IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-04, Issue-02 E-ISSN: 2347-2693

Comparison of Classification Techniques for Heart Health Analysis System

Karthika Jayprakash[1*],Nidhi Kargathra[2], Pranay Jagtap[3],
Suraj Shridhar[4]and Archana Gupta[5]

[1*,2,3,4,5]*Department of Computer Engineering, KJ Somaiya College Of Engineering, India,*

www.ijcseonline.org

Received: Jan /21/2016 Revised: Feb/04/2016 Accepted: Feb/15/2016 Accepted: Feb/29/2016

Abstract—Heart disease diagnosis is a difficult task which requires utmost accuracy. This accuracy is achieved through knowledge and experience in the field of medicine. This paper describes a heart diagnostic system which analyses several health parameters and medical test results to predict absence or presence of heart disease in terms of artery narrowing in the patient. The proposed system described in the paper is a computer based application which uses the patient information related to the various health parameters which govern the procedure of diagnosis for a heart disease. The system also relies on pre-processing of data. The system described in this paper has been created with a view to assist doctors and medical staff in diagnosis of heart related problems. The application also facilitates self-diagnosis for the common man. This paper deals with the internal functioning of the system which is based on the data mining techniques for classification. Few algorithms for classification based mining and association based mining of data and their comparisons have been incorporated in the paper. Classification is a data mining method used to classify data into pre-defined class labels. For instance, classification can be used to anticipate the weather on a specific day . Famous grouping procedures incorporate decision trees and neural systems.

Keywords—Classification;Data mining;ID3;Naïve Bayes

I. INTRODUCTION

Data mining is a used for analyzing large amounts of data in order to discover valuable patterns. Data mining hence automates the process of information discovery. Knowledge discovery in database is the procedure of finding out hidden patterns from data which is at a lower level [6]. Data mining is used in a large and varied number of applications. Large number of sciences, today, makes use of data mining techniques. Different application areas that make use of data mining are- Medical care, Finance, Economics, Telecommunication, Sales, Marketing, Recommendation Analysis and Fraud Detection.Data mining can also help in modeling future plans and strategies on the basis of past historical data. It is seen that in past few years the amount of company data, medical data and different types of data have grown. Human services industry today produces a lot of complex information about patients, clinical assets, diseases, electronic patient records, and so forth. The data mining strategies are exceptionally valuable to settle on restorative choices in curing diseases. We propose to build up a strategy to anticipate heart ailments with the assistance of indications utilizing data mining strategies. We intend to discuss various approaches of data mining on the basis of how accurately an approach will help in implementing a reliable heart ailment prediction system. This paper describes the advantage of using the method of classification for implementing a medical system. The importance of selecting the proper algorithm for classification has been emphasized as the algorithm impacts the accuracy of the system which is the key factor for a medical diagnostic system.

II. RELATED WORK

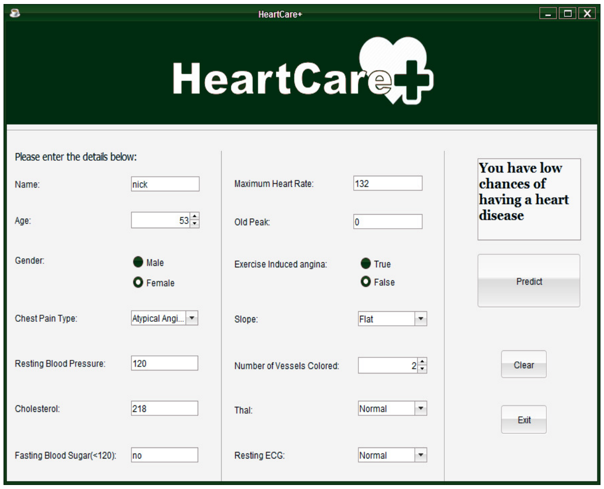
- An amalgam model is used model for classifying a PIDD (diabetic) database. It combines k-means algorithm and k-nearest neighbor i.e. KNN .Hence here pre-processing does not happen in a single step but it happens in multiple steps[1,8].
- In [2], the binning algorithm was implemented to deal with the noisy data. It is used to for data smoothing in grey relative analysis. It is also used to fill up the null values. It achieves this by building a function of grey relative coefficient for each null value and filling it up with the solution returned by the function. Another purpose is to find the outliers. This method is an application of grey systems in data cleaning.
- In [4], various data mining techniques are analyzed on database containing information of heart disease. Thesetechniques are: neural networks, decision induction trees, Naive Bayes, associative, classifiers, genetic algorithm.
- In [5], classifiers that are based on tree induction are used and their performance is analyzed. The parameters for performance are accuracy and time complexity.

