INFORMATION SYSTEMS DEPARTMENT
FACULTY OF COMPUTER AND INFORMATION SCIENCES
AIN SHAMS UNIVERSITY

A Semantic Web Based Platform for a Medical Decision Support System

A Thesis Submitted to Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University, Cairo, Egypt.

In partial fulfillment of the requirements for
the degree of master in Information system

By

Randa Adel El-Bialy
Information systems Department
Faculty of Informatics and Computer science
British University in Egypt

Under Supervision of

Prof. Dr. M. Essam Khalifa
Professor of Mathematics
Faculty of Computer and Information Sciences
Ain Shams University, Cairo, Egypt

Prof. Dr. Omar H. Karam
Dean of Informatics and Computer science
British University in Egypt

Dr. Ibrahim Fathy Moawad
Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University, Cairo, Egypt

Acknowledgement

I am indebted to many people who have influenced and inspired me in my research. Their enthusiasm, help and support have ultimately led to the completion of this work.

Firstly, I would like to express my sincerest gratitude to Prof. Ahmed Hamad for his constant support throughout the various obstacles that were encountered over the years, particularly the most recent. It is a fact that without this constant support and motivation, this work would not be where it is now.

Secondly, I would like to thank Prof. Mohamed Essam Khalifa for his constant exigency and motivation within the past few years, without which this work would have attained its maximum potential.

Thirdly, I would like to express my particular gratitude and deep appreciation to Prof. Omar Karam for his constant support and motivation as well as his invaluable advice and expertise that have allowed for the successful completion of this thesis dissertation.

I would like to further mention Dr. Nevine Makram's ceaseless efforts that have guided me throughout this work. Her constant encouragement and patience were an inspiration to push through.

It is a fact that Dr. Mostafa Salama's constant efforts were a huge contributor to the completion of this dissertation. He has provided insight and perspective that have helped this work immensely.

My deepest appreciation, gratitude and humbleness go to a dear colleague and friend, Dr. Marco Alfonse. He has truly provided a clear definition of friendship and integrity whilst sharing this long journey with me. No words can suffice to express my appreciation for his tremendous efforts.

Lastly, I would like to thank my family and dearest friends for enduring these past years and for never giving up on this work and on me. Their constant support could not have been replaced.

Abstract

Diagnosing Heart diseases is one of the problems that require high level of accurate analysis and prediction. Data sets dealing with the same medical problems such as coronary artery disease (CAD) may show different results when applying the same machine learning technique. Followed by searching for the best combination of classifiers in an ensemble that is generally suitable for all data sets of Heart diseases diagnoses. There exists multiple datasets targeting the same problem. On the other hand, several classifiers are available for the analysis of these data sets. This research aims to analyze the outcome of integrating the results of the common datasets in the same domain and as a second step the classification techniques applied in this domain. As for the first step, the two classifiers applied are decision tree algorithms (fast decision tree and pruned C4.5) the resulted trees are extracted from different data sets and compared. Common features among these data sets are extracted and used in the later analysis for the same disease in any data set. As for the second step, using ensemble methods in decision support systems provide an important help in analyzing this type of diseases. Six classifiers namely used which are (Bayesian Net, Naive Bayes, Multilayer perceptron, Sequential Minimal Optimization Algorithm (SMO), Decision Tree Algorithms (C4.5 and Fast Decision Tree (FDT)) to predict two different heart diseases subsequently. The results show that the classification accuracy of the collected dataset is 78.06% higher than the average of the classification accuracy of all separate datasets, which is 75.48%. On the Other hand, the classification accuracy results for the ensemble reached percentages higher than 90% accuracy. The best ensemble combination appears to be common for both datasets, composed of Bayesian Network, Naive Bayesian, Neural networks , C4.5 and SVM with 94% accuracy .

