu-Soud, *An Inductive Learning Algorithm for Production Rule Discovery*. Available online.
- [8] J. Johnson and P. Campeau, "Reverse Prediction," in *10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing RSFDGrC2005, Part II: In: J. G. Carbonell & J. Siekmann (Eds.) Lecture Notes in Computer Science, Vol. 3642*, 2005, pp. 88–97.
- [9] S. M. Crow, S. J. Hartman, E. N. Brockmann, and S. W. Henson, "The educational needs of healthcare managers and executives in the key strategic areas of healthcare.," *Hosp. Top.*, vol. 80, no. 3, pp. 16–20, Jan. 2002.
- [10] U. Srinivasan and B. Arunasalam, "Leveraging Big Data Analytics to Reduce Healthcare Costs," *IT Prof.*, vol. 15, no. 6, pp. 21–28, 2013.
- [11] X. Zhou, "Rough Set-based Reasoning and Pattern Mining for Information Filtering," Faculty of Science and Technology, Queensland University of Technology, Australia, 2008.
- [12] S. Rissino and G. Lambert-torres, "Rough Set Theory – Fundamental Concepts , Principals , Data Extraction , and Applications," *Data Min. Knowl. Discov. Real Life Appl.*, no. February, pp. 35–58, 2009.

- [13] A. Liang, B. Maguire, and J. A. Johnson, "Rough Set Based WebCT Learning," *H. Lu A. Zhou (Eds.), Lect. Notes Comput. Sci. Springer*, vol. 1846, pp. 425–436, 2000.
- [14] S. Tsumoto and S. Hirano, "Rough-Set based Criteria for Incremental Rule Induction," *Int. J. Hybrid Inf. Technol.*, vol. 5, no. 2, pp. 249–254, 2012.
- [15] I. B. Türkşen, "Belief, plausibility, and probability measures on interval-valued type 2 fuzzy sets," *Int. J. Intell. Syst.*, vol. 19, no. 7, pp. 681–699, Jul. 2004.
- [16] W.-Z. Wu, "A comparative study of belief and plausibility reducís in information systems with fuzzy decisions," *Mach. Learn. Cybern. (ICMLC), 2010 Int. Conf.*, vol. 2, pp. 552–557, 2010.
- [17] C. Joslyn and J. C. Helton, "Bounds on belief and plausibility of functionally propagated random sets," *Fuzzy Inf. Process. Soc. 2002. Proceedings. NAFIPS. 2002 Annu. Meet. North Am.*, pp. 412–417, 2002.
- [18] Z. Pawlak, "Rough Sets and Decision Algorithms," *Second Int. Conferance Rough Sets Curr. Trends Comput. RSCT'2000, Banff, Canada*, no. October 16–19, 2000.
- [19] H. Liu, H. Pan, and A. Wang, "Rough Sets Algorithm and its Application in Fault Diagnosis," *TELKOMNIKA*, vol. 11, no. 9, pp. 5471–5479, 2013.
- [20] J. Wang, "A Rough Set Approach of Mechanical Fault Diagnosis for Five-Plunger Pump," *Adv. Mech. Eng.*, vol. 2013, pp. 1–11, 2013.
- [21] Z. Geng and Q. Zhu, "Rough Set-Based Fuzzy Rule Acquisition and Its Application for Fault Diagnosis in Petrochemical Process," *Ind. Eng. Chem. Res.*, vol. 48, no. 2, pp. 827–836, Jan. 2009.
- [22] B. K. Tripathy, D. P. Acharjya, and V. Cynthia, "A Framework for Intelligent Medical Diagnosis Using Rough Set With Formal Concept Analysis," *Int. J. Artif. Intell. Appl.*, vol. 2, no. 2, 2011.
- [23] G. Ilczuk and A. Wakulicz-deja, "Rough Sets Approach to Medical Diagnosis System," in *Lecture Notes in Computer Science*, 2005, vol. 3528, pp. 204–210.
- [24] Y. Zhao, F. Luo, S. K. M. Wong, and Y. Y. Yao, "A General Definition of an Attribute Reduct," *Rough Sets Knowl. Technol. Second Int. Conf.*, vol. 4481, pp. 101–108, 2007.
- [25] A. E. Hassanien, A. Abraham, J. F. Peters, and G. Schaefer, "Rough Sets in Medical Informatics Applications," *Appl. Soft Comput.*, vol. 58, pp. 23–30, 2009.
- [26] W. Michalowski, S. Rubin, R. Slowinski, and S. Wilk, "Triage of the child with abdominal pain: A clinical algorithm for emergency patient management.," *Paediatr. Child Health*, vol. 6, no. 1, pp. 23–8, Jan. 2001.

- [27] P. Paszek and A. W. Deja, "Applying Rough Set Theory to Medical Diagnosing," in *Rough Sets and Intelligent Systems Paradigms - RSEISP*, 2007, pp. 427–435.
- [28] S. Tsumoto and S. Hirano, "Incremental Induction of Medical Diagnostic Rules Based on Incremental Sampling Scheme and SubRule Layers," *Fundam. Inform.*, vol. 127, no. 1–4, pp. 209–223, 2013.
- [29] J. Zaluski, R. Szoszkiewicz, J. Krysiński, and J. Stefanowski, "Rough Set Theory and Decision Rules in Data Analysis of Breast Cancer Patients," *Trans. Rough Sets I*, pp. 375–391, 2004.
- [30] J. L. Breault, C. R. Goodall, and P. J. Fos, "Data mining a diabetic data warehouse," *Artif. Intell. Med. Med. Data Min. Knowl. Discov.*, vol. 26, no. 1–2, pp. 37–54, 2002.
- [31] P. Pattaraintakorn and N. Cercone, "Integrating rough set theory and medical applications," *Appl. Math. Lett.*, vol. 21, no. 4, pp. 400–403, Apr. 2008.
- [32] M. Przybyla-kasperek and A. Wakulicz-deja, "Global Decisions Taking on the Basis of Dispersed Medical Data," *Rough Sets, Fuzzy Sets, Data Mining, Granul. Comput. RSFSGrC2013*, pp. 355–365, 2013.
- [33] J. W. Grzymala-Busse, "An Empirical Comparison of Rule Induction Using Feature Selection with the LEM2 Algorithm," *14th Int. Conf. Inf. Process. Manag. Uncertain. Knowledge-Based Syst. IPMU 2012*, pp. 270–279, 2012.
- [34] S. Tsumoto, "Rough Sets and Medical Differential Diagnosis," *Rough Sets Intell. Syst.*, vol. 42, pp. 605–621, 2013.
- [35] A. Abed-Elmdoust and R. Kerachian, "Wave height prediction using the rough set theory," *Ocean Eng.*, vol. 54, pp. 244–250, Nov. 2012.
- [36] Y. Cheng, D. Miao, and Q. Feng, "Positive approximation and converse approximation in interval-valued fuzzy rough sets," *Inf. Sci. (Ny)*, vol. 181, no. 11, pp. 2086–2110, Jun. 2011.
- [37] T. Tseng, C. Huang, and Y. Fan, "Autonomous rule induction from data with tolerances in customer relationship management," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 4889–4900, May 2011.
- [38] C. Huang, T. Tseng, Y. Fan, and C. Hsu, "Alternative rule induction methods based on incremental object using rough set theory," *Appl. Soft Comput.*, vol. 13, no. 1, pp. 372–389, Jan. 2013.
- [39] P. Yıldırım and M. R. Tolun, "Induction for Radiology Patients," *Lect. Notes Inst. Comput. Sci. Soc. Informatics Telecommun. Eng.*, pp. 208–220, 2009.

