Implementation of the Diagnosis System for Heart Diseases by using Decision Tree Induction Method

Lai War Zaw : Nilar Aye
University of Computer Studies, Yangon
zawandrew@gmail.com

Abstract

Computer have become an indispensable tool in the Health Care Industry. As technology grows rapidly many people take great interest in computer and then computer based method are used to improve the quality of the medical services. A decision support system is clearly an application that simply manipulates data or supports decision making. Decision tree induction algorithm has been successfully used in decision support system. Decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attributes, each branch represents an outcome of the test, and the leaf nodes represent classes or class distributions. There are over 100 types of heart diseases. This system provides 4 types of heart diseases. They are Heart Failure, Infective Endocarditis, Coronary Heart Disease and Valvular Heart Disease. In this paper, we implement the diagnosis system for heart diseases by using decision tree induction method. The heart diseases that we implement include (26) attributes (symptoms) and (4) classes (diseases). This system has 3000 train data and 1000 test data. This system provides the heart disease results to patients according to the symptoms that they are suffered and also provides some treatments that the patients should take.

1.Introduction

The effective identification of information from a large collection of data has been on a steady increase recently. Data mining is the task of discovering interesting patterns from large amount of data where the data can be stored in databases, data warehouse or other information repositories. Classification is the process of finding a set of modules. There are many techniques for data classification such as decision tree induction, Bayesian and so on. The efficiency of existing decision tree algorithm, such as ID3 and C4.5, has been well established for relatively small data sets. The ID3 algorithm is a well-known decision tree induction algorithm. A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.

Decision support systems (DSSs) are inherently complex in terms of both their data management and technology architecture[2]. DSS is part of a special category of information systems that are explicitly designed to enhance managerial decision making. Information systems that solve problems by capturing knowledge for a very specific and limited domain of human expertise. Decision support system can assist decision making by asking relevant questions and explaining the reasons for adopting certain actions. Well-designed decision support systems emulate the reasoning processes used by experts to solve problems, and are popularly used in medicine, business management, design, and searching for natural resources where decision support system is expressly designed for the support of individual and collective decision making and provide support for decision makers mainly in semi structured and unstructured situations by bringing together human judgment and computerized information[1].

In this paper, there are (26) attributes and (4) classes. This paper is proposed to analyze patients' symptoms and generate rules using decision tree induction algorithm. The system provides their types of heart diseases and treatments to the patients by using the generated rules.

2. Related Work

One of the earliest medical expert systems that use CBR techniques is CASEY, deals with heart failure diagnosis researched by L. Gierl et al[6]. The system uses these steps: first it searches for similar cases, and it determines the process concerning differences and their evidences between a current and a similar case, and then a transfer of the diagnosis of the similar to the current case or – if the differences between both cases are too important – an attempt to explain and modify the diagnosis.

(Sotiris A Pavlopoulos, Antonis CH Stasis, Euripides N Loukis)[7] researched that new technologies like echocardiography, color Doppler, CT, and MRI provide more direct and accurate evidence of heart disease than heart auscultation. However, these modalities are costly, large in size and operationally complex and therefore are not suitable for use in rural areas, in homecare and

generally in primary healthcare set-ups. Furthermore the majority of internal medicine and cardiology training programs underestimate the value of cardiac auscultation and junior clinicians are not adequately trained in this field. Therefore efficient decision support systems would be very useful for supporting clinicians to make better heart sound diagnosis.

Currently, many researches have been pursued for cardiovascular disease diagnosis using ECG so far. In (Heon Gyu Lee, et.al, Kiyong Noh, Bum Ju Lee, Ho-Sun Shon and Keun Ho Ryu)[8] they extracted multi-parametric features by HRV analysis from ECG, data preprocessing and heart disease pattern classification method. They studied analyzes the clinical information as well as the time and the frequency domains of HRV, and then discovers cardiovascular disease patterns of patient groups. In each group, its patterns are a large frequency in one class, patients with coronary artery disease but are never found in the control or normal group. These patterns are called emerging patterns.

The system (Abdel-Badeeh M. Salem, Mohamed Roushdy and Rania A. HodHod)[9] uses the Case Based Reasoning methodology to develop a case-based expert system prototype for supporting diagnosis of heart diseases. 110 cases were collected for 4 heart diseases namely; mitral stenosis, left-sided heart failure, stable angina pectoris and essential hypertension. Each case contains 207 attributes concerning both demographic and clinical data. After removing the duplicated cases, the system has trained the set of 42 cases for Egyptian cardiac patients. Statistical analysis has been done to determine the importance values of the case features.

