International Journal of Scientific Engineering and Science
Volume 1, Issue 10, pp. 15-17, 2017. ISSN (Online): 2456-7361

Disease Prediction System by Minimizing Number of Attributes

Meenakshi Sharma[1], Vijay Kumar Verma[2]
[1]M. Tech. (CSE) IV Semester, Lord Krishna College of Technology, Indore M.P. India
[2]Asst. Prof. CSE Dept., Lord Krishna College of Technology, Indore M.P. India

***Abstract*—** *In health industry, Data Mining provides numerous benefits such as detection of causes of diseases and identification of medical treatment methods. These help healthcare researchers for making efficient healthcare policies, drug recommendation systems, and developing health profiles of a person. The data generated by the health organizations is very huge and complex and also difficult to analyze. If this data is properly analyze important decision regarding patient health can be taken. This data contains details regarding hospitals, patients, medical claims, treatment cost etc. So, there is a need to develop powerful methods for analyzing and extracting important information from these complex data. In this paper proposed a new approach which predicts disease more accurately by using minimum number of responsible attribute. The proposed approach not only predicts the disease but classify into a particular class.*

***Keywords*—** *Disease, early prediction, diagnosis, symptoms, accuracy, class.*

I. INTRODUCTION

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data. The derived model may be represented in various forms, such as classification IF-THEN rules, decision trees, mathematical formulae, or neural networks. Classification methods can handle both numerical and categorical attributes. Constructing fast and accurate classifiers for large data sets is an important task in data mining and knowledge discovery. Classification predicts categorical class labels and classifies data based on the training set. Classification is two steps processes [1].

1. Model construction: describing a set of predetermined classes. Each tuple /sample is assumed to long to a predefined class, as determined by the class label attribute .The set of tuples used for model construction is training set .The model is represented as classification rules, decision trees, or mathematical formula.

2 Model usage: for classifying future or unknown objects .Estimate accuracy of the model .The known label of test sample is compared with the classified result from the model .Accuracy rate is the percentage of test set samples that are correctly classified by the model .Test set is independent of training set.

II. CLASSIFICATION TECHNIQUES

Three are several techniques are used for classification some of them are.

- Decision Tree,
- K-Nearest Neighbor,
- Support Vector Machines,
- Naive Bayesian Classifiers,
- Neural Networks.

A Decision Tree Classifier consists of a decision tree generated on the basis of instances. A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges.

K-Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. "Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between two points, $X=(x_1,x_2,\ldots\ldots,x_n)$ and $Y=(y_1,y_2,\ldots,y_n)$ is denoted by d(X, Y).

SVM is a very effective method for regression, classification and general pattern recognition. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high.

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. The Naive Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model identifies the characteristics of patients with heart disease.

III. LITERATURE REVIEW

In 2012 Qasem A. Radaideh et al. proposed "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance". They represent a study of data mining techniques and build a classification model to predict the performance of employees. They build CRISP-DM model. They used Decision tree to build the classification model. They perform several experiments using real data collected from several companies. The model is intended to be used for predicting new applicants [2].

In 2012 K. Rajesh et al proposed "Application of Data Mining Methods and Techniques for Diabetes Diagnosis". Their main aim mining the relationship in Diabetes data for efficient classification. They applied many classification algorithms on Diabetes dataset and the performance of those algorithms is analyzed. In future this works enhance of improvisation of the C4.5 algorithms to improve the classification rate to achieve greater accuracy in classification [3].

In 2013 M. Akhil Jabbar et al. proposed "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection". They introduced a classification approach based ANN and feature subset selection. They used PCA for preprocessing and to reduce no. Of attributes which indirectly reduces the no. of diagnosis tests which are needed to be taken by a patient. We applied our approach on Andhra Pradesh heart disease data base. Our experimental results show that accuracy improved over traditional classification techniques. This system is feasible and faster and more accurate for diagnosis of heart disease [4].

In 2013 Divya Tomar et al. proposed "A survey on Data Mining approaches for Healthcare". Survey explores the utility of various Data Mining techniques such as classification, clustering, association, regression in health domain. They represent a brief introduction of these techniques and their advantages and disadvantages. This survey also highlights applications, challenges and future issues of Data Mining in healthcare. Recommendation regarding the suitable choice of available Data Mining technique is also discussed [5].