- [40] O. Akgobek, Y. S. Aydin, E. Oztemel, and M. S. Aksoy, "A new algorithm for automatic knowledge acquisition in inductive learning," *Knowledge-Based Syst.*, vol. 19, no. 6, pp. 388–395, 2006.
- [41] M. R. Tolun, "ILA-2: an Inductive Learning Algorithm for Knowledge Discovery," *Cybern. Syst.*, vol. 30, no. 7, pp. 609–628, Oct. 1999.
- [42] M. R. Tolun, H. Sever, and M. Uludağ, "Improved Rule Discovery Performance on Uncertainty," *Lect. Notes Comput. Sci.*, vol. 1394, pp. 310–321, 1998.
- [43] G. de Niet, B. Tiemens, T. van Achterberg, and G. Hutschemaekers, "Applicability of two brief evidence-based interventions to improve sleep quality in inpatient mental health care," *Int. J. Ment. Health Nurs.*, vol. 20, no. 5, pp. 319–327, Oct. 2011.
- [44] C. Hopkins, "'But what about the really ill, poorly people?' (An ethnographic study into what it means to nurses on medical admissions units to have people who have harmed themselves as their patients).," *J. Psychiatr. Ment. Health Nurs.*, vol. 9, no. 2, pp. 147–154, Apr. 2002.
- [45] S. Kripalani, F. Lefevre, C. O. Phillips, M. V Williams, P. Basaviah, and D. W. Baker, "Deficits in Communication and Information Transfer Between Hospital-Based and Primary Care Physicians," *JAMA*, vol. 297, no. 8, pp. 831–841, 2007.
- [46] J. A. Martinez, A. Belastegui, I. Basabe, X. Goicoechea, C. Aguirre, N. Lizeaga, I. Urreta, and J. I. Emparanza, "Derivation and validation of a clinical prediction rule for delirium in patients admitted to a medical ward: an observational study," *Br. Med. J.*, vol. 2, pp. 1–7, Jan. 2012.
- [47] L. Grammatico, S. Baron, E. Rusch, B. Lepage, N. Surer, J. C. Desenclos, and J. M. Besnier, "Epidemiology of vertebral osteomyelitis (VO) in France: analysis of hospital-discharge data 2002-2003," *Epidemiol. Infect.*, vol. 136, no. 5, pp. 653–660, May 2008.
- [48] T. Alexander, G. Fuller, P. Hargovan, D. Clarke, D. Muckart, and S. Thomson, "An audit of the quality of care of traumatic brain injury at a busy regional hospital in South Africa," *South African J. Surg.*, vol. 47, no. 4, pp. 120–124, 2009.
- [49] J. Wilson, "Ward staff experiences of patient death in an acute medical setting," *Nurs. Stand. (Royal Coll. Nurs. (Great Britain))*, vol. 28, no. 37, pp. 37–45, 2014.
- [50] L. Polkowski, S. Tsumoto, and T. Y. Lin, *Rough set methods and applications. New developments in knowledge discovery in information systems series: Studies in fuzziness and soft computing*. New York: Physica-Verlag Heidelberg, 2000.
- [51] Z. Shi, *Advanced artificial intelligence*. London: World Scientific, 2011.

- [52] W.-Z. Wu, Y. Leung, and W.-X. Zhang, "Connections between rough set theory and Dempster-Shafer theory of evidence," *Int. J. Gen. Syst.*, vol. 31, no. 4, pp. 405–430, Jul. 2002.
- [53] Z. Pawlak, "Rough sets and intelligent data analysis," *Inf. Sci. (Ny)*, vol. 147, no. 1–4, pp. 1–12, Nov. 2002.
- [54] T. Connolly and C. Begg, *Database Systems: A Practical Approach to Design, Implementation and Management*, 5th ed. 2009, pp. 1–1242.
- [55] S. Kotsiantis and D. Kanellopoulos, "Discretization Techniques : A recent survey," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 32, no. 1, pp. 47–58, 2006.
- [56] Y. Yang and G. I. Webb, "A Comparative Study of Discretization Methods for Naive-Bayes Classifiers," *Proc. PKAW*, pp. 159–173, 2002.
- [57] C. and M. L. and Z. H. and Y. Z. and Y. Y. Wu, "Discretization Algorithms of Rough Sets Using Clustering," *Robot. Biomimetics, 2004. ROBIO 2004. IEEE Int. Conf.*, pp. 955–960, 2004.
- [58] I. H. W. and E. Frank, *Data mining: practical machine learning tools and techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.
- [59] L. Geng and H. J. Hamilton, "Interestingness measures for data mining," *ACM Comput. Surv.*, vol. 38, no. 3, Sep. 2006.
- [60] S. Greco, S. Roman, and I. Szcz, "Properties of rule interestingness measures and alternative approaches to normalization of measures," *Inf. Sci. an Int. J.*, vol. 216, pp. 1–16, 2012.
- [61] P. Ramasubramanian, V. Suresnkumar, P. Iyakutti, and P. Thangavelu, "Mining Analysis of SIS Database Using Rough Set Theory," *Conf. Comput. Intell. Multimed. Appl.*, vol. 2, pp. 81–87, 2007.

Appendix A: Ethics Approvals



APPROVAL FOR CONDUCTING RESEARCH INVOLVING HUMAN SUBJECTS Research Ethics Board – Laurentian University

This letter confirms that the research project identified below has successfully passed the ethics review by the Laurentian University Research Ethics Board (REB). Your ethics approval date, other milestone dates, and any special conditions for your project are indicated below.

TYPE OF APPROVAL / New X / Modifications to project / Time extension	
Name of Principal Investigator and school/department	Anwar Alenzi (Math and Computer Science)
Title of Project	Finding Patterns in Medical Ward Data Using Rough Sets
REB file number	2013-10-15
Date of original approval of project	November 1, 2013
Date of approval of project modifications or extension (<i>if applicable</i>)	
Final/Interim report due on	November 1, 2014
Conditions placed on project	Final report due on November 1, 2014

During the course of your research, no deviations from, or changes to, the protocol, recruitment or consent forms may be initiated without prior written approval from the REB. If you wish to modify your research project, please refer to the Research Ethics website to complete the appropriate [REB form](#).

All projects must submit a report to REB at least once per year. If involvement with human participants continues for longer than one year (e.g. you have not completed the objectives of the study and have not yet terminated contact with the participants, except for feedback of final results to participants), you must request an extension using the appropriate [REB form](#).

In all cases, please ensure that your research complies with [Tri-Council Policy Statement \(TCPS\)](#). Also please quote your REB file number on all future correspondence with the REB office. Congratulations and best of luck in conducting your research.

Susan James, Chair
Laurentian University Research Ethics Board



APPROVAL FOR CONDUCTING RESEARCH INVOLVING HUMAN SUBJECTS
Research Ethics Board – Laurentian University

This letter confirms that the research project identified below has successfully passed the ethics review by the Laurentian University Research Ethics Board (REB). Your ethics approval date, other milestone dates, and any special conditions for your project are indicated below.

TYPE OF APPROVAL / New <input checked="" type="checkbox"/> / Modifications to project / Time extension	
Name of Principal Investigator and school/department	Anwar Alenezi (Math and Computer Science) Julia Johnson (Supervisor)
Title of Project	Rules discovery for advising students on which courses to take
REB file number	2013-11-15
Date of original approval of project	December 4, 2013
Date of approval of project modifications or extension (<i>if applicable</i>)	
Final/Interim report due on	December 4, 2014
Conditions placed on project	Final report due on December 4, 2014

During the course of your research, no deviations from, or changes to, the protocol, recruitment or consent forms may be initiated without prior written approval from the REB. If you wish to modify your research project, please refer to the Research Ethics website to complete the appropriate [REB form](#).

All projects must submit a report to REB at least once per year. If involvement with human participants continues for longer than one year (e.g. you have not completed the objectives of the study and have not yet terminated contact with the participants, except for feedback of final results to participants), you must request an extension using the appropriate [REB form](#).

In all cases, please ensure that your research complies with [Tri-Council Policy Statement \(TCPS\)](#). Also please quote your REB file number on all future correspondence with the REB office.

Congratulations and best of luck in conducting your research.

Susan James, Chair
Laurentian University Research Ethics Board

الرقم : ٥٠٣/٢٧٥/

التاريخ : ١٤ / / ١٤

المشروعات :

Sept. 10th, 2013وزارة الصحة
Ministry of Health

المملكة العربية السعودية
وزارة الصحة
مديرية الشؤون الصحية بمحافظة حفر الباطن
مستشفى الملك خالد العام

Re: Letter of agreement to Laurentian University

King Khaled General Hospital in Hafer Albatin, Saudi Arabia is authorizing Anwar Alenezi to use the documents that have given to him. We have given him statistics data for two months, some of them typed in excel files, and some given hard copy of our manual documents system. So all documents need are given and he is allowed to use them in his research. The data are cleaned of all identifying information.

Do not hesitate to contact me if you have any future questions.

