ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Performance Analysis of Data Mining Classification Techniques on Public Health Care Data

Tanvi Sharma[1], Anand Sharma[2], Prof. Vibhakar Mansotra[3]

[1] M. Tech Research Student, Dept. of Computer Science & I.T., University of Jammu, Jammu, J&K, India

[2]Research Scholar, Dept. of Computer Science & I.T., University of Jammu, Jammu, J&K, India

[3]Professor, Dept. of Computer Science & I.T., University of Jammu, Jammu, J&K, India

ABSTRACT: Public health care includes preventing disease, increasing life span and upholding the health through organized efforts. A large amount of data is available related to healthcare. Different relevant hidden information can be extracted from the public healthcare data using various data mining techniques. Data mining techniques proved to be very effective in extracting information from the public healthcare data. The present study focused on the application of various data mining classification techniques using different machine learning tools such as WEKA and Rapidminer over the public healthcare dataset for analysing 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.

KEYWORDS: Knowledge discovery in databases; Data mining; WEKA; Rapidminer

I. INTRODUCTION

In today's world, a vast amount of data is collected and stored daily. There is an important need to analyze this data but without any analytical tool, this seems impossible. This has led to the development of Knowledge Discovery in Databases (KDD) which transforms the low level data to high level knowledge. KDD consists of various processes at different steps and Data mining is one of those processes. Data mining is the process of discovering interesting knowledge from large amount of data stored in databases, data warehouses or other information repositories [1]. The main aim of data mining process is to extract information from a dataset and transform it into an understandable form in order to aid decision making [14]. A huge amount of data is available in healthcare sector but the knowledge extraction is poor. Thus, the analysis of the healthcare data is must. Knowledge Discovery in databases is becoming popular research tool for public healthcare data.

In this paper, we have done the performance analysis of different data mining classification techniques on healthcare data. This work helped in finding out the best data mining classification technique in terms of accuracy on the particular dataset. For thus, we have trained the public healthcare dataset taken from HMIS portal of MoHFW. The scrutinized classification techniques are K-nearest neighbour (KNN), Naive Bayes, Decision tree. The performance of these techniques is measured based on their accuracy. This study will help the future researchers to get efficient results after knowing best data mining classification technique for specific dataset.

II. LITERATURE REVIEW

Data mining has been widely applied in the medical field as this provide huge amount of data. Various researchers had applied the different data mining techniques on healthcare data. Choi *et al.*[9], applied 5 classification algorithms i.e. decision tree, artificial neural network, logistic regression, Bayesian networks and naive Bayes and stacking-bagging method for building classification models and compared the accuracy of the plain and ensemble model to predict whether a patient will revisit a healthcare centre or not. From results, the best classification model depends on data set

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

i.e. ANN in 3M data set, decision tree in 6M and logistic regression in 12M data set. Soni *et al.* [11] compared the data mining with traditional statistics and states some advantages of automated data system. This paper gives an overview of how data mining is used in health care and medicine. Patil *et al.* [3], determines whether a person is fit or unfit based on historical and real time data using clustering algorithms viz. K-means and D-stream are applied. The performance and accuracy of D-stream algorithm is more than K-means Al-Radaideh *et al.* [2], used decision tree to build a classification model for predicting employee's performance. To build a classification model CRISP-DM was adopted. Based on performance, job title is strongest attribute then university followed by other attributes. Jabbar *et al*. [8], proposed a decision support system to identify a risk score for predicting the heart disease. An associative classification algorithm using genetic approach is proposed for prediction. Experimental results show that the most of the classifier rules help in best prediction of heart disease. Garchchopogh *et al.* [6] explained the utilization of medical data mining in determining when we should perform surgery. The decision tree algorithm designed for this study generates correct prediction for more than 86.25% tests cases.D.K *et al.* [4], applied decision tree J48 to find the hidden patterns for Classification of women health disease (Fibroid). Decision tree J48 algorithm is implemented using WEKA 3.75 data miner. It classified the data into correctly and incorrectly instance. Hearty *et al.* [7] evaluated the usability of supervised data mining to predict dietary quality. Artificial Neural Networks and Decision trees were used. The ANN had a slightly higher accuracy than the decision tree.Sundar *et al.* [17] analyzed the performance of the Naive Bayes and WAC (weighted associative classifier) to predict the likelihood of patients getting a heart disease. This system uses CRISP-DM methodology to build the mining models. These methods depict that the WAC gives highest percentage of correct predictions for diagnosing patients with a heart disease.Zurada *et al.* [19], examined and compared the effectiveness of neural networks, decision tree, logistic regression, memory based reasoning and the ensemble model in evaluating whether the bad debt is likely to be repaid. They employed SAS Enterprise Miner to build initial and final model. Computer simulation shows that the logistic regression, neural network model and ensemble model produced best overall classification accuracy. Koç *et al*. [10], applied ANN and logistic regression to predict if the client will subscribe a term deposit or not after marketing campaign. ANN classifies 84.4% data correctly while logistic regression classifies 83.63% data correctly but LR takes 54 seconds and ANN takes 11 seconds to run. Thus, with more data and higher dimensional feature space, using ANN will be more efficient. Haghanikhameneh *et al*. [5] compared the various classification algorithms to predict the bandwidth usage pattern in different time intervals among different groups of users in the network comparison of different classification algorithms including. Decision Tree and Naïve Bayesian using Orange is done. The Decision Tree algorithm achieved 97% accuracy and efficiency in predicting the required bandwidth inside the network. Sakshi *et al.* [15] provided a complete analysis of different data mining classification techniques that includes decision tree, Bayesian networks, k-nearest neighbor classifier & artificial neural network. Performance of these algorithms is analyzed based on accuracy, ability to handle corrupted data and speed... Elsid *et al*.[18], in this paper, the knowledge is retrieved from a huge amount of data about students using an efficient technique of data mining to help the administration to make a quick decision. Rani *et al.* [13] analyzed the efficiency of different classification algorithm in data mining using blood transfusion dataset. The comparison of various algorithms in classification is done .The algorithm Random tree has shows 93.18% accuracy within short duration when compared with other algorithms in classification. Pujari *et al*. [12] described the performance analysis of different data mining classifiers such as classifiers Logistic Regression, SVM and Neural Network before and after feature selection on binomial data set. The classification performance of all classifiers is based on various statistical performance measures like accuracy, specificity and sensitivity. Gain chart and R.O.C chart are also used to measure the performances of the classifiers.