In 2014 Dr. B Rosiline et al. proposed "Efficient Classification Method for Large Dataset by Assigning the Key Value in Clustering". They proposed classification method to discover data of big difference from the instances in training data, which may mean a new data type. The generalize Canberra distance for continuous numerical attributes data to mixed attributes data, and use clustering analysis technique to squash existing instances, improve the classical nearest neighbor classification method [6].

In 2015 S. Olalekan Akinola et al. proposed "Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study". They determine how data mining classification algorithm perform with increase in input data sizes. Three data mining classification algorithms Decision Tree, Multi-Layer Perceptron (MLP) Neural Network and Naïve Bayes were subjected to varying simulated data sizes. The time taken by the algorithms for trainings and accuracies of their classifications were analyzed for the different data sizes. By the result show that Naïve Bayes takes least time to train data but with least accuracy as compared to MLP and Decision Tree algorithms [7].

In 2016 Jaimini Majali et al. proposed "Data Mining Techniques for Diagnosis and Prognosis of Cancer". They used data mining techniques for diagnosis and prognosis of cancer. They proposed a system for diagnosis and prognosis of cancer using Classification and Association approach in Data Mining. They used FP algorithm in Association Rule Mining (ARM) to conclude the patterns frequently found in benign and malignant patients. They also used Decision Tree algorithm under classification to predict the possibility of cancer in context to age [8].

IV. Problem Statement

Classification techniques provide benefit to all the people such as doctor, healthcare insurers, patients and organizations who are engaged in healthcare industry. Decision tree, Bays Naive classification, Support Vector Machine, Rule based classification, Neural Network as a classifier etc. The main problem related to classification techniques are

- *Correctness:* This includes accuracy of the classifier in term of predicting the class label, guessing value of predicted attributes.
- *Speed:* This include the required time to construct the model (training time) and time to use the model (classification/prediction time)
- *Strength:* This is the ability of the classifier or predictor to make correct predictions given noisy data or data with missing values.
- *Scalability:* Efficiency in term of database size.

V. Proposed Method

First we assign most recommended value to every attribute as per suggested by the physician for heart attack condition cording to the given conditions. In seconds step we calculate total value for each tuple. Now we take an unknown tuple and apply the proposed method. The working process of proposed model is shown the figure 1.

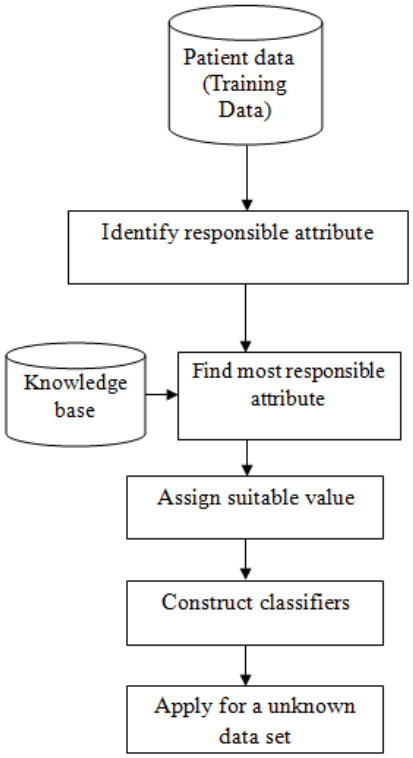

Fig. 1. Working of proposed approach.

International Journal of Scientific Engineering and Science
Volume 1, Issue 10, pp. 15-17, 2017. ISSN (Online): 2456-7361

Let D patient database. The proposed method used following step to classify the given unknown tuple.
(1) First find out the responsible attributes for the disease.
(2) Find out most responsible attribute from the list of attributes, using knowledge base.
(3) Assign value to the attribute suggested by physician.
(4) Calculate the total of the most responsible attributes value.
(5)Divide total value with the responsible attribute values.

VI. Experimental Analysis

We used VB dot net 2013 and SQL server 2010 R2 for experimental analysis. We have taken 5 attribute and 100 records of different patient with corresponding attribute and tested the proposed method. We are different parameter for our Experimental analysis one of them is number of records are correctly classified. We compare the proposed method with Bayesian Classification.