Manager of Computer and Statistics department

Mohammad H. Thany

[Handwritten signature]
٢٠١٣/٩/١٠



Appendix B: Experiment tools

Occupancy Decision Rules 1:

1. (remain := LARGE) AND (discharged := LOW) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
2. (remain := LARGE) AND (discharged := NORMAL) AND (admitted := LARGE) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 3/89]
3. (remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := NONE) ---> (occupancy := BUSY)[certainty = 1.0][coverage = 2/89]
4. (remain := NORMAL) AND (discharged := LOW) AND (admitted := LARGE) AND (death := 1) -----> (occupancy := BUSY)[certainty = 1.0][coverage = 1/89]
5. (remain := NORMAL) AND (discharged := LOW) AND (admitted := NORMAL) AND (death := NONE) -----> (occupancy := BUSY)[certainty = 0.2727272727272727][coverage = 11/89]
6. (discharged := LARGE) AND (death := NONE) AND (remain := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/89]
7. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 5/89]
8. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 3/89]
9. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.9333333333333333][coverage = 15/89]
10. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.7272727272727273][coverage = 11/89]
11. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 0.5][coverage = 6/89]
12. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LOW) -----> (occupancy := NORMAL)[certainty = 0.8333333333333334][coverage = 6/89]
13. (admitted := LOW) AND (remain := LOW) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 5/89]
14. (admitted := LOW) AND (remain := NORMAL) AND (discharged := LARGE) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 2/89]
15. (admitted := NORMAL) AND (remain := NORMAL) AND (discharged := LARGE) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]
16. (admitted := LOW) AND (remain := NORMAL) AND (discharged := NORMAL) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.6666666666666666][coverage = 3/89]
17. (admitted := NORMAL) AND (remain := LOW) AND (discharged := LOW) AND (death := NONE) -----> (occupancy := NOT_BUSY)[certainty = 0.5][coverage = 6/89]

Death Decision Rules 2:

1. (remain := LARGE) AND (admitted := NORMAL) AND (occupancy := NOT_BUSY) ---> (death := 2)[certainty = 1.0][coverage = 1/89]
2. (remain := NORMAL) AND (admitted := LARGE) AND (occupancy := NORMAL) AND (discharged := LOW) ---> (death := 2)[certainty = 1.0][coverage = 1/89]
3. (remain := LARGE) AND (admitted := LOW) AND (occupancy := NORMAL) AND (discharged := NORMAL) ---> (death := 2)[certainty = 0.25][coverage = 4/89]
4. (discharged := LARGE) AND (admitted := NORMAL) AND (occupancy := BUSY) --->(death := 1)[certainty = 1.0][coverage = 1/89]
5. (discharged := NORMAL) AND (admitted := LARGE) AND (occupancy := BUSY) AND (remain := LARGE) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/89]
6. (discharged := NORMAL) AND (admitted := LOW) AND (occupancy := NORMAL) AND (remain := LARGE) ---> (death := 1)[certainty = 0.25][coverage = 4/89]
7. (discharged := LOW) AND (admitted := LARGE) AND (occupancy := BUSY) AND (remain := NORMAL) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/89]
8. (discharged := LOW) AND (admitted := NORMAL) AND (occupancy := NOT_BUSY) AND (remain := LOW) ---> (death := 1)[certainty = 0.25][coverage = 4/89]
9. (discharged := LOW) AND (admitted := LOW) ---> (death := NONE)[certainty = 1.0][coverage = 12/89]
10. (discharged := NORMAL) AND (admitted := LARGE) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 5/89]
11. (discharged := NORMAL) AND (admitted := NORMAL) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 15/89]
12. (discharged := NORMAL) AND (admitted := LARGE) AND (remain := LARGE) AND (occupancy := BUSY) ---> (death := NONE)[certainty = 0.6666666666666666][coverage = 3/89]
13. (discharged := LOW) AND (admitted := NORMAL) AND (remain := NORMAL) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.8888888888888888][coverage = 9/89]
14. (discharged := LOW) AND (admitted := NORMAL) AND (remain := LOW) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.75][coverage = 4/89]
15. (discharged := LOW) AND (admitted := NORMAL) AND (remain := LOW) AND (occupancy := NOT_BUSY) ---> (death := NONE)[certainty = 0.75][coverage = 4/89]

Appendix C: Decision Rules and their evaluation table

Decision Rules 1:

1. (admitted := NORMAL) AND (discharged := NORMAL) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 14/77]
2. (admitted := LOW) AND (discharged := LOW) ---> (death := NONE)[certainty = 1.0][coverage = 11/77]
3. (admitted := NORMAL) AND (discharged := LOW) AND (occupancy := NORMAL) AND (remain := NORMAL) ---> (death := 1)[certainty = 0.1111111111111111][coverage = 9/77]
4. (admitted := LARGE) AND (discharged := NORMAL) AND (occupancy := BUSY) AND (remain := LARGE) ---> (death := 1)[certainty = 0.5][coverage = 4/77]
5. (admitted := NONE) ---> (death := NONE)[certainty = 1.0][coverage = 1/77]
6. (remain := NORMAL) AND (admitted := LARGE) AND (occupancy := NORMAL) AND (discharged := LOW) ---> (death := 2)[certainty = 1.0][coverage = 1/77]
7. (admitted := NORMAL) AND (discharged := LOW) AND (remain := LOW) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 4/77]
8. (remain := LARGE) AND (admitted := NORMAL) AND (occupancy := NOT_BUSY) ---> (death := 2)[certainty = 1.0][coverage = 1/77]
9. (admitted := LARGE) AND (discharged := LOW) AND (occupancy := BUSY) AND (remain := NORMAL) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/77]
10. (admitted := LARGE) AND (discharged := LOW) AND (remain := NORMAL) AND (occupancy := BUSY) ---> (death := NONE)[certainty = 0.6666666666666666][coverage = 3/77]
11. (admitted := LOW) AND (discharged := LARGE) AND (remain := NORMAL) AND (occupancy := NOT_BUSY) ---> (death := NONE)[certainty = 0.5][coverage = 2/77]
12. (admitted := NORMAL) AND (discharged := LOW) AND (remain := NORMAL) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.8888888888888888][coverage = 9/77]
13. (remain := LARGE) AND (admitted := LOW) AND (occupancy := NORMAL) AND (discharged := NORMAL) ---> (death := 2)[certainty = 0.25][coverage = 4/77]
14. (admitted := LOW) AND (discharged := NORMAL) AND (remain := NORMAL) AND (occupancy := NOT_BUSY) ---> (death := NONE)[certainty = 0.6666666666666666][coverage = 3/77]
15. (admitted := LOW) AND (discharged := NORMAL) AND (remain := LARGE) AND (occupancy := NORMAL) ---> (death := NONE)[certainty = 0.5][coverage = 4/77]
16. (admitted := LOW) AND (discharged := NORMAL) AND (occupancy := NORMAL) AND (remain := LARGE) ---> (death := 1)[certainty = 0.25][coverage = 4/77]
17. (admitted := LARGE) AND (discharged := NORMAL) AND (remain := LARGE) AND (occupancy := BUSY) ---> (death := NONE)[certainty = 0.5][coverage = 4/77]
18. (admitted := LARGE) AND (discharged := LOW) AND (remain := LOW) ---> (death := NONE)[certainty = 1.0][coverage = 3/77]
19. (admitted := LOW) AND (discharged := NORMAL) AND (occupancy := NOT_BUSY) AND (remain := NORMAL) ---> (death := 1)[certainty = 0.3333333333333333][coverage = 3/77]

20. (admitted := NORMAL) AND (discharged := NORMAL) AND (occupancy := BUSY) AND (remain := LARGE) ---> (death := 1)[certainty = 0.5][coverage = 2/77] >

21. (admitted := LARGE) AND (discharged := NORMAL) AND (remain := NORMAL) ---> (death := NONE)[certainty = 1.0][coverage = 5/77]

22. (admitted := LOW) AND (discharged := LARGE) AND (occupancy := NOT_BUSY) AND (remain := NORMAL) ---> (death := 1)[certainty = 0.5][coverage = 2/77]

Table C.1: Evaluation by the criteria of the coverage, and the ranking of each pattern