III. OVERVIEW OF THE SYSTEM

The user enters his information in the GUI.(Fig 1).

The system recieves the different attribute values from the user . Pre-processing is applied on the data received and then data mining classification techniques is applied on the cleansed data and the result is given to the user.(Fig 2)

Few basic iterative steps that have been defined for data mining are- data preprocessing, integration of data from various sources, selection of the data required,transforming the data according to the requirements odf the target data, data mining,evaluation of patterns to select the useful ones and representation of knowledge.For our system we have implemented the data mining step.



Please enter the details below:

Name: nick

Age: 53

Gender: ● Male ○ Female

Chest Pain Type: Atypical Angi...

Resting Blood Pressure: 120

Cholesterol: 218

Fasting Blood Sugar(<120): no

Maximum Heart Rate: 132

Old Peak: 0

Exercise Induced angina: ● True ○ False

Slope: Flat

Number of Vessels Colored: 2

Thal: Normal

Resting ECG: Normal

You have low chances of having a heart disease

Figure 1: System Interface

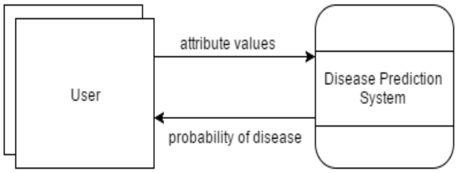


Figure 2: Basic block diagram

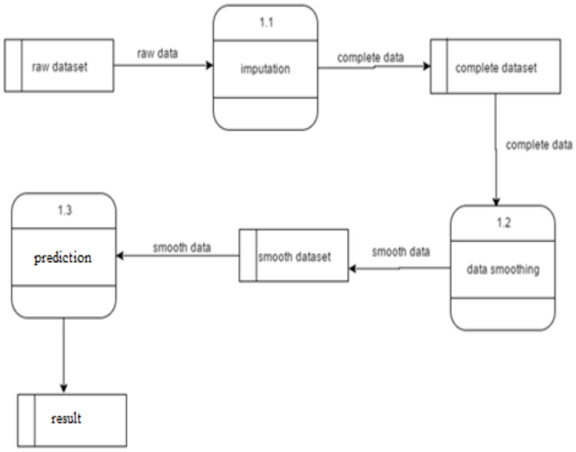


Figure 3:System Components.

The components of the system are:

A) Pre-processing
B) Classification
C) Results and Solutions
D) Database Updation

A. Pre-processing of data:

i. Imputation: Data consists of missing values due to data entry problems,etc.Missing values causes hindrances in data analysis.Hence, it is essential to rectify the incomplete data set by insertion of missing values using Knn algorithm.(Fig.3 module 1.1)

ii. Data Smoothening: Outliers are nothing but anomalies or noisy data that affect the accuracy of the system.In order to increase the accuracy of the system smoothening of data is doneto remove the ouliers.This is implemented using binning algorithm.(Fig.3 module 1.2)

B. Classification:

The system gives a yes or no result to the user i.e. our system contains two classes with labels : yes and no.In order to classify the given user query into these predefined classes, classifiers are used. ID3 is the classifier that is implemented by us.(Fig.1b module 1.3)

C. Results and Solutions :

Apart from providing the results of the query,our system also provides solution to the client by fetching articles from trusted websites using JSoup libraries.

D. Database Updation:

The system updates the database on the basis of new data values encountered. This addition of data is done to increase the accuracy of the system.

The input to the system is given in the following data description.

