Risk Level Prediction for Heart Disease using Decision Tree Induction

Naing Naing Khin, Win Thein Lwin
Computer University(Sittwe)
naingk86@gmail.com, wintheinlwin007@gmail.com

Abstract

Heart Disease was the major cause of causalities in most of the countries. According to the medical records, heart disease kills one person in very sort time. Classification and prediction are the forms of data analysis that can be used to extract models for important classes or to predict future data trends. In this paper, decision tree induction algorithm is used to classify the risk level for heart disease. 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. This system generates the understandable rules for user and estimates the accuracy for classifier. Depending on the attribute values of the data set, this system can classify the risk level of heart disease whether it is in serious or normal conditions for patients. Thus, the user can test his or her medical check concerned with their heart. Moreover, the system can provide the classifier accuracy by using Holdout Method.

1. Introduction

The effective identification of information from a large collection of data has been on a steady increase recently. Medical diagnosis is considered to a significant intricate task that needs to be carried out precisely and efficiently. Data mining is the process of using tools such as classification, association rule mining, clustering, etc. Decision tree induction algorithm is one on the most popular algorithms in the mining classification. The primary intent of the system is to design and develop an efficient approach for extracting decision rules, which are important for heart disease, from the heart disease database. The heart disease data set consists of attributes to classify the risk level of the heart disease into three classes. These classes are the risk level of heart disease: No Risk, Medium Risk, and High Risk. No Risk is the condition of patient is normal. For Medium Risk, the patient has to care for his heart but not serious. And a patient with High Risk must care for him and should make the detail medical check in experienced doctors. A majority of areas related to medical services such as prediction of effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data also make use of Data Mining methodologies. This can be attained by the making use of proper computer-based information and prediction systems. Data mining is the central process of knowledge discovery that is the transformation of data into knowledge for decision making. The proposed system aims to utilize decision tree induction algorithm and extracted the correct rules for patients.

2. Related Work

Decision tree classifier is a simple yet widely used classification techniques. Minos.G, Dongjoon.H, Rajeev R. and Kyuseok S. [10] used decision tree for efficient algorithms for constructing decision tree. Quinlan, J. [6] developed decision tree induction algorithm in machine learning. And Kamber, L.Winstone, W.Gong, S.Cheng, J. Han [8] applied decision tree induction algorithm for effective classification in data mining in 1997. K.Viikki, 1, 4 Martti Juhola, 1 I1mari Pyykk0, 2 and Pekka Honkavaara3[7] described about Decision Tree Induction algorithm for evaluating data suitability. Myo Myo Than Naing [11] used Decision Tree Induction Algorithm in Decision Making for Poultry Diseases. Soe San Oo [12] analyzed decision tree induction algorithm for Diagnosis of Acute Diarrhoea in Children. There are so many related works by using decision tree induction.

3. Data Mining

Data Mining is the process of discovering meaningful, new correlation patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition techniques as well as statistical and mathematical techniques. The primary goals of data mining in practice tend to be prediction and description. Data mining serves as an essential step in the process of knowledge discovery in the databases. Diverse fields such as marketing, customer relationship management, engineering, medicine, crime analysis, expert prediction, Web

mining, and mobile computing, besides others utilize data mining. It combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases.[5] Data mining is used for a variety of purposes in both private and public sectors such as in banking, insurance, medicine, and the development of enhanced search-related techniques. In medical community, data mining is used to predict the diagnosis of diseases for effective solutions.[2]

3.1 Classification

Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. The goal of classification is to describe the data or to make future prediction. Many classification methods have been proposed in machine learning, expert systems, statistics and neurobiology. Data mining community inherits the classification techniques are applied in real world problems. There are many practical situations in which classification is of immense use. Basic techniques for data classification are decision tree induction, Bayesian classification and Bayesian belief networks, and neural networks. [1]

4. Decision Tree Induction

Decision tree learning is one of the most popular classification algorithms. A decision tree is a tree structured prediction model where each internal node denotes a test on an attribute, each outgoing branch represents an outcome of the test and the leaf nodes represent classes or class distributions.[6] A decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node, which provides the classification of the instance.[12]

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node, which provide the classification of the instance. The information gain measure is used to select the test attribute at each node in the tree. Such a measure is used to as an attribute selection measure or a measure of goodness of split. The attribute with the highest information gain is chosen as test attribute for the current node.[3],[4]

Let S be a set consisting of a s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, Ci (for i=1,…,m). Let si be the number of samples of S in the class Ci. The expected information needed to classify a given sample is given by

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (4.1)$$

where p_i is the probability that an arbitrary sample belongs to the class C_i and is estimated by s_i/s. Note that a log function to the base 2 is used since the information is encoded in bits.

