© IJCIRAS | ISSN (O) - 2581-5334
May 2020 | Vol. 2 Issue. 12

CORONARY ARTERY DISEASE PREDICTION SYSTEM BY USING DECISION TREE (DT) CLASSIFIER

Yin Yin Htay[1], Ya Min[2]

Faculty of Information Science, Computer University (Magway), Myanmar

Abstract

Today, the diagnosis of diseases is a vital and intricate job in medicine. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. Regrettably all doctors do not possess expertise in every sub specialty and moreover there is a shortage of resource persons at certain places. In this situation, an automatic medical diagnosis system is beneficial by bringing all of them together. For this diagnosis system, many classifiers are essential and needed for disease classification. So, this system is proposed as the coronary artery disease prediction system. For disease prediction, this system uses the decision tree (DT) classifier. This system is useful for medical domain.

Keyword: Decision Tree, Classification, Disease

1.INTRODUCTION

Healthcare industry generates the large amount of data about patient, disease diagnosis etc. However, there is a lack of effective analysis tools to discover hidden relationships in data. Data mining provides a set of techniques to discover hidden patterns from data. Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. A knowledge discovery process includes data cleaning, data integration, data selection, data mining, pattern evaluation and knowledge presentation.

Data mining system can be classified according to the kinds of databases mined, the kinds of knowledge mined, the techniques used or the applications adapted. Data mining is the process of classification, association rule mining, clustering, etc. K-nearest neighbor(KNN), naive bayesian (NB) and decision tree (DT) classifiers are the most popular algorithms in the mining classification. Major challenge facing Healthcare industry is quality of service. Quality of service implies diagnosing disease correctly and provides effective treatments to patients. Poor diagnosis can lead to disastrous consequences which are unacceptable.

So, this system is proposed to predict whether the patient is having coronary artery disease or not by using decision tree (DT) classifier that is data mining technique. After classifying according to DT classifier, this system calculates the accuracy of DT classifier. In the remote areas like rural regions or country sides, the proposed system is also a user friendly, scalable and reliable system that can be implemented to imitate like human diagnosis expertise for treatment of heart disease.

2. RELATED WORK

In 2016, M. Panwar, A. Acharyya and R. A. Shafik [1] presented a new methodology based on novel preprocessing techniques, and K-nearest neighbor classifier. The effectiveness of the proposed methodology is validated with the help of various quantitative metrics and a comparative analysis, with previously reported studies using the same UCI dataset focusing on pima-diabetes disease diagnosis.

In 2016, D. VijayaKumar and V. J. R. Krishniah [2] used decision tree classification model for diagnosis of three brain diseases namely ischemic stroke, hemorrhage and hematoma, and tumor. This system helped the physicians to identify the type of human brain hemorrhage and hematoma and the type of brain tumors for further treatment.

In 2017, N. R. Gorrepati and N. R.Uppala [3] compared the performances of different classifiers on diagnosis of the Erythemato-Squamous disease. The classifiers

examined here are support vector machine, discriminant classifier, K-nearest neighbor and decision tree. They have performed their analysis with two well-known multiclass implementation techniques. They demonstrated that the most reliable performance has been achieved using support vector machine classifier.

3. CLASSIFICATION

For decision making procedure, data mining is a very favorable and constructive method. Classification is a very simple and mostly used data mining technique. Knowledge of training data is mandatory for understanding of classification. There are two phases of classification procedure:

- Development of a model for training
- Evaluating the model using testing data.

For classification, there are various classifier. Bayesian classifier uses frequentist technique. The essence of frequentist technique is to apply probability to data. Bayesian calculations go straight for the probability of the hypothesis. K-nearest neighbor is a non parametric method which depends on the use of distance measurement. All available cases can be stored in it and whenever a new case entered, it can be classified based on the distance function. According to the decision tree based classifier, there is a requirement of construction of a tree to model classification process [4].

4. DECISION TREE CLASSIFIER

In the decision tree (DT) classifier, information gain measure (entropy) is used to select the test feature at each node in the decision tree. This classifier is precise about classification because it is based on the entropy logic. DT algorithm is as follows:
Algorithm : Generate_decision_tree.