Data description: The datasets are taken from [8]. There are 14 attributes listed as follows:

1. age: age
2. gender:
3. ch_pain: angina type
4. rest_bp: resting blood pressure (in mm/Hg)
5. choles: cholstrol in miligram/dl
6. bs_f: blood sugar(fasting)
7. rest_ecg: resting electrocardiographic
8. thal_ach: max rate of the heart
9. ex_ang: angina induced due to exercise (one: yes; zero: no)

10. old_peak = ST decrease induced due to exercise

11. slope: (one:upward slope; two:no slope; three: downward slope)
12. ca: number of major artery
13. thale: 3 = normal; 6 = fixed defect; 7 = reversible defect
14. number: result of the system
    - 0: less than 50% narrowing in major artery radius
    - 1: greater than 50% narrowing in major artery radius

IV. PRE-PROCESSING METHODS

A) *Imputation*

K-NN (k-nearest neighbors) [1,7] algorithm is very useful for filling up the missing values that exists in the data set. K-NN uses Euclidean distance formula to calculate the distances. It then finds out the 'k' nearest neighbors that is the k number of cases in the training data set which give the lowest distances. Fine tuning can be done by varying the value of 'k'. We have taken k as 3. The next stage involved in K-NN algorithm is the response stage or voting. Here each of the 3 cases i.e.all three neighbors vote for valuesthat ought to be filled in the missing space. The value with highest votes is predicted as the missing value and filled. K-NN has a good accuracy however its speed decreases with increase in data sets. Prediction of heart disease is a critical task and requires high precision. It is preferable to have higher accuracy over speed. Hence, K-NN is a good algorithm used for imputation. Grey relative analysis approach can also be used for data pre-processing [2].

B) *Data Smoothing*

Binning algorithm is used for smoothing data. This is done immediately after imputation. Raw data after imputation is noisy i.e. it contains errors or outliers. Outliers are nothing but exceptions which affects the accuracy. In order to get rid of outliers, data smoothing is done. In binning algorithm, the first step is to fetch the given data and sort it. This sorted data is then separated into different bins of fixed sizes. We have implemented a size of 5. Mean of each bin is calculated and all the data stored inside the bin is replaced by this mean of the bin. An output from binning gives a smooth data without outliers.

Other method for performing binning is 'Binning by bin boundary' wherein all the values in the bin are replaced by either lower boundary values or upper boundary values of the respective bin depending upon the proximity of the value from the boundary.

In grey scale systems, binning is implemented to reduce the affect of outliers on the accuracy i.e. data smoothing is done.

V. CLASSIFICATION METHODS

A) *ID3*

ID3 (Iterative Dichotomiser 3)[3] is used to generate a decision tree from the dataset into pre-defined class labels. It uses entropy of each attribute to select the best attribute for the tree generation. Entropy gives the amount of randomness or impurity. An entropy value close to 1 indicates high level of impurity which is not desirable while values close to 0 indicate low impurity. This is used to identify gain of the attribute. Gain indicates how much information we gained using this attribute for the split. Attribute having maximum gain is chosen as the root for the sub tree.It is a top down approach wherein the attribute with highest gain takes the role of root for the sub part of the tree and this process iteratively continues until one of the two stopping criteria are met. The two stopping rules are: The entropy of an attribute gives the value 0; All the attributes are used up and the tree is complete after which voting will be done to pick the correct class. Handling of continuous attributes is supported by C4.5. In case of ID3, domain knowledge is used to set basic thresholds for data discretizationto convert the numeric data into nominal. Domain knowledge along with classification algorithm, ID3, is used for predicting the outcome.

i. Entropy

It is basically the measure of how impure some self assertive accumulation of data objects is. Say, there are about n diverse values, then the entropy s with respect to the classification described is defined as

$$\text{Entropy}(s) = \sum - P_i \log_2 P_i \qquad (1)$$

Where P_i is the likelihood of S fitting in with class i.

ii. Information gain

Information gain puts a value to the amount of information that is useful from the attribute if it is used for splitting. ID3 uses gain information to build a tree.

$$G(D, S) = H(D) - \sum P(D_i)H(D_i) \qquad (2)$$

B) *Naive Bayesian*

Naive Bayesian algorithm is based on the Bayes' posterior probability theorem i.e. P(X/H) where X is an attribute and H is a predefined condition. Bayesian classifiers are used to classify the data into predefined class labels [3,4].