3. Background Theory

In this section, this paper introduces the background theories used to implement the Decision Support System.

3.1 Decision Support System

Decision support systems (DSS) are a subset of computer-based information systems . The general term 'computer-based information systems' is a constellation of a variety of information systems such as office automation systems, transaction processing systems, management information systems and management support systems[1].

A decision support system involves an interactive analytical modeling process for example, using a DDS software package for decision support may result in a series of displays in response to alternative what –if changes entered by a manger. Decision support system can provide intelligent access to relevant knowledge, and aiding the process of structuring decision. Instead, they use the DSS to find the information they need to help them make a decision. That is the essence of the decision support system concept[2].

A Decision Support System (DSS) can be described as a computer-based interactive human–computer decision-making system that:

1. supports decision makers rather than replaces them;
2. utilizes data and models;
3. solves problems with varying degrees of structure: non-structured(unstructured or ill-structure semi-structured semi-structured and unstructured
4. Focuses on effectiveness rather than efficiency in decision processes.

3.2 Decision Tree Induction Method

A decision tree is a flowchart –like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distribution. It is a method for approximating discrete valued target functions. The learned function is represented by a tree structure. It processes search of a completely expressive hypothesis space.

It builds the tree from the top down, with no backtracking. Information Gain is used to select the most useful attribute for classification. Decision Tree Induction method searches through the attributes of the training instances and extracts the attribute that best separates the given examples. Decision Tree Induction method stops if the attribute perfectly classifies the training sets. The algorithm uses a greedy search, that is, it picks the best attribute and never looks back to reconsider earlier choices. The objective of the decision tree algorithm is selecting which attribute to test at each node in the tree . For the selection of the attribute uses the concept of entropy[3].

3.2.1 Entropy

Entropy measures the impurity of an arbitrary collection of examples. For a collection S, entropy is given as:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad \ldots\ldots\ldots\ldots(1)$$

P_i is the probability that an arbitrary in D belongs to class C_i *Info(D)* is also known as the Entropy of *D.*

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D_j) \qquad \ldots\ldots\ldots\ldots(2)$$

The term $\frac{|D_j|}{|D|}$ acts as the weight of the j^{th} partition.

$Inof_A(D)$ is the expected information required to classify a tuple from *D* based on the partitioning by $_A$[4].

3.2.2 Information Gain

It measures the expected reduction in entropy. The higher information gain (IG), the more is the expected reduction in entropy.

$Gain(A) = Info(D) - Info_A(D)$ ……(3)

Where Values (A) is the set of all possible values for attribute A. First the entropy of the total dataset is calculated. The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy. The attribute that yields the largest Information Gain is chosen for the decision node[5].

3.2.3 Decision Tree Algorithm

Create a node N;
if *samples* are all of the same class, C **then**
return *N* as a leaf node labeled with the class *C;*
If *attribute-list* is empty **then**
return *N* as a leaf node labeled with the most common class in *samples;*
Select *test-attribute*, the attribute among *attribute-list* with the highest information gain;
Label node *N* with *test-attribute;*
For each known value a_i **of** *test-attribute*
Grow a branch from node *N* for the condition *test-attribute* = a_i ;
Let s_i be the set of samples in *samples* for which *test-attribute* = a_i*;*
If s_i is empty **then**
Attach a leaf labeled with the most common class in *samples;*
Else
attach the node returned by Generate decision tree (s_i, *attribute-list-test-attribute);*

4. System Overview

Medical diagnosis is considered to a significant intricate task that needs to be carried out precisely and efficiently. Heart disease was the major cause of causalities in most of countries . According to the record, heart disease kills one person in very short time.

We implement the diagnosis system for heart diseases by using decision tree induction method. This system has two main parts, administrator module and user module. Administrator can register new administrator registration, update and delete existing administrator. Administrator can insert, update and delete train data and test data. Administrator can extract decision tree and decision rules by applying the train data. In this system, there are 3000 train data and 1000 test data. Administrator can monitor decision rules, test data and system accuracy. Administrator can also enquiry heart disease by entering symptoms of these diseases.

User can view user information, disease rules statement and system accuracy. By inserting symptoms, the user can view heart diseases information.