- create a node *N*;
- if *samples* are all of the same class, *C* then
  return *N* as a leaf node labeled with 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;

4.1. Information Gain Measure

Information gain measure is used to select the test feature at each node in the tree with the highest information gain. Let s_i be the number of samples of S in class C_i. The expected information needed to classify a given sample is given by

$$I(s_1,s_2,\ldots,s_m) = \sum_{i=1}^{m} p_i \log(p_i) \quad (1)$$

where p_i is the probability that an arbitrary sample belongs to C_i and is estimated by s_i/s. Let attribute A have v distinct values, $\{a_1, a_2,\ldots, a_v\}$. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j},\ldots,s_{mj}}{s} I(s_{1j},\ldots,s_{mj}) \quad (2)$$

For a given subset S_i,

$$I(s_{1j},s_{2j},\ldots,s_{mj}) = \sum_{i=1}^{m} p_{ij} \log(p_{ij}) \quad (3)$$

Encoding information would be gained by branching on A:

$$Gain(A) = I(s_1,s_2,\ldots,s_m) - E(A) \quad (4)$$

The feature with the highest information gain is chosen as the test feature for given set S [5, 6].

5. PROPOSED SYSTEM DESIGN

In this system, the user must first put the patient information (the patient suffered disease symptom). For classification, this system extracts the training coronary artery disease data into the system. Then, this system classifies the coronary artery disease stage by using decision tree (DT) classifier.
After classifying, this system produces and displays the coronary artery disease stage. Then, this system

measures the processing time of each classifier and calculates the accuracy of each classifier by using Holdout method. Finally, this system displays the result to the user. Proposed system design is shown in Figure 1.

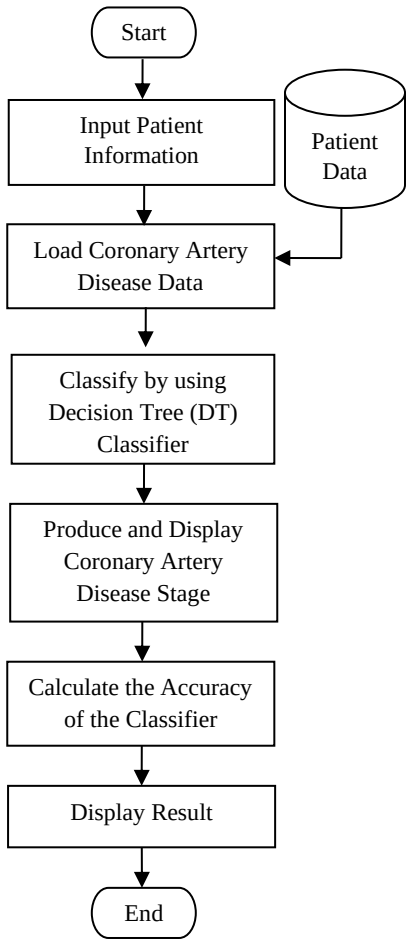

Figure 1. Proposed System Design

6. EXPLANATION OF THE SYSTEM

For classification, this system uses the coronary artery disease dataset. In this dataset, there are 13 (symptoms) attributes. As a sample, this system uses 10 records that are obtained from 10 patients who suffer coronary artery disease. Coronary artery disease dataset includes five class levels that are Normal (N), Level I (I), Level II (II), Level III (III) and Level IV (IV). Sample coronary artery disease dataset is shown in Table I.

Table 1. Coronary Artery Disease Dataset

Age	Sex	Chest Pain Type	Trestbps	Chol	Fasting Blood	Rest ECG	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Class
63	1	4	0	260	140	1	112	1	3	2	0	0	II
44	1	4	0	209	130	1	127	0	0	0	0	0	N
60	1	4	0	218	132	1	140	1	1.5	3	0	0	II
55	1	4	0	228	142	1	149	1	2.5	1	0	0	I
66	1	3	1	213	110	2	99	0	1.3	2	0	0	**N**
66	1	3	0	0	120	1	120	0	0.5	1	0	0	N
65	1	4	1	236	150	1	105	1	0	0	0	0	III
62	1	3	0	0	180	1	140	1	1.3	2	0	0	N
60	1	3	0	0	120	0	141	1	2	1	0	0	III
60	1	2	1	267	160	1	157	0	0.5	2	0	0	I

The patient input coronary artery disease symptom into system. The inputted information includes age (60), sex (1), chest pain type (4), trestbps (0), chol (260), fasting blood sugar (140), restECG (1), thalach (140), exang (1), oldpeak (1.5), slope (3), ca (0) and thal (0). Then, this system calculates and classifies the coronary artery disease stage that the patient suffers.

6.1. Decision Tree Classification Process

To produce the decision rules, this system calculates the gain for each attribute. The attribute is chosen as root node if this attribute has highest gain result. In this sample. This system obtains the decision tree after finishing second iteration. According to the decision tree, this system produces the decision rules to produce the result. Decision tree is shown in Figure 2.