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Chapter (1)

1. Introduction

1.1. Heart Disease Diagnosis

The leading cause of death in the world, over the past 10 years is heart disease. The first and most leading cause of death in high and low income countries was reported by the World Health Organization to be heart disease [1]. Strokes, heart attacks and other circulatory diseases account for 41% of all deaths as reported by the European Public Health Alliance [2].Most lives in Asia, are lost to non- infectious diseases such as cancers ,cardiovascular diseases, diabetes and chronic respiratory diseases, as stated by the Economical and Social Commission of Asia and the Pacific [3]. In Australia, according to the Australian Bureau of Statistics 33.7% of all deaths are caused by heart and circulatory system diseases [4]. Heart disease is a general term that consists of a wide range of diseases that affect the heart such as shortness of breath, chest pain, heart attack and other symptoms [5]. Due to the insufficient blood flow received by the heart muscles, chest pain occurs. Heart disease includes cardiovascular disease which comprises of a wide diversity and variety of circumstances that distresses the blood vessels and the heart and the mechanism in which blood is pumped and circulated in the body [6]. Consistent with statistics, heart disease is one of the most leading and significant causes of deaths all over the world. Therefore, it was considered in the industry of health care that heart disease diagnosis is a crucial and important matter. This led to various attempts for developing an efficient Medical Decision Support System (MDSS) in order to assist the physicians in their diagnosis. The main aim of developing MDSS is to improve the diagnosis accuracy, enhance and further improve the time taken by the physician to make a diagnosis, in addition to support the progressive increase of the more complicated diagnosis decision process. [7, 8].The succession of information technology, system integration as well as software development, techniques have shaped a pioneering generation of comprehensive computer systems. Health care system is an instance of

these systems. Recently, the use of data mining technologies in healthcare systems gained an enlarged awareness. The main idea is to recommend a computerized method for diagnosing heart diseases based on prior data and information [9].

1.2. Problem Definition

Data Mining is a technique used for problem solving, in which uses the database in order to analyze the data stored in it. Data Mining is the process of analyzing the data in order to discover, classify and find hidden patterns in this data. Having a variety of different types of problems, data classification assists in the process of decision making. According to researchers and recent research, data mining techniques are proven to be tremendously helpful in several disease diagnoses such as heart disease [10], stroke [11], cancer [12], and diabetes [13]. The problem is to make useful predications that are beneficial for the decision support process; this includes discovering the relationships between data attributes and finding algorithms that are suitable to be applied for a specific field. In other words Data sets dealing with the same medical problems like Coronary artery disease (CAD) may show different results when applying the same machine learning technique. The heterogeneity datasets produce heterogeneity data types. These datasets provide information complementary to the medical diagnosis, and so it can enhance the classification accuracy percentages [14]. The diagnosis of patients is based on the accuracy of the medical data collected. Different problems exist for the medical data, which constitutes a real challenge in the field of medical informatics. These problems can be summarized as follows: [15, 16]

- Incorrect, sparse and temporal information.
- Small sized samples.
- Manual data collection inconsistencies.
- Missing values.
- Professionalism of medical analysts / practitioners / technicians in diagnosing the disease.
- Accuracy of machines or instruments used in the diagnosis.

Moreover, different data mining techniques for heart disease diagnosis have been introduced mostly by using a single data mining technique with level of accuracy that needs improvement; accordingly, we seek better accuracy by combining different classifiers together.

1.3. Objective

The aim of this work is to develop an intelligent heart disease diagnosis and prediction system and apply an integration of the results of the machine learning analysis applied on different data sets targeting the CAD disease. This will avoid the missing, incorrect, and inconsistent data problems that may appear in the data collection. Fast decision tree and pruned C4.5 tree are applied where the resulted trees are extracted from different data sets and compared. Common features among these data sets are extracted and used in the later analysis for the same disease in any data set. The results show that the classification accuracy of the collected dataset is 78.06% higher than the average of the classification accuracy of all separate datasets, which is 75.48%. Moreover an Ensemble data mining model has been introduced that performs better than individual classifiers, which proves to have high prediction accuracy and consistent performance. Six different classification techniques have been used Naïve Bayes, Bayes Net, Neural Network, Support vector machine, C4.5 and FDT. The classification accuracy results reached percentages higher than 90%.

