International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 07 | July -2017 www.irjet.net p-ISSN: 2395-0072

Early Identification of Diseases Based on Responsible Attribute Using Data Mining

Mr. Sudhir M. Gorade[1], Prof. Ankit Deo[2], Prof. Preetesh Purohit[3]

[1] *Dept. Of Computer Science and Engineering, SVCE Indore, M.P, India*
[2] *Professor, Dept. Of Computer Science and Engineering, SVCE Indore, M.P., India*
[3] *H.O.D., Dept. Of Computer Science and Engineering, SVCE Indore, M.P., India*

Abstract - *Now a day's Data Mining is becoming a common tool in healthcare field. Data mining tools help in analytical methodology for detecting valuable information. Data Mining provides several benefits in health industry. Detection of the fraud in health insurance, availability of medical solution to the patients at lower cost. Recognition of causes of diseases and identification of medical treatment methods. It also helps the healthcare researchers for making efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals etc. The data generated by the health organizations is very vast and complex and it is difficult to analyze the data in order to make important decision regarding patient health. This data contains details regarding hospitals, patients, medical claims, treatment cost etc. So, there is a need to generate a powerful tool for analyzing and extracting important information from this complex data. In this paper proposed a classification based algorithm which reduce number of attribute and classify a known record to a correct class.*

Keyword: - Prediction, Classification, Diagnosis, Symptoms, accuracy.

Classification used two steps in the first step a model is

1. INTRODUCTION

Classification is a data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. Classification divides data samples into target classes. The classification technique predicts the target class for each data points. For example, patient can be classified as "high risk" or "low risk" patient on the basis of their disease pattern using data classification approach. It is a supervised learning approach having known class categories. In binary classification, only two possible classes such as, "high" or "low" risk patient may be considered. Multiclass approach has more than two classes for example, "high", "medium" and "low" risk patient. Data set is partitioned as training and testing dataset. Using training dataset, we trained the classifier. Correctness of the classifier could be tested using test dataset. Classification is one of the most widely used methods of Data Mining in Healthcare organization. Different classification method such as decision tree, SVM and ensemble approach is used for analyzing data. Classification techniques are also used for predicting the treatment cost of healthcare services which is increases with rapid growth every year and is becoming a main concern for everyone [1,12].

Classification is the task of generalizing known structure to apply to new data. The classification task can be seen as a supervised technique where each instance belongs to a class, which is indicated by the value of a special goal attribute or simply the class attribute. The goal attribute can take on categorical values, each of them corresponding to a class. One of the major goals of a Classification algorithm is to maximize the predictive accuracy obtained by the classification model when classifying examples in the test set unseen during training 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.

2. 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 M. Akhil jabbar et. al proposed "Heart Disease Prediction System using Associative Classification and Genetic Algorithm". They proposed efficient associative classification algorithm using genetic approach for heart disease prediction. Their main motivation for using genetic algorithm in the discovery of high level prediction rules is that the discovered rules are highly comprehensible, having

high predictive accuracy and of high interestingness values [3].

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 [4].

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 [5].

In 2013 V. Krishnaiah et al proposed "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" They briefly examine the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Byes and Artificial Neural Network to massive volume of healthcare data. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information. For data preprocessing and effective decision making One Dependency Augmented Naïve Byes classifier (ODANB) and naive creedal classifier 2 (NCC2) are used. This is an extension of naïve Byes to imprecise probabilities that aims at delivering robust
[6].

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 [7].

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 [8].

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 Perception (MLP) Neural Network and Naïve Byes 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 [9].

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 [10].

In 2016 Tanvi Sharma et al proposed "Performance Analysis of Data Mining Classification Techniques on Public Health Care Data". They focused on the application of various data mining classification techniques used in different machine learning tools such as WEKA and Rapid miner over the public healthcare dataset for analyzing the health care system. The percentage of accuracy of every applied data mining classification technique is used as a standard for performance measure. The best technique for particular data set is chosen based on highest accuracy [11].

