UNIVERSITI PUTRA MALAYSIA

CLASSIFICATION SYSTEM FOR HEART DISEASE USING BAYESIAN CLASSIFIER

ANUSHA MAGENDRAM.

FSKTM 2007 9

CLASIFICATION SYSTEM FOR HEART DISEASE USING BAYESIAN CLASSIFIER

BY

ANUSHA MAGENDRAM

Thesis submitted in Partial Fulfilment of the Requirement for the Degree of Master of Science in the Faculty of Computer Science and Information Technology University Putra Malaysia

DEC 2007

APPROVAL SHEET

This thesis is received, checked and approved for submission in partial fulfilment of the requirement for the degree of Master of Science in the Faculty of Computer Science and Information Technology, University Putra Malaysia, Serdang, Selangor Darul Ehsan.

... ...

(Prof. Madya Dr. Md. Nasir Sulaiman)

(Supervisor)

Faculty of Computer Science and Information Technology

University Putra Malaysia

Serdang Selangor Darul Ehsan

... ...

Date

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations, which have been dully acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions.

Signature

... ...

Anusha Magendram

ABSTRACT

Increase of hearth problem in this world is rising each day. Classification system for heart disease is a system that able to justify whether a patient has heart problem or not. This is a new approach that able to use by doctors to rectify the heart problem.

This system was developing base on to three main part which is data processing, testing and implementation of the algorithm. In this system a Bayesian algorithm was used in order to implement the system.

This system was mainly developing using java programming. Apache Tom cat was used as a server in order to run the application smoothly.

This system is tested using the test and training data set. It has proven that the system able to provide an accurate result on justifying whether the patient has has problem or not.

As a conclusion by using this system doctors able to improvise the effectiveness in medical field.

ABSTRAK

Pada masa kini ,penyakit jantung semakin meningkat dari masa ke semasa.Sistem klasifikasi untuk penyakit jantung merupakan sebuah sistem yang mapumegesan sama ada pesakit itu menghadapi masalah jantung.Ia juga merupakan kaedah baru yang dapat membantu doctor megenal pasti jantung yang bermasalah.

Sistem ini dibina berdasarkan tigafasa iaitu pemprocessan data , pengujian dan implementasi berdasarkan algoritma. Algoritma Bayesian telah di gunakan dalam sistem ini untuk process pengimplementasian.

Secara keseluruhannya,sistem ini telah di bangunkan mengunakan program java.Apache tom cat digunakan sebagai terminalpelayan supaya perjalanan sistem lebih lancar.

Sistem ini telah diuji mengunakan kaedah pengujian data dan pengujian latihanbahawa sistem ini mampu memberikan keputusan yang tepat dalam menjustifikasi masalah pesakit.

Kesimpulanny , pengunaan sistem ini boleh memberikan manfaat kepada doctor untuk memperbaiki keberkesanan sesuatu process di bidang perubatan.

Blessing of

My Family & Best Friend

For their loving and caring inspired me:

"There's No Success without Hard work"

ACKNOWLEDGEMENTS

This thesis means a lot to me. Many people have played their role to assist me in all manners. Support and encouragement was the main thing I ever needed.

Deepest thanks to my supervisor; Prof. Madya Dr. Md. Nasir Sulaiman for his assistance, guidance and support throughout the work of this thesis is greatly appreciated.

Dedications to dear mum, dad and all my family members who had been the very supportive, my best friend for his encouragement and others who were always there with help whenever I need it.

Last but no least, my thanks to all that are involved in this project either directly or indirectly. I hope that god blesses all of the efforts being put on this project by the whole of you.

TABLE OF CONTENTS

		Pages
APPROVAL SHEET		ii
DECLARATION		iii
ABSTRACT		iv
DEDICATION		vi
ACKNOWLEDGEMENT.		vii
TABLE OF CONTENTS		xi
LIST OF FIGURES		xii
LIST OF TABLES		xiii
CHAPTERS		
1 INTRODUCTION		1
1.1	Background	1
1.1	Problem Statement	3
1.3	Objective	4
1.4	Scope	4
1.5	Organization of the thesis	5
2 LITERATURE REVIEW		7
2.1	Introduction	7

2.2	Classification	9
2.3	Two Tier Software Architectures	10
	2.3.1 Technical Details	10
	2.3.2 Usage of two tier architecture	12
2.4	Previous Researches	13
	2.4.1Ensemble Feature Selection with the Simple Bayesian Classification in Medical Diagnostics	13
	2.2.4 A global optimization approach to classification in medical diagnosis and prognosis	15
	2.4.3Mining association rules using asthma patient profile data set	17
METHODOLOGY		18
3.1	Introduction	18
3.2	Software Development Life Cycle (SDLC)	19
	3.2.1 User Requirement analysis	20
	3.2.2 Design	20
	3.2.3 Coding	21
	3.2.4 Implementation	21
	3.2.5 Testing	21
	3.2.6 Maintenance and Operation	21
3.3	Guide Line of The Research Work	22
3.4	Data Preparation	25
	3.4.1 Raw Data Structure	25
	3.4.2 Pre – Processing Data	26