Let attribute A have v distinct values, $\{a_1, a_2, \ldots, a_v\}$. Attribute A can be used to partitions S into v subsets, $\{S_1, S_2, \ldots, S_v\}$, where S_j contains those samples in S that have value a_j of A. If A were selected as the test attribute (i.e., the best attribute for splitting), then these subsets would correspond to the branches grown from the node containing the set S. Let s_{ij} be the number of samples of S in class Ci in a subset Sj. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = -\sum_{j=1}^{v} \frac{sij + \ldots.. + smj}{s} I(s1j, \ldots, smj) \qquad (4.2)$$

The term $\frac{sij + \ldots.. + smj}{s}$ act as the weight of the jth subset and is the number of samples in the subset(i.e., having the value aj of A) divided by the total number of samples in S. The smaller entropy value, the greater the purity of the subset partitions. Note that for a given subset Sj,

$$I(s_{1j}, s_{2j}, \ldots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij}),$$

Where $p_{ij} = \frac{sij}{|sj|}$ and p is the the probability that a sample in Sj, belongs to class Ci. The encoding information that would be gained by branching on A is

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A) \qquad (4.3)$$

The algorithm computes the information gain of each attribute.

4.1 Decision Tree Construction

Decision Tree uses the gain ratio criterion selects, from among those attributes with an average or better gain, the attribute that maximizes the ratio of its gain divided by its entropy. The algorithm is applied recursively to form sub-trees, terminating when a given subset contains instances of only one class. It is an approximation discrete function method and can yield lots of useful expressions. The decision tree induction algorithm is as follow:

1. Check for base cases
2. For each attribute *a*
    1. Find the normalized information gain from splitting on *a*
3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recurse on the sublists obtained by splitting on *a_best*, and add those nodes as children of *node*

5. Overview of the Proposed System

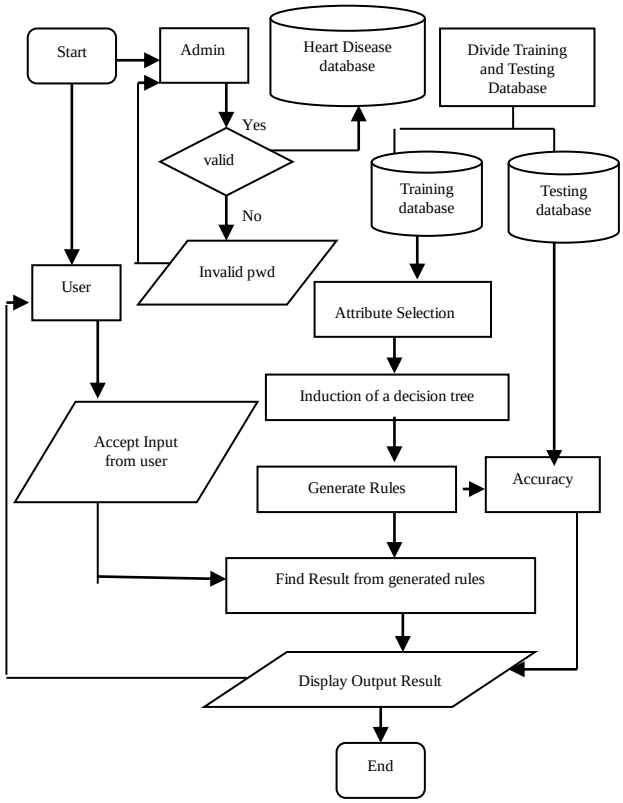

Fig:1 System Flow of the Heart Disease Prediction System

6. System Implementation

The heart disease data set consists of attributes to classify the risk level of the heart disease and three classes. These classes are the risk level of heart disease: No Risk, Medium Risk, and High Risk. The user who wants to know the risk level of their heart can test with the given attributes in the heart disease data set. The system compute the information gain of each attribute by using Equation 1 and compute the entropy or expected information of each attribute by using Equation 2. By using equation, the highest information gain among the attribute is selected and created as root node as in fig:2. Finally, the system can generate the decision tree by using decision tree induction algorithm. The knowledge represented in decision trees can be extracted and represented in the form of classification IF-THEN rules which are described in fig: 3. After generating the rules, the system can classify the risk level for heart disease. The system can estimate the classifier accuracy test result using Holdout method. The new patient can also test the condition of their heart in this system as in figure: 4. This system is developed on Microsoft Access 2003 for database and implemented using Java programming language, Jdk 1.6.