There are over 100 types of heart diseases. This system provides only 4 types of heart diseases. The heart diseases defected by this system are Heart Failure, Infective Endocarditis, Coronary Heart Disease and Valvular Heart Disease. The heart diseases that we implement include (26) attributes (symptoms) and (4) classes (diseases). By using 3000 train data, we implement the diagnosis by applying decision tree induction. First, the system searches entropy for all train data. Then, it calculates information gain(IG) for each attribute. The maximum IG is denoted as the root node. The maximum attribute is removed from next calculation of IG. The system calculates entropies and IGs from left attributes repeatedly. Finally, the system generates the complete decision tree by using train data. And then by using decision tree, the system generates the decision rules and stores in the rule database. When the user enters the symptoms that he/she suffered, the system checks patients' symptoms with decision rules in the rule database. If patients' symptoms are matched with the disease rules in the rule database, the system displays the heart disease type and treatments according to the symptoms that he/she suffered. If patient's symptoms are not matched with the disease rule in the rule database, the system returns "This system is not able to diagnosis about this kind of Heart Disease, it may be other Heart Disease or other Disease" message will be appeared. To calculate the system's accuracy, the system has to test using 1000 test data. The system matches attributes of these test data with disease rules in the rule database. System accuracy is 84.4%.

5. Process Flow of the System

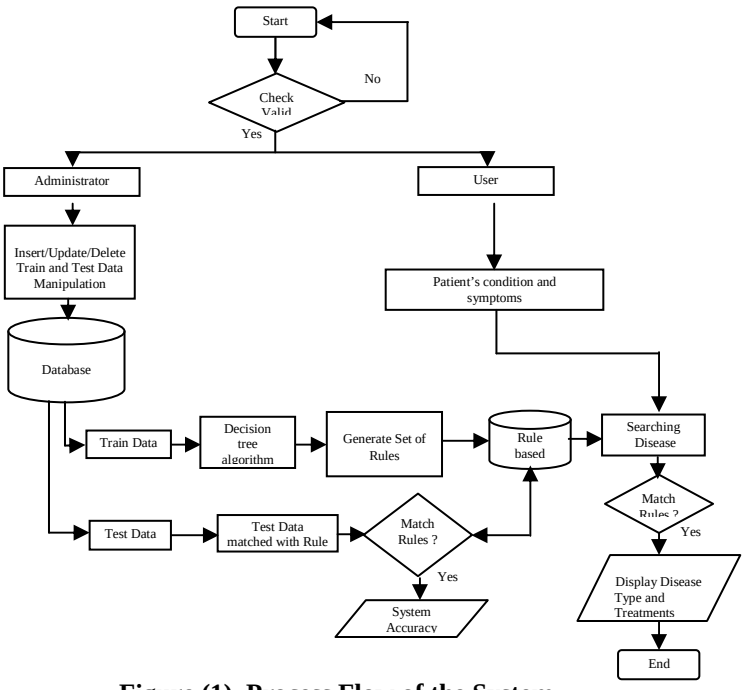

Figure (1). Process Flow of the System

6. Attribute and Class List of the system

In this system, there are 26 attributes and 4 classes.

No.	Attribute Name	Attribute Value
1	Chest pain aggravated by	1. exertion 2. emotion 3. exterme weather 4. none
2	Fever	1. low 2. high 3. none
3	Breathlessness	1. dyspnoea or fatigue 2. orthop or paroxysamul nocturnal 3. syncope 4. none
4	Gastrointestinal symptoms	1. abdominal pain 2. anorexia or nausea or vomiting 3. none
5	Lower extremities	1. sweating 2. edema 3. clubbing 4. none
6	Upper extremities	1. sweating 2. oslar node or splinter hemorrhage 3. cyanosis 4. none
7	Eye	1. pallor 2. jaundice 3. conjunctiva congestion 4. none
8	Neck	1. JVP increased 2. prominent carotid artery pulsation 3. none
9	Heart rate	1. tachycardia 2. bradycardia 3. none
10	Pulse rhythm	1. sinus 2. irregular
11	Pulse volume	1. small 2. high 3. normal
12	Pulse character	1. pulses atteran 2. collapsing pulse 3. no special character
13	Apex beat site	1. normal 2. displaced
14	Apex beat character	1. normal 2. tapping 3. heaving 4. thrill
15	Heart sound	1. 1+2 sound 2. 3 sound 3. 4 sound
16	Heart rhythm	1. sinus 2. triple 3. gallop
17	Murmurs	1. ESM 2. MDM 3. PSM 4. EDM 5. complex 6. changing 7. none
18	Repertory symptoms	1. cough 2. crepitating 3. basal crepitating 4. wheezing 5. none
19	Abdomen	1. enlarged tender liver 2. spleenomegaly
20	Urinary symptoms	3. ascities 4. none 1. oliguria 2. haematuria 3. ARF 4. none
21	Age	1. <=30 2. >=30 - <=40 3. >=40 - <=60 4. >60
22	Sex	1. male 2. female
23	Habit	1. alcohol 2. smoking 3. both 4. none
24	Underlying Disease	1. hypertension 2. diabetes mellitus 3. both 4. none
25	Past history of rheumatic	1. present 2. absent
26	Central chest pain	1. heaviness constriction in nature 2. dull or squeezing in nature 3. sharp stabbing in nature 4. none