Rule#	Reliability	Utility	Conciseness	Generality	Novelty	Applicability	Diversity	Peculiarity	Surprisingness	Coverage	Interesting Rank
1	√	√	√	√		√				14	9
2	√	√	√	√	√	√				11	9
3	√	√		√		√				9	7
4	√	√		√		√				4	7
5	√	√	√		√				√	1	5
6	√	√			√	√	√			1	4
7		√		√	√	√				4	3
8	√	√	√				√		√	1	3
9		√					√			3	1
10		√		√	√	√	√			3	1
11		√				√	√			2	0
12				√			√			9	-1
13		√		√			√		√	4	-2
14		√				√				3	-2
15				√			√		√	4	-5
16				√				√	√	4	-6
17				√	√			√	√	4	-6
18							√		√	3	-6
19					√			√	√	2	-7
20								√	√	3	-7
21				√			√	√	√	5	-8
22							√	√	√	2	-9

Results found using RS1 when selecting occupancy as a decision attribute with 77 rows and ordered from most interesting to least interesting are as follows:

Decision Rules 2:

1. (admitted := LOW) AND (remain := LOW) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 5/77]

2. (remain := LARGE) AND (admitted := LARGE) --> (occupancy := BUSY)[certainty = 1.0][coverage = 4/77]

3. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 14/77]
4. (admitted := LOW) AND (remain := NORMAL) AND (discharged := NORMAL) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 3/77]
5. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LOW) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 5/77]
6. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 3/77]
7. (admitted := LOW) AND (remain := NORMAL) AND (discharged := LARGE) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 2/77]
8. (remain := LARGE) AND (admitted := NORMAL) AND (discharged := LOW) --> (occupancy := BUSY)[certainty = 1.0][coverage = 2/77]
9. (discharged := LARGE) AND (death := NONE) AND (remain := LARGE) --> (occupancy := NORMAL)[certainty = 1.0][coverage = 1/77]
10. (discharged := LOW) AND (death := NONE) AND (remain := NORMAL) AND (admitted := NORMAL) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 8/77]
11. (discharged := NORMAL) AND (death := NONE) AND (remain := NORMAL) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 5/77]
12. (remain := LARGE) AND (admitted := NORMAL) AND (discharged := NORMAL) --> (occupancy := BUSY)[certainty = 1.0][coverage = 2/77]
13. (admitted := NORMAL) AND (remain := LARGE) AND (discharged := LARGE) AND (death := 2) -----> (occupancy := NOT_BUSY)[certainty = 1.0][coverage = 1/77]
14. (discharged := NONE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 2/77]
15. (discharged := LOW) AND (death := NONE) AND (remain := LOW) AND (admitted := LARGE) -----> (occupancy := NORMAL)[certainty = 1.0][coverage = 3/77]

Table C.2: Evaluation by the criteria of the coverage, and the ranking for each rule above

Rule#	Reliability	Utility	Conciseness	Generality	Novelty	Applicability	Diversity	Peculiarity	Surprisingness	Coverage	Rank
1	√	√	√	√		√				5	9
2	√	√	√	√		√				4	9
3	√	√		√		√				14	7
4	√	√				√				3	6
5		√		√		√				5	3
6		√				√				3	2
7		√			√		√			2	1
8		√			√		√			2	1
9		√				√	√			1	-1
10				√					√	8	-3

11				$\sqrt{\quad}$					$\sqrt{\quad}$	5	-3
12					$\sqrt{\quad}$	$\sqrt{\quad}$	$\sqrt{\quad}$			2	-3
13					$\sqrt{\quad}$				$\sqrt{\quad}$	1	-4
14			$\sqrt{\quad}$		$\sqrt{\quad}$			$\sqrt{\quad}$	$\sqrt{\quad}$	2	-5
15							$\sqrt{\quad}$	$\sqrt{\quad}$	$\sqrt{\quad}$	3	-9

Appendix D: One month data discretized

Table D.1: Female Medical Ward (FMW) data for one month

Day	Rem.	New Adm.	Death	Dis	OCC
1	18	1	0	7	12
2	12	2	0	1	13
3	13	6	0	3	16
4	16	6	0	4	18
5	18	5	0	6	17
6	17	2	0	5	14
7	14	3	0	8	9
8	9	4	0	2	11
9	11	1	0	2	10
10	10	5	0	3	12
11	12	4	1	2	13
12	13	2	0	3	12
13	12	2	0	1	13
14	13	5	0	7	11
15	11	4	0	2	13
16	13	3	0	0	16
17	16	3	0	4	15
18	15	4	0	4	15
19	15	10	0	2	23
20	23	3	1	5	20
21	20	4	0	3	21
22	21	4	0	8	17
23	17	4	0	2	19
24	19	3	0	1	21
25	21	2	1	5	18
26	18	5	0	5	18
27	18	4	0	4	18
28	18	5	0	1	22
29	22	3	2	10	13

Table D.2: Discretized table for Female Medical Ward data for one month

Day	Rem.	New Adm.	Death	Dis	OCC
1	NORMAL	LOW	NONE	LARGE	NOT_BUSY
2	LOW	LOW	NONE	LOW	NOT_BUSY
3	LOW	LARGE	NONE	LOW	NORMAL
4	NORMAL	LARGE	NONE	NORMAL	NORMAL
5	NORMAL	NORMAL	NONE	NORMAL	NORMAL
6	NORMAL	LOW	NONE	NORMAL	NORMAL
7	NORMAL	NORMAL	NONE	LARGE	NOT_BUSY
8	LOW	NORMAL	NONE	LOW	NOT_BUSY
9	LOW	LOW	NONE	LOW	NOT_BUSY
10	LOW	NORMAL	NONE	LOW	NOT_BUSY
11	LOW	NORMAL	1	LOW	NOT_BUSY
12	LOW	LOW	NONE	LOW	NOT_BUSY
13	LOW	LOW	NONE	LOW	NOT_BUSY
14	LOW	NORMAL	NONE	LARGE	NOT_BUSY
15	LOW	NORMAL	NONE	LOW	NOT_BUSY
16	LOW	NORMAL	NONE	NONE	NORMAL
17	NORMAL	NORMAL	NONE	NORMAL	NORMAL
18	NORMAL	NORMAL	NONE	NORMAL	NORMAL
19	NORMAL	LARGE	NONE	LOW	BUSY
20	LARGE	NORMAL	1	NORMAL	BUSY
21	LARGE	NORMAL	NONE	LOW	BUSY
22	LARGE	NORMAL	NONE	LARGE	NORMAL
23	NORMAL	NORMAL	NONE	LOW	BUSY
24	LARGE	NORMAL	NONE	LOW	BUSY
25	LARGE	LOW	1	NORMAL	NORMAL
26	NORMAL	NORMAL	NONE	NORMAL	NORMAL
27	NORMAL	NORMAL	NONE	NORMAL	NORMAL
28	NORMAL	NORMAL	NONE	LOW	BUSY
29	LARGE	NORMAL	2	LARGE	NOT_BUSY

Appendix E: Hospital Forms Descriptions

INFORMATION CENTER DEPARTMENT
KING KHALED GENERAL HOSPITAL
DATE : 5 / 7 / 1433
DAY: SATURDAY

1433

DAILY FLOUR CENSUS

ADMISSIONS						DICHARGES										
WARD	REM. P.		NEW ADM.		TRA.IN	TRA.OUT	TRA.TO OH	DEATH	DIS.		TOT. REM. P.		NO.O. BEDS	L.O.S	BED. OCC. RATE	
	S	N S	S	N S					S	N S	S	NS				
M.M.W	10	1	2	0	0	0	0	0	3	1	9	0	33	14	27%	
F.M.W	17	2	5	0	0	2	0	0	8	2	12	0	36	18	33%	
M.S.W	29	3	4	1	1	0	0	0	10	1	24	3	35	11	77%	
F.S.W	13	3	2	0	1	0	0	0	2	1	14	2	36	8	44%	
O.B.S	42	1	20	2	0	0	0	0	32	1	30	2	55	38	87%	
NEURS	25	2	3	0	0	0	0	0	2	1	26	1	35	18	77%	
PED.W	32	1	7	0	0	0	0	0	10	1	29	0	53	22	55%	
P.ICU	2	0	0	0	0	0	0	0	0	0	2	0	4	0	50%	
I.C.U	6	1	2	0	1	1	0	0	0	0	8	1	9	3	100%	
C.C.U	3	0	0	0	1	1	0	0	0	0	3	0	4	3	75%	
TOTAL	179	14	45	3	4	4	0	0	68	8	157	9	300	117	55%	
193		48								75		166				
BED OCCUPANCY RATE : %55						ADULT CAPACITY %52						NERS/PEDIA: %63				