6.1 Attribute Information

The attributes which are important for heart disease prediction system are described. This system can be used attributes selection for patient data. [11] In training data set 300 patients records to evaluate the performance of the classification system, These record consists of 10 attributes for three classes. The detailed description of the parameters and their corresponding values are given as follows:

Table 1 : Heart Disease parameters with corresponding values

No.	Parameters	Values
1.	Age	<=30 >30
2.	Sex	Male Female
3.	Blood Pressure	Normal Low High
4.	Smoking	Never Past Current
5.	Heart Rate	Normal Bradycardia Tachycardia
6.	Serum Cholestrol	>240 mg/dl <240 mg/dl
7.	Central Chest Pain	No >15 min <10 min
8.	Electrocardiography	Normal Ischaemic Changes
9.	Condition of Vessels Status	Normal Reversible Irresversible
10.	Coronary Angiography	Normal >50% of CA <50% of CA

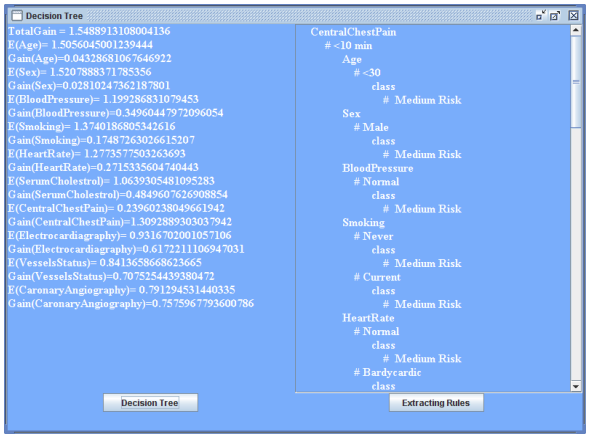

Fig:2 Attribute Selection and Decision Tree Form

We have calculated the information gain by using the attributes from data set and generate decision tree from the resulted information gain measure.

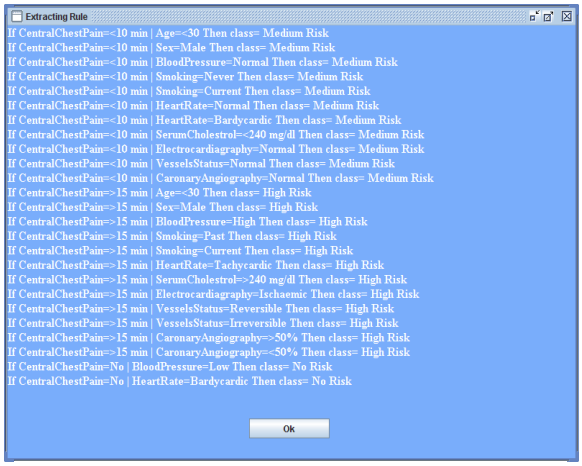

Fig: 3 Decision rules form of the system

After generating decision tree, we can extract the classification rules as in above figure 3.

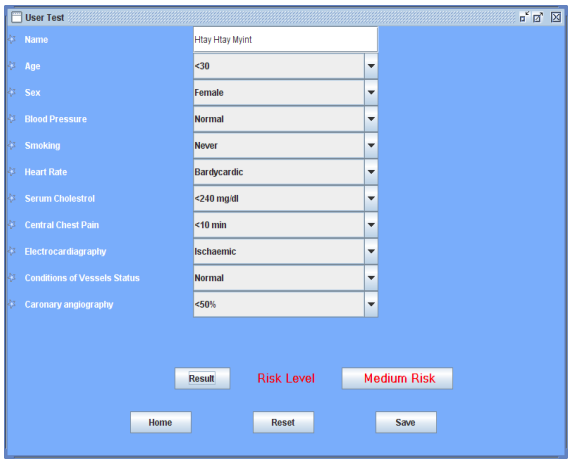

Fig: 4 User Test form of the Heat Disease Prediction System

We are going to choose the attributes for new patient and click on Result to find out the result button, then the system displays the risk level of heart disease and give detail information about the risk level.