Table (1) Attribute Table

No.	Age	Sex	Habit	Underlying Disease	Past history of rheumatic	Central chest pain	……	Class List
1.	2	1	4	1	2	3	…..	IE
2.	1	1	2	3	2	2	…..	HF
3.	1	2	3	4	1	3	…..	IHD
4.	2	1	1	4	2	2	…..	IE
5.	1	2	2	2	1	1	…..	RHD
6.	3	1	4	1	1	4	…..	HF
7.	2	2	1	3	2	1	…..	RHD

Table (2) Class Table

HF = Heart Failure
IE = Infective Endocarditis
IHD = Coronary Heart Disease
RHD = Valvular Heart Disease

7. System Accuracy

Classification Accuracy of a rule set is the ratio of the number of correctly classified objects from the test set and all objects from the test set. ID3 uses a set of test samples independent from the training set to estimate rule accuracy. This would result in an optimistic estimate of rule accuracy.

$$\text{Accuracy} = \frac{\text{Number of correctly classified objects}}{\text{Overall object from the test set}} \quad \ldots\ldots\ldots (4)$$

By testing 1000 data, 844 test records are correct. So, the accuracy of the system is 84.4%.

8. Conclusion

The decision support system uses the advantages of Decision Tree Induction method to build the knowledge-based system that can support the decision making in medical diagnosis. When the user inputs the symptoms, the system will give advice for disease. Human experts in medical field are frequently in great demand to diagnose the diseases. It is impossible for a person to consult a medical specialist whenever she feels a symptom of heart diseases. Thus, computer-based diagnostic system will play an increasingly importance in health care.

Diseases include many symptoms and are difficult to decide accurately which disease the patients have been suffered. Diagnosis for diseases should be combined with computer's ability and computer methodology to get accurate results. This paper is focused on developing architecture for the diagnosis of heart diseases by using decision tree induction method. The patients can know disease types accurately. If the estimation heart disease is correct, the patient and the doctor also get a lot of benefits. Therefore, diagnosis of heart diseases using Decision Tree Induction method is suitable and very useful for patients.

References

[1] J. R. Quinlan, Induction of decision trees, *Machine Learning*, 1, 1986, 81-106.

[2]J. Han and M. Kamber (2000), *Data Mining: Concepts and Techniques*, Academic Press, San Diego, CA.

[3] Jiawei Han and Micheline Kamber: *Data Mining :*

[4] Fredda Weinberg , Rule induction: "Ross Quinlan's ID3 algorithm", CIS 718X, Fall 2005, Professor Kopec

[5] Andrew Colin, Dr. Dobbs Journal, June 1996 "Building Decision Trees with the ID3 Algorithm",

[6] L. Gierl et al., "CBR in Medicine". In M. Lenz et al., *Case-Based Reasoning Technology, From Foundations to Applications*, ISBN: 3-540-64572-1, Springer, Berlin, 1998, pp. 273-297.

[7]. Sotiris A Pavlopoulos, Antonis CH Stasis, Euripides N Loukis "A decision tree – based method for the differential diagnosis of Aortic Stenosis from Mitral Regurgitation using heart sounds".

[8]. Heon Gyu Lee, Kiyong Noh, Bum Ju Lee, Ho-Sun Shon and Keun Ho Ryu, "Cardiovascular Disease Diagnosis Method by Emerging Patterns".

[9]. Abdel-Badeeh M. Salem, Mohamed Roushdy and Rania A. HodHod, "A CASE BASED EXPERT SYSTEM FOR SUPPORTING DIAGNOSIS OF HEART DISEASES" , *Computer Science Department, Faculty of computer & Information Sciences, Ain Shams University, Abbassia, Cairo, Egypt*