1.4. Methodology

The proposed intelligent heart disease diagnosis and prediction system is composed of two parts. The first of the 2-part system involved the integration of different datasets of the CAD disease using the machine learning techniques. In addition, common features among these data sets are extracted and used in the later analysis for the same disease in any data set.

- Data sets that deal with the same medical problem are gathered from different resources. The number and type of attributes may differ from one data set to the other, but the classification category or target class label must be the same.

- Two classification techniques are applied; C4.5 and the fast decision tree, to extract the corresponding decision tree for each of the data sets and then to compare and detect the classification accuracy percentages.
- Five Common features among all the decision trees are extracted. For example, the Cleveland dataset as shown in fig.2 and fig.3 for C4.5 and fast decision trees classifiers. After building the C4.5 and Fast decision Tree significant difference in tree size and processing time between the fast decision tree and the C4.5 are shown in table 2, therefore the slight difference in accuracy between them can be ignored.
- These five most common features are used to build a new integrated data set from all the input data sets. The new pruned collected dataset includes only these extracted common features. These features shows highest information gain values that are selected to avoid over-fitting in the new generated and integrated decision tree.
- C4.5 and the fast decision trees are applied to this integrated data set, and a new decision trees are generated. In the processes of creating these new trees, another layer of pruning or feature selection, is applied. Only four features frequently appear in the new decision tree. The fast decision tree selects only ca, age, cp and thal.
- It may be noticed that for small number of instances, the classification accuracy of the decision tree applied on all features will be higher than that of the four-featured decision tree. But for big data of large number of instances, the tree size and so the processing time will be apparently low in the case of the four-featured data sets.
- One last step is applied to show the proposed model efficiency, which is the comparison between the average of classification accuracies of all data sets in separate and the classification accuracy of the integrating decision tree.

The second part of the 2-part system involves the integration of different and variant classification techniques using Ensemble data-mining models which performs better than the individual classifiers. The aforementioned further proved to have a higher prediction

accuracy and more consistent performance when compared to each technique individually. The proposed work shows different combinations of the classifiers and apply them to detect the one of the best prediction accuracy. Combining several classifiers, which differ in their properties and results, is the main focus of the proposed ensemble framework.

The proposed system uses UCI benchmark heart disease dataset and heart sound signals dataset for the detection of heart valve disease in order to identify, discover and extract the hidden knowledge associated with heart disease. It can differentiate between sick and healthy individuals with a high level of accuracy.

- Two datasets have been used the collective heart disease dataset (CAD)[53] and the heart sound signals dataset for the detection of heart valve disease [54].
- Detecting and removing the outliers and extreme values for both datasets is an important step to remove the noisy data using the Inter-quartile range method. As for the CAD dataset no outliers or extreme values were detected but as for the heart valve disease dataset outliers and extreme values were detected and removed.
- Choosing the classification algorithms will be used in the ensemble according to the most used with the heart disease datasets.
- Six classifiers were commonly used by other researchers [55][56], Naïve Bayes, Bayesian Network, Multilayer perceptron (a classifier that uses backpropagation to classify instances; Neural network) , sequential minimal optimization algorithm (SMO)for training a support vector classifier and Decision tree algorithms (C4.5 and Fast decision tree).
- Combining classifiers together using ensemble techniques: bagging, boosting and stacking in order to enhance the accuracy of the classification by merging them sequentially one by one.
- After combining classifiers together determining the best combination of classifiers according to their accuracies for the heart disease datasets in general.
- All the above steps are repeated for both of the data sets.

1.5. Organization of this thesis

This current thesis is organized as follows:

- Chapter 2 contains key concepts and previous work related to the topic.
- In chapter 3-architecture is presenter.
- Methodology is in chapter 4.
- The conclusion is provided in chapter 5.