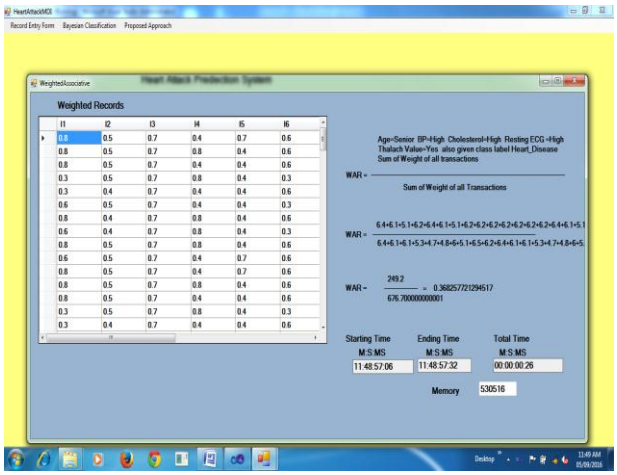

Fig. 2. Proposed approach with 100 records.

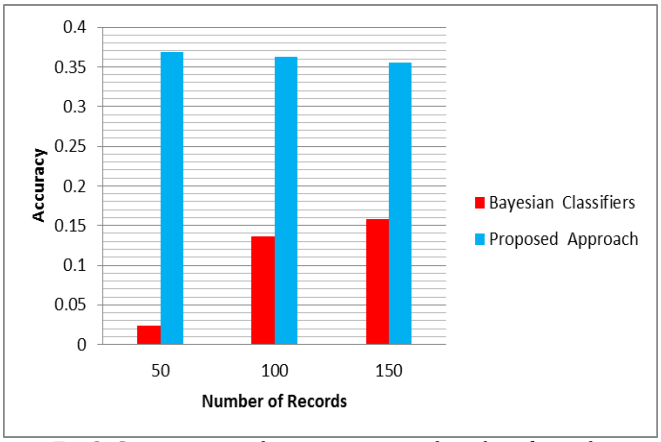

Fig. 3. Comparison graph using accuracy and number of records.

VII. Conclusion and Future Works

There are several techniques are available to predict heart disease problem like Decision trees, Bayesian classifiers, classification by back propagation, support vector machines, nearest-neighbor classifiers and case-based reasoning classifiers These techniques are compared on basis of Sensitivity, Specificity, Accuracy, Error Rate, True Positive Rate and False Positive Rate. The proposed method reduces number of attribute and reduces complex calculation. In future we also used fuzzy data set to include more desecrate value for the attribute.

Reference

[1] J. Han and M. Kamber, *Data mining, Concepts and Techniques*, Academic Press, 2003.
[2] Q. A. Al-Radaideh and E. Al Nagi, "Using data mining techniques to build a classification model for predicting employees performance," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 3, no. 2, pp. 144-151, 2012.
[3] K. Rajesh and V. Sangeetha, "Application of data mining methods and techniques for diabetes diagnosis," *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 2, issue 3, pp. 224-229, 2012.
[4] M. Akhil Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of heart disease using artificial neural networkand feature subset selection," *Global Journal of Computer Science and Technology Neural & Artificial Intelligence*, vol. 13, issue 3 version 1.0, 2013.
[5] D. Tomar and S. Agarwal, "Survey on data mining approaches for healthcare," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241-266, 2013.
[6] Dr. B. Rosiline Jeetha, "Efficient classification method for large dataset by assigning the key value in clustering," *International Journal of Computer Science and Mobile Computing*, vol. 3, issue. 1, pp. 319–324, 2014.
[7] S. Olalekan Akinola and O. Jephthar Oyabugbe, "Accuracies and training times of data mining classification algorithms: An empirical comparative study," *Journal of Software Engineering and Applications*, vol. 8, issue 9, pp. 470-477, 2015.
[8] J. Majali, R. Niranjan, and V. Phatak, "Data mining techniques for diagnosis and prognosis of cancer," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, issue 3, pp. 613-616, 2015.
[9] T. Sharma and A. Sharma, "Performance analysis of data mining classification techniques on public health care data," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, issue 6, pp. 11381- 11386, 2016.
[10] N. N. Salvithal and R. B. Kulkarni, "Appraisal management system using data mining classification technique," *International Journal of Computer Applications*, vol. 135, no. 12, pp. 45-50, 2016.