3. PROBLEM STATEMENT

There are various classification techniques that can be used for the identification and prevention of heart disease. The performance of classification techniques depends on the type of dataset that we have taken for doing experiment. 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

- **Accuracy:** - 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)
- **Robustness:** -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.

4. ARCHITECTURE FOR DISEASE DIAGNOSIS

Predication should be done to reduce risk of disease. Diagnosis is usually based on signs, symptoms and physical examination of a patient. Almost all the doctors are predicting heart disease by learning and experience. The diagnosis of disease is a difficult and tedious task in medical field. Predicting disease from various factors or symptoms is a multi-layered issue which may lead to false presumptions and unpredictable effects

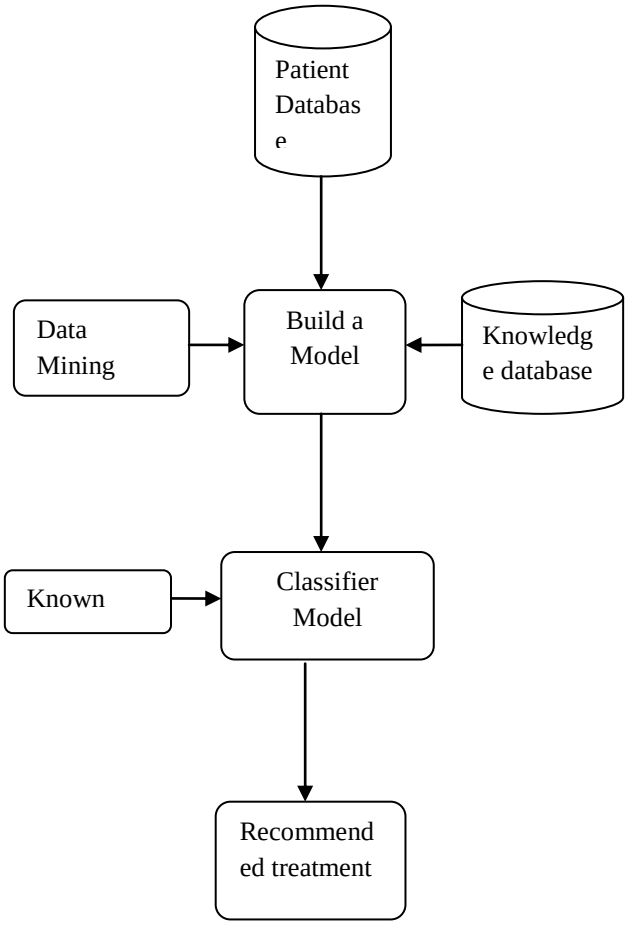

Figure 1 Decease prediction system

5. 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 2

Let D heart patient database and most recommended Value. The proposed method used following step to classify the given unknown tuple.

(1) First we assign the suggested most recommended value to each attribute suggested by the physician for heart attack.
(2) Find total of most recommended of each tuple
(3) Take an unknown tuple which has to be classified.
(4) Calculate the sum of the most recommended value of those tuple which satisfy the given conditions.
(5) Divide this value with the most recommended value of all tuple in the database.