7. Estimating Classifier Accuracy

Estimating the classifier accuracy is the necessity for the system to know how much a given classifier correct and accurate. We have use about 300 patients records as data set to evaluate the performance of the classification system. Holdout method estimates the classifier performances in this system. The given data are randomly partitioned into two independent sets, a training set and a test set. Typically, two third of the data are allocated to the training set and the remaining one third is allocated to the test set.[9]

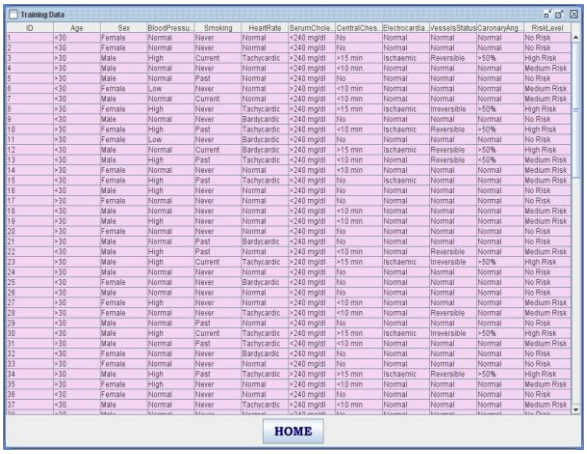

Fig: 5 Data set of the System

The separate class accuracy of i-th single decision class is calculated as:

$$ACC_i = \frac{T_i}{T_i + F_i} \qquad (7.1)$$

Where T stands for "true" cases(i.e correctly classified objects) and F stands for false cases(i.e not correctly classified objects) and the average accuracy over all decision classes is calculated as

$$ACC_i = \frac{1}{v} \cdot \sum_{i=1}^{v} \frac{T_i}{T_i + F_i} \qquad (7.2)$$

Where v represents the number of decision classes. Heart Disease database system use 100 patients records for testing data. There are 32 patients for No Risk, 38 patients for Medium Risk

and 30 patients for High Risk. The system can estimate the classification accuracy as follow:

Table 2: Accuracy Test Result

Classes	True	False	Accuracy
No Risk	32	0	1
Medium Risk	37	1	0.97
High Risk	28	2	0.93

ACC = 1/3(1+0.97+0.93) = 0.96

Thus, estimating accuracy for the system is 96 %.

8. Conclusion

This system focus on the classification rules mining that based on decision tree induction algorithm. This system is intended to develop an effective solution for heart disease. Risk level for heart disease can be generated as decision rules by using decision tree induction algorithm and can evaluate estimate classifier accuracy by using Holdout method. The capability and practical use of this system was proved in testing the heart disease patients. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Another use of decision trees is as a descriptive means for calculating conditional probabilities. The system can determine the risk level of heart disease for patients to know their heart conditions. This system is mainly support for risk level for heart disease to make a valid prediction. User can get specific result and detail information for their medical check.

9. References

[1] Aijun ,A "Classification Methods",
New York University , Canada .
http://en.wikipedia.org/wiki/Granular-computing

[2] D.A. Keim, "Knowledge Discovery and Data Mining New Port Beach , USA, 1997".

[3] http://decision tree learning applet.htm
[4] http:// decisiontree.net

[5] Han, Jiawei & K.Micheline "Data Minining Concepts and Techniques".

[6] J. R. Quinlan, C4.5: "Programs for Machine Learning, Morgan Kaufmann, 1993".

7] K.Viikki,1,4 Martti Juhola, 1 I1mari Pyykk0, 2 and Pekka Honkavaara3, "Evaluating Training Data Suitability for Decision Tree Induction", Journal of Medical Systems, Vol 25, No.2, 2001.

[8] Kamber, L.Winstone, W.Gong, S.Cheng, J.Han applied decision tree induction algorithm in "Generalization and Decision Tree Indction: Efficient Classification in Data Mining", 1997.

[9] Khaing Nay Kyi, "Classification of Industry Test by Decision Tree Induction".

[10] Minos.G, Dongjoon.H, Rajeev R. and Kyuseok S. "Efficient Algorithms for Constructing Decision Tree with constraints".

[11] Myo Myo Than Naing, "Decision Making for Poultry Diseases using Decision Tree Induction Algorithm".

[12] Soe San Oo, "Diagnosis of Acute Diarrhoea in Children by using Decision Tree Induction".