Figure E.1: Daily Flour Census form

In this form, the hospital meant accurate or precise instead of Flour. The daily Flour Census form (Figure E.1) gives a daily admissions and discharges statistics from King Khaled General Hospital in Saudi Arabia. The hospital has 10 wards, as shown in Figure E.1. Table E.1 expands the ward acronyms. Every ward is described by several columns with values filled by numbers. The form of Figure E.1 captures data for one day, as shown by the date in the upper left corner. 1433 is Arabic for 2012. This form (Figure E.1) is generated daily. Every ward row has acronym, an abbreviation for their name shortcut letters of the ward name; see below:

Table E.1: Ward acronym expansions

Ward	Description
M.M.W	Male Medical Ward
F.M.W	Female Medical Ward
M.S.W	Male Surgery Ward
F.S.W	Female Surgery Ward
O.B.S.	Obstetrical
NEURS	Nursery
PED.W	Pediatric Ward
P.ICU	Pediatric Intensive Care Unit
IC.U	Intensive Care Unit
C.CU	Coronary Care Unit

The description of every column in Figure E.1 follows:

REM. P. stands for Remaining Patients from the previous day, which has Saudi and Non-Saudi patients.

NEW ADM. stands for newly admitted patients in the ward that determined in the row, so they are counted separately, rather than with the remaining patients.

TRA.IN stands for the number of patients transferred from other wards to this ward

TRA.OUT stands for patients were discharged from this ward to other ward in the hospital.

TRA.TO OH stands for the patients who transferred from this hospital to other hospitals.

DEATH: patients died in this ward.

Figure E.2 is more detailed than Figure E.1. It contains the hospital number, age, gender, name, admitted time, and diagnosis, for all patients in a specific ward. However, Figure E.1 has only

the number and statistics of patients in the ward, in contrast with the textual data given in Figure E.2. Notice that the form is stamped by ward name at the top of it.

Table E.2: ADMISSIONS box, on left top of the form of Figure E.2

Field	Full Name	Description
Hosp. No.	Hospital Number	The hospital number or file number is a patient's unique number, used for the hospital purposes only
Age	Age	Physicians may need patient age for medical purpose
Sex	Sex or gender	The patient gender is needed in Saudi hospitals
Admitted	Admitted name	Full name of the newly admitted patient
Time	Time	Time that the patient is admitted
Diagnosis	Diagnosis	Determined by physician only

Received by Transfer From Other Ward and Died boxes are same as ADMISSIONS box

Table E.3: DISCHARGES box, on right top of the form of Figure E.2

Field	Full Name	Description
Hosp. No.	Hospital Number	The hospital number or file number is a patient's unique number used for hospital purposes only
Age	Age	Physicians may need patient age for

		medical purpose
Sex	Sex or gender	The patient's gender is needed in Saudi hospitals
Discharged	Discharged name	Full name of Discharged patient
Time	Time	The time that the patient is discharged out of the ward or the hospital
Length of Stay	Length of Stay	The number of days the patient was admitted in the ward
Diagnosis	Diagnosis	Determined by physician only

The Discharged By Transfer to Other Ward box is identical to the DISCHARGES box

23/9/1433 H
Saturday

ADMISSIONS

CLINIC	M.	FEM.	PED.	S	N S	S	N S	NEW	OLD	TOTAL
CARDIO (1)	12	10				20	2	1	21	22
CARDIO (2)										
NEURO . S	13	17				30		6	24	30
M . DERMA	13	12				22	3	18	7	25
F . DERMA	6	5				9	2	3	8	11
GYNE 1		17				17			17	17
GYNE 2		20				19	1		20	20
O . B . S (1)		23				23		3	20	23
O . B . S (2)		13				13		2	11	13
O . B . S (3)		10				10		10		10
F . MED.	3	13				16			16	16
M . MED	4	3				7			7	7
OPHTHALMO (1)	16	8				23	1	14	10	24
ORTHO (1)	18	19				37		3	34	37
ORTHO (2)	14	21				33	2	5	30	35
ORTHO(3)										
E . N . T (1)	11	10				21			21	21
E . N . T (2)										
PAEDIA (1)			8			8		4	4	8
PAEDIA (2)			22			21	1	10	12	22
PED. S.			15	2		15		5	10	15
ENDOSCOPY										
URO 1	6	6		3		12		8	4	12
URO 2	15	5				18	2	8	12	20
Eye2	10	6				16		16		16
VAS.S										
PLASTIC SURG										
NEUROLOGY	7	8				14	1		15	15
VAS.SURG										
THORCIC.SURG	4					3	1	2	2	4
NEURO .SURG 2										
CHEST										
SUR. (1)	2	9				11		3	8	11
SUR. (2)	6	9				15		2	13	15
ENDO (1)		4				4			4	4
#CLINIC(1)										
ENDO		6				6			6	6
DIABETIC	6	51				57		8	49	57
DIABETIC FOOT	9	4				13		1	12	13
O.B.S DIABETIC										
PEDIA DIABETIC		12				12			12	12
TOTAL	175	321	45	5		525	16	132	409	541
DROP IN										

GRAND TOTAL OF E.R VISITS (MAIN E.R + OBS E.R) = (802)

MAIN ER :726 S: 718 NS : 8			OBS ER :76		
ER.CASES : 80 COLD :646		ADM :26 S:26 NS :0		S: 72 NS: 4	
MED :4 SUR :13 PED :8 N PICU :1 RTA :0 PW.IMC:0				GYNE: 6 OBS : 70	
ICU :0 CCU :0 NUR :0 AKU: 0 OT:0 ASS :5				ADM:31 S : 27 NS :4	

Figure E.3: Clinics Form is for one day of statistics and the day was Saturday

The hospital has 41 clinics in operation every business day, which do not operate on weekends. The form of Figure E.3 shows statistics for one day only. It has the number of patients that visited a clinic per day. It specifies the visiting patient as male, female, or child.

Table E.4: Field of clinics hospital form of Figure E.3

Field	Full Name	Description
CLINIC	Clinic name	The name or the shortcut of the clinic
M	Male	Male patient
FEM.	Female	Female patient
PED	Pediatric	Child patient
ADMISSIONS S	Saudi admitted	Doctor in the clinic admitted a Saudi patient
ADMISSIONS NS	Non-Saudi admitted	Doctor in the clinic admitted a Non-Saudi patient
S	Saudi visitor	Saudi patient visited
NS	Non-Saudi visitor	Non-Saudi patient visited
New	New patient	New patient visited the clinic for the first time
OLD	Old patient	Follow up patient visited the clinic
TOTAL	Total number	Total number visited the clinic by the end of the day

The explanation for the first row in the form, is that the Cardiology clinic received 12 male visitors, 10 female visitors, and 0 child visitors. The clinic admitted no patient in that day. Patients comprise 20 Saudi nationalities and 2 Non-Saudi nationalities. There is 1 new patient

and 21 old patient visitors (follow up patients), so the total numbers who visited the clinic are 22 patients.

Appendix F: Webpage Form Descriptions

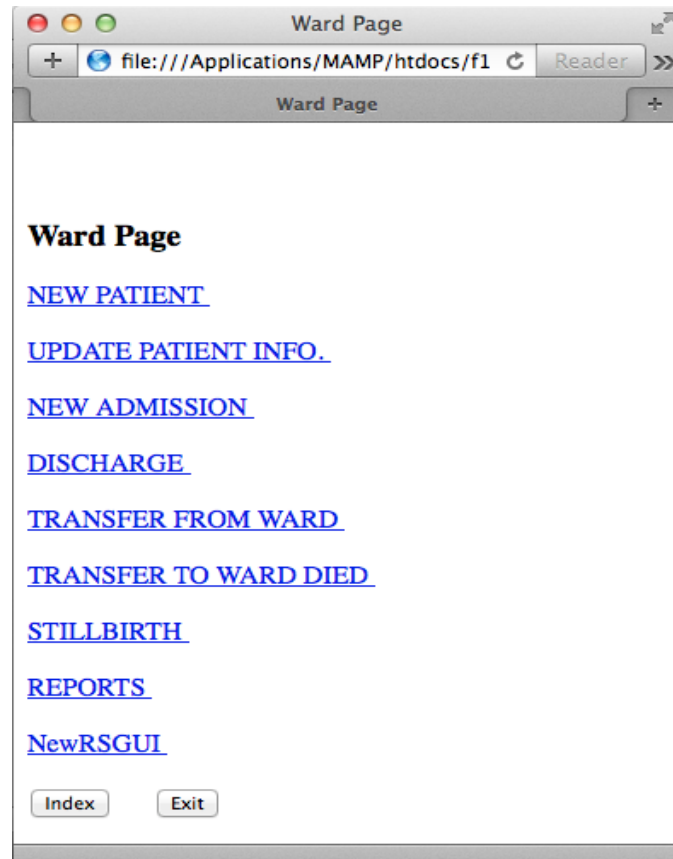


Figure F.1: Webpage Used by Ward Receptionist

The webpage of Figure F.1 is the front page for a ward. If New Patient link is clicked another page appears that can be seen in Figure F.2. The purpose of this link is to help gather data about a new patient. If Update Patient Info link is clicked another page appears that can be seen in Figure F.4. The purpose of this link is to update the patient information whom is already in the ward.