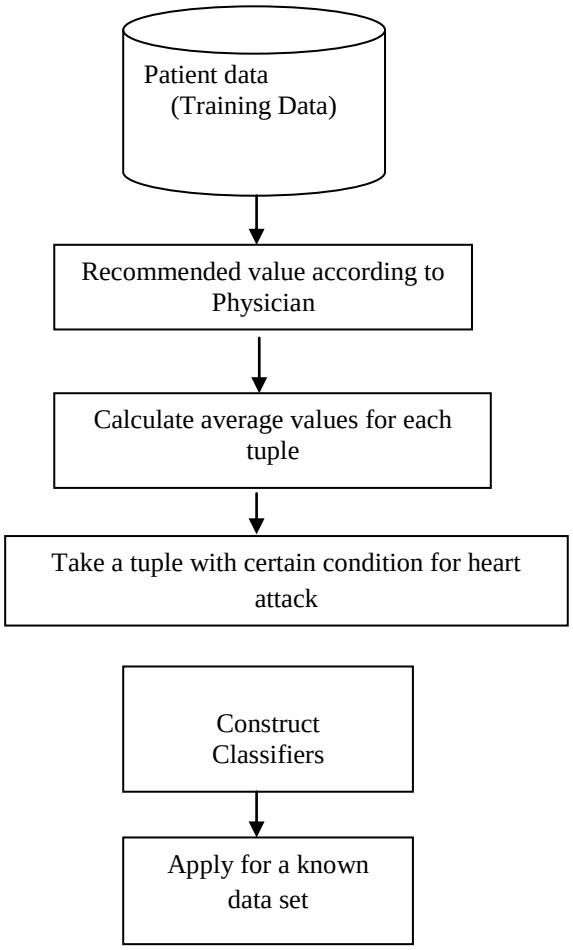

Figure 2 Architecture of proposed system

6. 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.

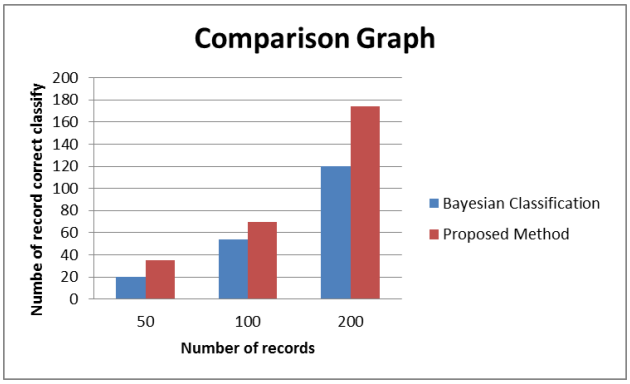

Figure 3. Comparison Graph

7. 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

8. References

[1] J. Han, M. Kamber, Data mining, Concepts and techniques, Academic Press, 2003.

[2] Qasem A. Radaideh et al proposed 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, 2012.

[3] In 2012 M. Akhil jabbar et. al proposed "Heart Disease Prediction System using Associative Classification and Genetic Algorithm" International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012.

[4] In 2012 K. Rajesh et al proposed "Application of Data Mining Methods and Techniques for Diabetes Diagnosis". ISSN: 2277-3754 ISO 9001:2008 Certified International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

[5] In 2013 M. Akhil Jabbar et al proposed "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection". Global Journal of Computer Science and Technology Neural & Artificial Intelligence Volume 13 Issue 3 Versions 1.0 Year 2013.

[6] In 2013 V. Krishnaiah et al proposed "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques". (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1), 2013, 39 45.

[7] In 2013 Divya Tomar et al proposed "Survey on Data Mining approaches for Healthcare" International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266 http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25

[8] In 2014 Dr. B Rosiline et al proposed Efficient Classification Method for Large Dataset by Assigning the Key Value in Clustering. IJCSMC, Vol. 3, Issue. 1, January 2014, pg.319 324 International Journal of Computer Science and Mobile Computing a Monthly Journal of Computer Science and Information Technology.

[9] In 2015 S. Olalekan Akinola et al proposed Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study. Journal of Software Engineering and Applications, 2015, 8, 470-477 Published Online September.

[10] In 2016 Jaimini Majali et al proposed "Data Mining Techniques for Diagnosis and Prognosis of Cancer". International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 3, March 2015.

[11] In 2016 Tanvi Sharma et al proposed "Performance Analysis of Data Mining Classification Techniques on Public Health Care" Data International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 6, June 2016.

[12] Appraisal Management System using Data mining. Classification Technique International Journal of Computer Applications (0975 8887) Volume 135 – No.12, February 2016