3.5	Generating Of Experiment & Test Data Set	30
3.6	Implementation of The Algorithm	30
3.7	Summary	31
4	**IMPLEMENTATION AND EXPERIMENTAL DESIGN**	32
4.1	System Architecture	32
4.2	System Module	33
4.3	User interface	38
	4.3.1 Interface Design	38
	4.3.2 User Input	41
	4.3.3 Output Design	41
4.4	Connection With The Server	42
4.5	Connection With The Database	43
4.6	Hardware & Software Requirement	45
4.7	System Work Flow	46
4.8	Conclusion	56
5	**TESTING**	57
5.1	Testing	57
	5.1.1 Unit Testing	57
	5.1.2 Integration Testing	59
	5.1.3 System Testing	60
6	**RESULT AND DISCUSSIONS**	64
6.1	Performance of the System Data Set Conversion	64

6.3 Summary 69

7 CONCLUSION AND FUTURE WORKS 70

7.1 Conclusion 70

7.2 Contribution 71

7.3 Future Works 71

REFERENCES 73

LIST OF FIGURES

FIGURE NO	TITLE	PAGE
2.1	Two Tier Client Server Architecture	11
3.1	Software Development Life Cycle	19
3.2	Guide line Chart	23
3.3	Pre –Processing Steps	26
4.1	Architecture Design	32
4.2	System Module	33
4.3	Main Page Interface Design	38
4.4	Uploading Screen Interface	38
4.5	Data Processing Interface Design	40
4.6	System work Flow Chart	46
4.7	Main screen	47
4.8	Uploading Data Screen	47
4.9	Display Data Screen	48
4.10	Missing Data Screen	49
4.11	Algorithm / Data Concept Implementation Screen	50
4.12	Option To calculate Screen	51
4.13	Calculation Screen	52
4.14	Report Screen	53
4.15	Statistic Report	53

4.16	Gender Comparison Report	54
4.17	Clean Data Set	55
6.1	Converted Data Set	66
6.2	Optional Calculation	67
6.3	Bayesian Calculation	68

LIST OF TABLES

TABLE NO	TITLE	PAGE
4.1	Data Set Options	36
4.2	Data Dictionary	44
4.3	Hardware & Software Requirement	45
5.1	System Testing	60
6.1	Actual Data Set	65
6.2	Test Data Set	67

CHAPTER 1

INTRODUCTION

1.1 Background

Health is the most important aspect in human race. Everybody wants to live without disease but there's no escape to it. The only way out is precautions and early detection. There are many medical databases created with patient details and records. It became tans of collection .But no body has time to always over look at this. This is because there are too many data's and they are not organized. How can they summarize it, so that it is easier for doctors to read it?

By using Data mining concepts. Data mining also known as knowledge discovery in database. Data mining is a process of analyzing data from different perspectives and summarizing it into useful information, which can be used to increase revenue and cut cost. It has gained considerable attention among the practitioners and researchers. There many kinds of data mining concepts such as association, classification, cluster, outlier and evolution. These concepts able to help to arrange the data in a systematic approach. Further more able to discover the trend and hidden pattern within the data's that could significantly enhanced our understanding of the patients ,diseases and the diagnose.

In this research classification is going to be the concept. Classification is the a suitable data mining method in order to analysis the raw data. It is an important area of research

in data mining. Classification partitions massive quantities of data into sets of common characteristic and properties. The classification model can be used to categorize future data samples, as well as providing a better understanding of the database contents. Classification is particularly useful when a database contains examples that can be used as the basis for future decision making, e.g. for assessing credit risks, for medical diagnosis, or for scientific data analysis.

Time and cost is the main issue when coming to saving lives .The quicker we save a life the cheaper the cost is. Well this is where the concept lays. By analysis the dieses and the relations with symptoms and diagnose quicken a doctors task. Well here is the method to analysis the disease.

1.2 Problem statement

Everybody has their own limitation , reading from whole list of data doesn't tell us anything. We are puzzle when looking at whole lots of details in one go. How do we solve this?

The main aim is to improvise the way of detecting a heart disease problem. Currently it is done manually .It will be good and much more effective if there is a system which can compile all the test result in order to detect the heart problem .There few question rise when it come to detecting heart problem. For an example : How can a doctor notice a patient has heart problem immediately? Is there a way to analysis the test results?

1.3 Objectives of the Research

The objective of this research is to develop a classification system for detecting heart disease using Bayesian classifier. Basically by using a data mining approach. Expect the system to detect heart disease problem using the Bayesian algorithm. This system also able to detect the missing value on the data and able to generate reports in graph forms.

1.4 Scope of the Research

This research is mainly focus on heart disease. It is to detect the present of heart disease on individual patient. In order to test this , the raw data is taken from the internet. The raw data is separated into two potions. A training data set and a test data set. The system will be tested using this two set of data.

1.5 Organization of the Thesis

The thesis has seven chapters which covers introduction , literature review ,methodology ,implementation and experimental design, testing , result and discussions and finally conclusion and future works.