The screenshot shows a web browser window titled 'Hospital Website' with the address bar displaying 'http://localhost:8888/med.php'. The page content is titled 'New Patient' and contains the following form fields:

- Saudi ID:
- File No.:
- First Name:
- Last Name:
- D.O.B.:
- Sex: ☐ Male ☐ Female
- Citezen:
- Tel:
- Address:
- City:

At the bottom of the form are three buttons: 'Save', 'Reset', and 'Back'.

Figure F.2: Web interface for gathering New Patient Information

Once New patient link in first page is clicked, we are going to see the New patient page illustrated in Figure F.2 which has several fields as described in Table F.1.

Table F.1: New patient page Description

Ward	Description
Saudi ID	A patient ID in Saudi Arabia and it should be 10 numbers
File No.	Patient file number given by the hospital
First Name	Patient first name
Last name	Patient last name
D.O.B	Patient date of birth and the year is enough

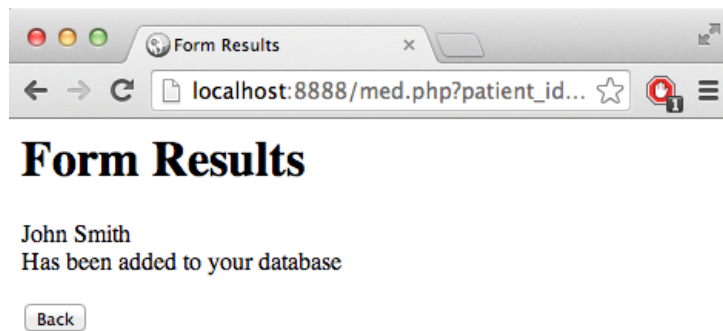


Figure F.3: Results of adding a new patient to the database

It tells if the fields filled and the Saudi ID field should be unique, if is not, it will show an error.

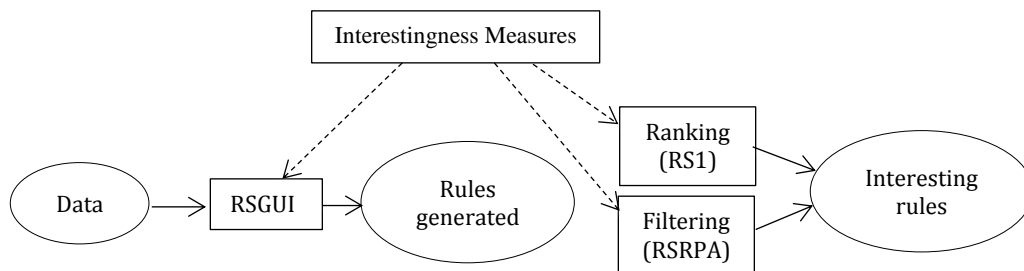


Figure F.4: Interestingness measures in play for rule generation

Figure F.4 is repeated here for convenience and for its usefulness to see how the Rough Set Graphical user Interface can be integrated with the processes of knowledge acquisition and creation of information flow through the hospital.

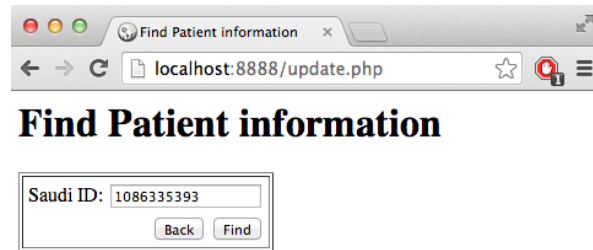


Figure F.5: Web page for search of patient information in the database

This webpage is used to search and find patient information by his/her Saudi ID. Pressing Find button sends the user to existing information to let the user update and review patient information.

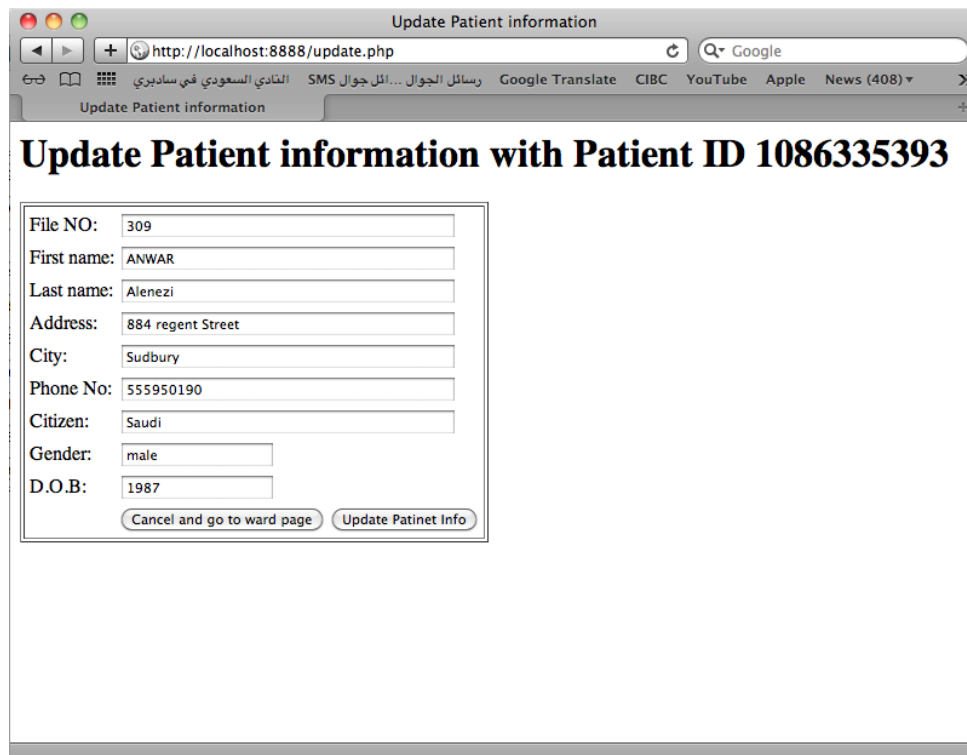
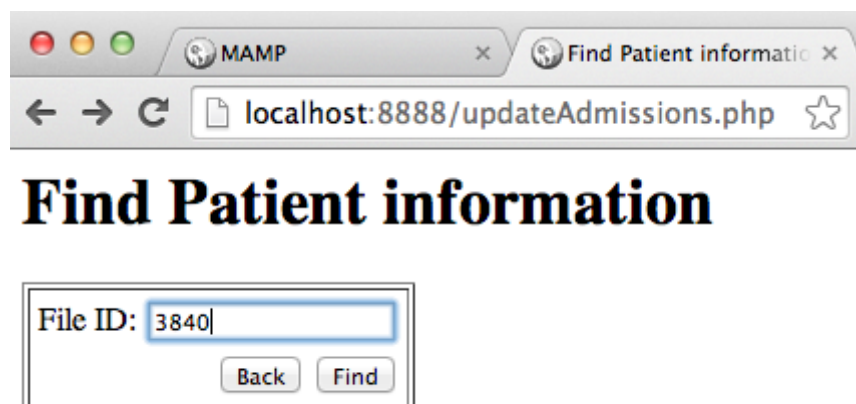