In our system, given a query containing the values of the various attributes. Naive Bayes attempts to classify it into the two classes i.e. yes or no (yes meaning the person has a heart disease).

Here two variables a positive response and a negative response are defined.

$$p = P(X1/yes) * P(X2/yes) *\ldots.* P(X13/yes)$$

$$n = P(X1/no) * P(X2/no) *\ldots.* P(X13/no) \qquad (3)$$

whereXi is the value of the ith attribute.

If p>n then it is classified into the yes class else, it is classified into the no class

VI. Comparison of Algorithms

For classification, two approaches have been followed: ID3 and Naive Bayes'.

ID3 which is a decision tree approach and Naive Bayes which is a Bayes' posterior probability based approach.

Testing was carried out by dividing dataset into 80/20 ratio for training and testing respectively.

As shown in the table below the accuracy of ID3 is more than that of Naive Bayes'.

Algorithms	Analysis		
	No of queries tested	*No of correct predictions*	*Accuracy*
ID3	180	148	82.22%
Naive Bayes	180	128	71.11%

Confusion matrix:

ID3:

yes	no	←classified as
84	17	yes
15	64	no

Naive Bayes:

yes	no	←classified as
66	35	yes
17	62	no

In Naive Bayes, class independency is required but in medical systems the classes are interrelated and hence the accuracy is low. Hence, for medical systems ID3 is the preferred approach.

VII. Conclusion

Since the available information is not clean and has a lot of missing values and outliers,preprocessing of data is performed to enhance quality of data. Removal of missing values is done using K-NN algorithm and binning removes outliers, thus producing a data set that is synonymous to the actual data with a good accuracy factor. Predicting the presence or absence of heart ailments is performed using classification techniques of ID3 which is a decision tree model and Naïve Bayes which is a conditional probability model. The combination of K-NN, Binning and ID3 algorithms produce a better accuracy rate as compared to the Naïve Bayes algorithm approach. The degree of interrelation between classes of the system provides an insight into the classification mechanism to be used for that respective system.

VIII. References

[1] NirmalaDevi, M.; Appavu, S.; Swathi, U.V., “An amalgam KNN to predict diabetes mellitus”, Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), 2013 International Conference on, pages 691 – 695, 25-26 March 2013.

[2] S.X. Wu, S.F. Liu, M.Q. Li, “The Method of Data Pre-processing in Grey Information Systems”, ControlAutomation, Robotics and Vision, 2006. ICARCV ‘06.9th International conference on, pages 1-4,5-8 Dec. 2006.

[3] Ranganatha S.; Pooja Raj H.R.; Anusha C.;Vinay S.K,, “Medical data mining and analysis for heart disease dataset using classification techniques”, Research & Technology in the Coming Decades (CRT 2013), National Conference on Challenges in, pages 1 – 5, 27-28 Sept. 2013.

[4] Sudhakar, K.; Manimekalai, Dr. M., "Study of Heart Disease Prediction using Data Mining," International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1,ISSN: 2277 128X, pages 1157-1160, January 2014.

[5] D.Lavanya; Dr.K.Usha Rani. "Performance Evaluation of Decision Tree Classifiers on Medical Datasets"International Journal of Computer Applications (0975 – 8887),Volume 26– No.4,pages 1-4, July 2011.

[6] Dr. K. Usha Rani, “Analysis Of Heart Diseases Dataset Using Neural Network Approach”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5, September 2011.

[7] Missing values in data mining, “Soft Computing and Intelligent Information Systems”, 25 June 2015. http://sci2s.ugr.es/MVDM#Imputation%20Methods

[8] Heart disease dataset, “{UCI} Machine Learning Repository”, 9 April 2015. https://archive.ics.uci.edu/ml/datasets/Heart+Disease