Introduction chapter covers the background of the research , information of heart disease and methodology . In here , the objective and scope of the research also is covered.

The chapter two which is literature review covers the introduction of the overview of data mining concepts and classification concept . In here it is also explained about the architecture chosen , the purpose of it and the technical details. Further more , this chapter also covers the work of previous researchers which are used as a guide line for this research.

Chapter three covers the methodology used to implement this system. The introduction explains the over view of the whole concepts in methodology. In this chapter it also talks about the way of system development, the requirement , design and steps taken to implement the system.

The following would be chapter four. This chapter covers the implementation of the system . It also explains the over view and functions of the system.

Chapter five covers testing part. In this chapter it is explain what kind of the testing is available and should be done in a system. Also included the test done in this system and the result of it.

Result and discussion is chapter six. In this chapter , it is discuss the over view of the whole result of the system and the out put.

Finally the last chapter , chapter seven is the conclusion and future works. It explains the over view of whole research based and explores promising directions for future works and development.

CHAPTER 2

Literature Review

2.1 Introduction

In today's world, knowledge is power. An important source of knowledge is the data stored in databases. Data allows us to learn from the past and to predict the future. With the rapid computerization of businesses and organizations, a huge amount of data has been collected and stored in databases, and the rate at which data is stored is growing at a phenomenal rate. As a result, traditional ad hoc mixtures of statistical techniques and data management tools are no longer adequate for analyzing this vast collection of data.

Data mining (or knowledge discovery in databases or KDD in short) has emerged as a growing field of multidisciplinary research for discovering interesting/useful knowledge from large databases. KDD is defined as the extraction of implicit, previously unknown, and potentially useful patterns from data. Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table

2.2 Classification

Classification enables the categorization of data (or entities) into pre-defined classes (of course there should be two or more classes pre-defined before running categorization). The use of classification algorithms involves a training set consisting of pre-classified examples. The classifier calibration algorithm uses the pre-classified examples to determine a set of parameters required for proper discrimination between the classes.

The algorithm then encodes these parameters into a model called a classifier. Once such a classifier is calibrated, it can assign new filings to either of the classes. There are many algorithms that can be used for classification, such as decision trees, neural networks, logistic regression, etc. Using this method Data Mining system learns from examples or the data (data warehouses, databases etc) how to partition or classify certain objects (it can be an object, an action, or any other information, that can be formalized). As a result, data mining software formulates classification rules.

2.3 Two Tier Software Architectures

Two tier software architectures were developed in the 1980s from the file server software architecture design. The two-tier architecture is intended to improve usability by supporting a forms-based, user-friendly interface. The two-tier architecture improves scalability by accommodating up to 100 users (file server architectures only accommodate a dozen users), and improves flexibility by allowing data to be shared, usually within a homogeneous environment. The two-tier architecture requires minimal operator intervention, and is frequently used in non-complex, non-time critical information processing systems (Darleen Sadoski,2007).

2.3.1 Technical Details

Two tier architectures consist of three components distributed in two layers: client (requester of services) and server (provider of services). The three components are as below:

1. User System Interface (such as session, text input, dialog, and display management services)

2. Processing Management (such as process development, process enactment, process monitoring, and process resource services)

3. Database Management (such as data and file services)

The two-tier design allocates the user system interface exclusively to the client. It places database management on the server and splits the processing management between client and server, creating two layers. Figure 1 depicts the two-tier software architecture.

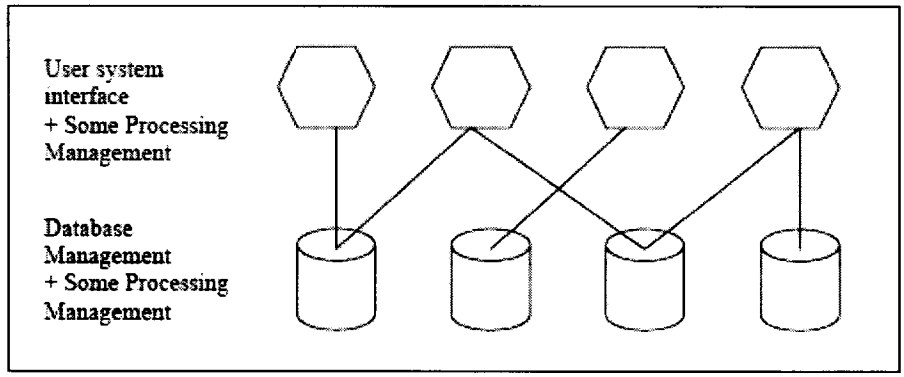

Figure 2.1: Two Tier Client Server Architecture

In general, the user system interface client invokes services from the database management server. In many two-tier designs, most of the application portion of processing is in the client environment. The database management server usually provides the portion of the processing related to accessing data (often implemented in store procedures). Clients commonly communicate with the server through SQL statements or a call-level interface. It should be noted that connectivity between tiers could be dynamically changed depending upon the user's request for data and services.

As compared to the file server software architecture (that also supports distributed systems), the two-tier architecture improves flexibility and scalability by