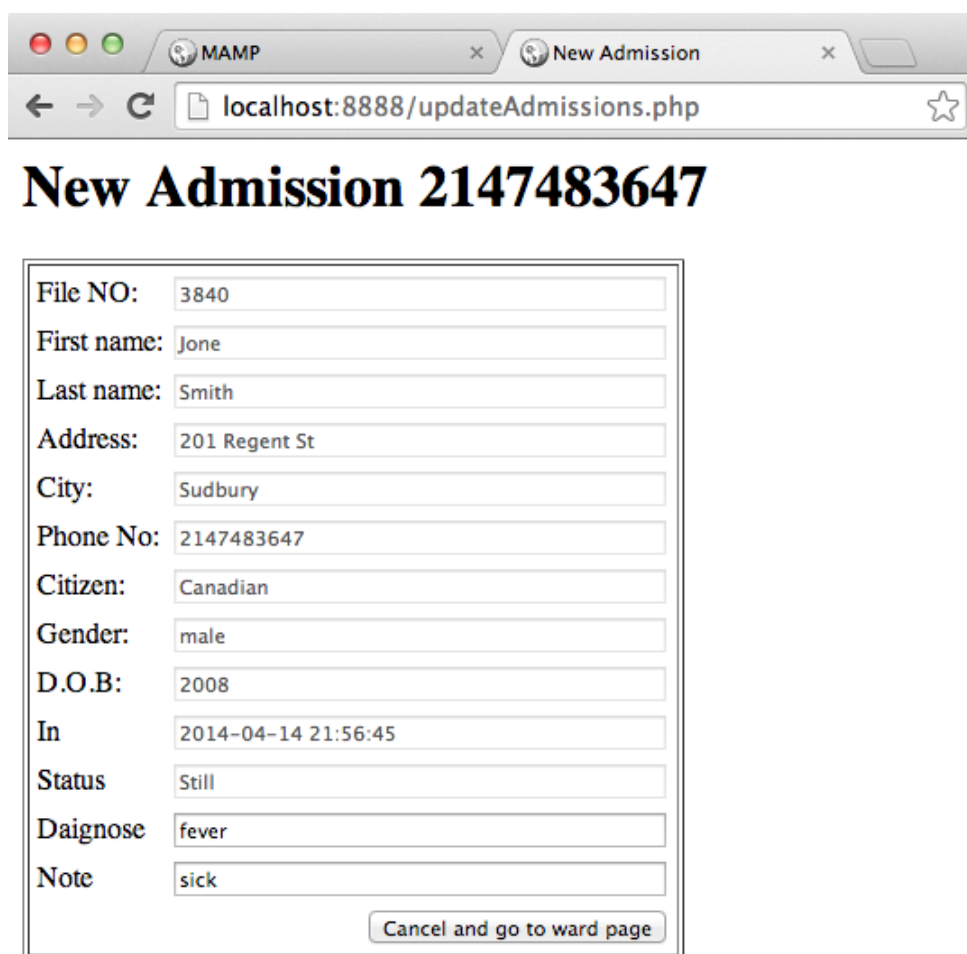
Figure F.6: Webpage for review and update of patient information



Find Patient information

File ID:

Figure F. 7: New admitted patient diagnosis



New Admission 2147483647

File NO:

First name:

Last name:

Address:

City:

Phone No:

Citizen:

Gender:

D.O.B:

In:

Status:

Daignose:

Note:

Figure F.8: Admitted patient's status and diagnosis

Patient's diagnosis could be seen by the webpage in Figure F.6 and it will open the new webpage which is Figure F.7 to check what time patient was admitted, status, diagnosis of the patient, and doctor note if there any.

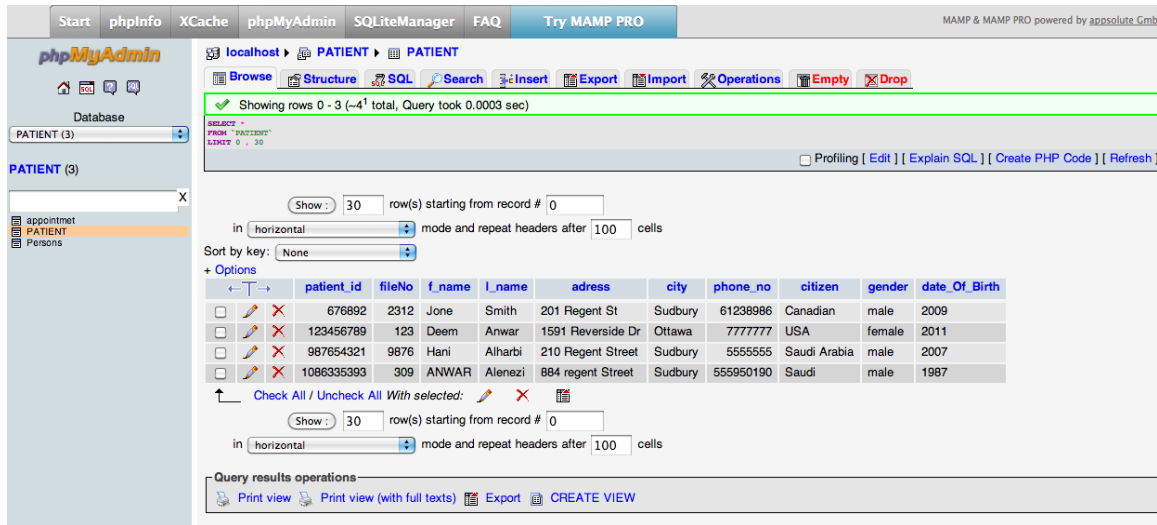


Figure F.9: Knowledge acquisition by means of webpage/MySQL prototype

In the end, a report of the ward could be printed with number of patients, admit time, discharge time, status, and the length of patient stay by the search in Figure F.10.

The screenshot shows a web browser window with the address bar displaying 'localhost:8888/reports.php'. The page title is 'Reports'. Below the title, there are three input fields for search filters:

- Day: 14
- Month: 4
- Year: 2014

At the bottom of the form, there is a 'search' button.

Figure F.10: Daily ward report

Appendix G: NewRSGUI java code

G.1 NewRSGUI snippets java code

This appendix shows snippets of java codes of NewRSGUI that I am using in this chapter to analyze the student records.

The following code is to create the text field to identify the address or the Server of the database that is going to be used or analyzed using the methods that exist in the RSGUI.

```
JLabel labelURL = new JLabel("Database URL:");
JTextField textFieldURL = new JTextField(15);
textFieldURL.setMaximumSize(textFieldURL.getPreferredSize());
Box hboxURL = Box.createHorizontalBox();
hboxURL.add(labelURL);
hboxURL.add(Box.createHorizontalStrut(35));
hboxURL.add(textFieldURL);
```

The following code is to create the text field of the Server's port of MySQL. To connect to the database in MySQL, it is needed to define the port to access the database. The code is most likely similar to the URL database code.

```
JLabel labelPort = new JLabel("Database Port:");
JTextField textFieldPort = new JTextField(15);
textFieldPort.setText("3306");
textFieldPort.setMaximumSize(textFieldPort.getPreferredSize());
Box hboxPort = Box.createHorizontalBox();
hboxPort.add(labelPort);
hboxPort.add(Box.createHorizontalStrut(35));
hboxPort.add(textFieldPort);
```

The following code is to create the text field for the Database Schema. Database Schema is the database name on the Server and it must be provided to run the program.

```
JLabel label1 = new JLabel("Database Schema:");
final JTextField textField1 = new JTextField(15);
textField1.setMaximumSize(textField1.getPreferredSize());
Box hbox1 = Box.createHorizontalBox();
hbox1.add(label1);
hbox1.add(Box.createHorizontalStrut(15));
hbox1.add(textField1);
```

The following codes are to create the text fields of the Username and its Password for the

database.

```
JLabel label2 = new JLabel("Username:");
JTextField textField2 = new JTextField(15);
textField2.setMaximumSize(textField2.getPreferredSize());
Box hbox2 = Box.createHorizontalBox();
hbox2.add(label2);
hbox2.add(Box.createHorizontalStrut(60));
hbox2.add(textField2);
```

```
JLabel label3 = new JLabel("Password:");
JPasswordField textField3 = new JPasswordField(15);
textField3.setMaximumSize(textField3.getPreferredSize());
Box hbox3 = Box.createHorizontalBox();
hbox3.add(label3);
hbox3.add(Box.createHorizontalStrut(60));
hbox3.add(textField3);
```

Figure G1 shows how the final screen looks after being generated by the above codes. This is much better than the command line bunches of RSGUI.