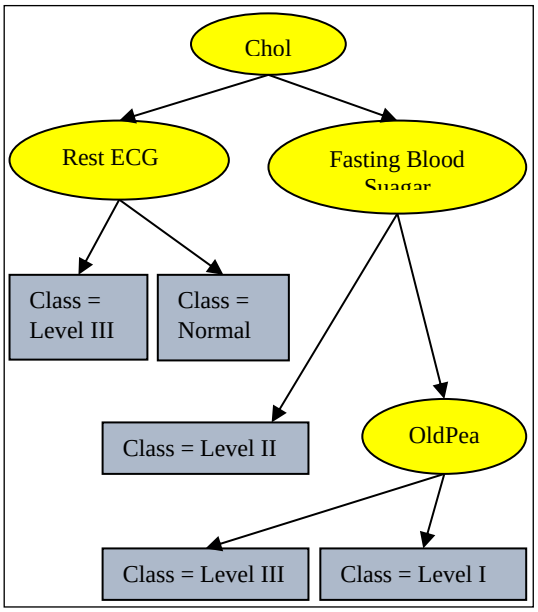

Figure 2. Decision Tree

Decision rules are generated from the decision tree. These rules are as follows:

- Rule 1 is {IF "Chol <= 163.1" AND "RestECG <= 1" THEN "Class= Level III"}.
- Rule 2 is {IF "Chol <= 163.1" AND "RestECG > 1" THEN "Class= Normal"}.
- Rule 3 is {IF "Chol > 163.1" AND "Fasting Blood Sugar <= 141" THEN "Class = Level II"}.
- Rule 4 is {IF "Chol > 163.1" AND "Fasting Blood Sugar > 141" AND "OldPeak <= 1.5" THEN "Class = Level III"}.
- Rule 5 is {IF "Chol > 163.1" AND "Fasting Blood Sugar > 141" AND "OldPeak > 1.5" THEN "Class = Level I"}.

According to the Rule 3, the coronary artery disease level that the patient suffered is **"Level II"**.

7.EXPERIMENTAL RESULT

This system uses the coronary artery disease dataset from the UCI website. This system is tested 500 records. By using hold out method, this system calculates the accuracy of each classifier. The experimental result of the system is shown in Table 2.

Table 2. Experimental Result of the System

Testing Data	Accuracy (Correct Rate)
150	97%
220	92%
270	89%
300	87%
330	83%

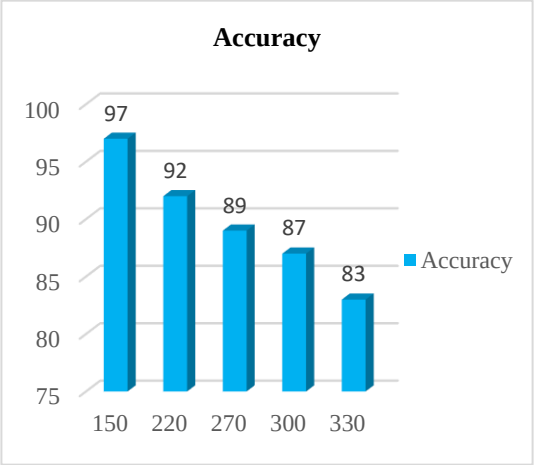

Figure 3. Accuracy of the System

8.CONCLUSION

This system is proposed an effective coronary artery disease prediction system by using decision tree classifiers. This system points out the decision tree classier that can quickly produce the coronary artery disease result. Finally, this system is helpful for practice to confirm his/ her decision during coronary artery disease prediction.

REFERNCES

[1] M. Panwar, A. Acharyya and R. A. Shafik, "K-Nearest Neighbor Based Methodology for Accurate Diagnosis of Diabetes Mellitus", IEEE, 2016.

[2] D. VijayaKumar and V. J. R. Krishniah, "An Automated Framework for Stroke and Hemorrhage Detection using Decision Tree Classifier", IEEE, 2016.

[3] N. R. Gorrepati and N. R.Uppala, "Comparative Performance Analysis of Different Classifiers on Diagnosis of Erythmato-Squamous Diseases", International Conference on Innovations in Information, Embedded and Communication Systems,

IEEE, 2017.
[4] C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification", International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, pp. 243-247, 2019.
[5] H. Jiawei and K. Micheline, "Data Mining Concepts and Techniques", Simon Fraser University, United States of America, 2001.
[6] M.S. Basarslan and I. D. Argun, "Classification of A Bank Data Set on Various Data Mining Platforms", IEEE, 2018.