Once the interface is displayed, there is a function that acquires data from the database identified in the interface, and here is the function code as follows:

```
public static Connection getMySQLConnection(String durl,
    String port, String sch, String un, String pwd) throws Exception {
    String driver = "org.gjt.mm.mysql.Driver";
    String url = "jdbc:mysql://" + durl + ":" + port + "/" + sch;
    String username = un;
    String password = pwd;
    conn = DriverManager.getConnection(url, username, password);
    return conn
}
```

The SelectTable class is implemented as shown below, which is a Java frame class. In this class, the input is the connection. It lets the user to select one table as an input for the main function; also, there is a text field, which asks user to input the decision attributes number, that how many columns will be treated as decision attributes in the selected table. See Figure G2.



Figure G1:Select table and number of decision attributes interface

```

public SelectTable(final Connection Conn)
try {
    DatabaseMetaData dmd = Conn.getMetaData();
    ResultSet rs = dmd.getTables(null, null, "%", null);
    while (rs.next()) {
        comboBoxTable.addItem(rs.getString(3));
    }
    comboBoxTable.setMaximumSize(comboBoxTable.getPreferredSize());
    hboxTables.add(comboBoxTable);
    decisions.setMaximumSize(decisions.getPreferredSize());
    decisions.setText("1");
    hboxAttributes.add(Box.createHorizontalStrut(10));
    hboxAttributes.add(decisions);
} catch (SQLException e) {
    String message = "Tables retrieval failure, try again...";
    JOptionPane.showMessageDialog(new JFrame(), message, "Error",
        JOptionPane.ERROR_MESSAGE);
}

try {
    DatabaseMetaData dmd = Conn.getMetaData();
    ResultSet rsColumns = null;
    rsColumns = dmd.getColumns(null, null, comboBoxTable.getSelectedItem().toString(), null);
    int columnNum = 0;
    String[] columnHeader = new String[1000];
    while (rsColumns.next()) {
        columnHeader[columnNum++] = rsColumns.getString("COLUMN_NAME");
    }
    int decisionNum = Integer.parseInt(decisions.getText().toString());
    String columns = "";
    while (decisionNum > 0) {
        columns = columnHeader[--columnNum] + " " + columns;
        decisionNum--;
    }
    if (JOptionPane.showConfirmDialog(new JFrame(),
        "The decision attributes is(are) " + columns, "Confirm",
        JOptionPane.YES_NO_OPTION) == JOptionPane.YES_OPTION) {
        DecisionTable dt = new DecisionTable(Conn, comboBoxTable.getSelectedItem().toString(),
            decisions.getText().toString());
        try {

```

```

        new RSGUI(dt);
        close();
    } catch (IOException e) {
        String message = "Initial RSGUI failure, try again...";
        JOptionPane.showMessageDialog(new JFrame(), message, "Error",
            JOptionPane.ERROR_MESSAGE);
    }
}
} catch (SQLException e) {
    String message = "Tables columns names retrieval failure, try again...";
    JOptionPane.showMessageDialog(new JFrame(), message, "Error",
        JOptionPane.ERROR_MESSAGE);
}
}

```

The DecisionTable class is designed. Snippets of the class are shown below. All in all, the DecisionTable class tries to define the number of rows and columns, generates the real data array, sorts the array, and creates the attributes and so on.

```

public DecisionTable(Connection pConn, String tablename, String decisionNum) {
    Conn = pConn;
    tableName = tablename;
    DecisionAttributes = Integer.parseInt(decisionNum);
    newsetRowsAndColumns();
    newsetDataArray();
    newcreateAttributes();
    setDecisionAttributes(getDecisionAttributes());
}

```

The new method to set the number of rows and columns

```

private void newsetDataArray() {
    data = new String[rows][columns];
    try {
        Statement stat = Conn.createStatement();
        ResultSet rs1 = stat.executeQuery("Select * from " + tableName);
        for (int i = 0; i < rows; i++) {
            rs1.next();
            for (int j = 0; j < columns; j++) {
                data[i][j] = rs1.getString(j + 1);
            }
        }
    } catch (SQLException e) {
        String message = "Data generating failure, try again...";
        JOptionPane.showMessageDialog(new JFrame(), message, "Error",
            JOptionPane.ERROR_MESSAGE);
    }
}

```

The NewRSGUI method of creating attributes is shown as follow:

```

private void newcreateAttributes() {
    attributes = new Attribute[columns];
    //String rowOne = null;
    DatabaseMetaData dmd;
    try {

```

```

dmd = Conn.getMetaData();
ResultSet rsColumns = null;
rsColumns = dmd.getColumns(null, null, tableName, null);
String columnName = null;
int a = 0;
while (rsColumns.next()) {
    columnName = rsColumns.getString("COLUMN_NAME");
    attributes[a++] = new Attribute(columnName);
}
for (int c = 0; c < columns; c++) {
    for (int r = 0; r < rows; r++) {
        if (!valueInArray(attributes[c].values, data[r][c])) {
            attributes[c].addAttributeValue(data[r][c]);
        }
    }
}
for (int c = 0; c < columns; c++) {
    attributes[c].values = sortValues(c);
}
} catch (SQLException e) {
    String message = "Attributes generating failure, try again...";
    JOptionPane.showMessageDialog(new JFrame(), message, "Error",
        JOptionPane.ERROR_MESSAGE);
}
}
The new method to get attributes
private int[] getDecisionAttributes() {
    int number = DecisionAttributes;
    int[] dec = new int[number];
    for (int i = 0; i < number; i++) {
        dec[i] = columns - number + i;
    }
    return dec;
}

```

Appendix H: Web interfaces design

Web interface design in this appendix is the final result, which is agreed upon the Manager of Computer Science and Statistical Department and the director of the hospital in king Khaled General Hospital. Web interface designed by hand written that illustrate it below.

Login Page

User Name:

Password:

Female Medical Ward

[Admissions](#)[Discharges](#)[Transfer from ward](#)[Transfer to ward](#)[Died](#)[Stillbirth](#)[New patient](#)[Index](#)[Exit](#)

New Patient

Saudi ID	<input type="text"/>	File No.	<input type="text"/>
First Name	<input type="text"/>	Second Name	<input type="text"/>
Family Name	<input type="text"/>	Sex:	<input type="checkbox"/> Male <input type="checkbox"/> Female
D.O.B.	<input type="text"/>		
Citizen	<input type="text"/>	Phone No.	<input type="text"/>
Address	<input type="text"/>		

Save**Clear****Back**

FMW Admissions

Saudi ID File No. First Name Second Name Family Name Sex D.O.B. Citizen Phone No.

Address

Admitted Time Diagnose

Note

Save**Cancel****Back**

FMW Discharges

Saudi ID File No. First Name Second Name Family Name Sex D.O.B. Citizen Phone No. Discharged Time Length of stay Diagnose

Note

FMW Transfer from another ward

	Transfer From	<input type="text" value="▼"/>
Saudi ID	<input type="text"/>	File No. <input type="text"/>
First Name	<input type="text"/>	Second Name <input type="text"/>
Family Name	<input type="text"/>	Sex <input type="text"/>
D.O.B.	<input type="text"/>	
Citizen	<input type="text"/>	Phone No. <input type="text"/>
Transferred Time	<input type="text"/>	
Diagnose	<input type="text"/>	
Note	<input type="text"/>	

Save**Cancel****Back**

Transfer to another ward

Saudi ID File No. First Name Second Name Family Name Sex D.O.B. Citizen Phone No. Transferred Time Length of stay Diagnose

Note

Transferred To **Save****Cancel****Back**

FMW Death

Saudi ID	<input type="text"/>	File No.	<input type="text"/>
First Name	<input type="text"/>	Second Name	<input type="text"/>
Family Name	<input type="text"/>	Sex	<input type="text"/>
D.O.B.	<input type="text"/>		
Citizen	<input type="text"/>		
Death Time	<input type="text"/>		
Diagnose	<input type="text"/>		
Note	<input type="text"/>		

Save**Cancel****Back**

Stillbirth

Saudi ID ^{New} File No.

First Name Second Name

Family Name Sex

D.O.B.

Citizen

Time

Diagnose

Note

Save

Cancel

Back