III. RESEARCH METHODOLOGY

In this study, three data mining techniques for predictive data mining task that includes Decision tree, K-NN, Naïve Bayes. These methods are used for generating knowledge to make it useful for decision making. Each method will produce different results to classify the districts into focused or non-focused states comprising the available variables in dataset .The experimentation is performed using WEKA and Rapidminer..

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

A. J48 Decision Tree

J48 decision tree is an open source java implementation of commonly known C4.5 supervised classification algorithm in WEKA. It is an evolution and extension of ID3 algorithm developed by Quinlan. It is a fraction between information gain and its splitting information.

B. Naïve Bayes

This method is based on probabilistic knowledge. The naïve Bayes rule outputs probabilities for the predicted class of each member of the set of test instance. Naïve Bayes is based on supervised learning. The goal is to predict the class of the test cases with class information that is provided in the training data.

C. K-NN

K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. K-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

IV. PUBLIC HEALTHCARE DATASET

The dataset used for this experiment was taken from HMIS portal of MoHFW. In this data file, each record represents the complete details of facility services available and used in each of the district. Table 1 shows the description of dataset selected for this work.

Table 1. Dataset Description

DATASET	NO. OF ATTRIBUTES	INSTANCES	CLASSES
Health Facility Services Data	16	739	2

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, the experimental results and analysis done for this study are shown. The research methodology has been explained in the previous section. For the experiments, various data mining classification techniques have been applied on the public healthcare dataset. In this study, WEKA and Rapidminer machine learning tools for data mining are used to accomplish the objectives. The percentage of accuracy rate and error rate of data mining Classification techniques are used as the measurement parameters for analysis. These parameters suggest that the classifier having a higher accuracy rate and lower value of error rate classify the dataset in highly corrected manner and vice-versa. In this research, the data is firstly divided into training data and testing data. The training set is used to construct the classifier and test set used for validation. In this research, the percentage of dataset used for training and testing data are 40% and 60% respectively. Then, the 10 fold cross validation method is applied to generate the classifiers using previously mentioned machine learning tools. Finally, the results are documented in terms of accuracy rate and error rates. The results are shown below:

Table 2 and 3 displays the results for classification techniques applied on health facility services data in WEKA and Rapidminer respectively. Considering accuracy and error rates as performance measure the classification techniques with highest accuracy are obtained for Health Facility Services Data in given different machine learning tools

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

TABLE 2. RESULTS USING WEKA

TECHNIQUE USED	ACCURACY RATE	ERROR RATE
NAIVE BAYES	77.94	22.05
Kstar(K-NN)	69.42	30.58
DECISION TREE (J48)	77.26	22.73

TABLE 3. RESULTS USING RAPIDMINER

TECHNIQUE USED	ACCURACY RATE	ERROR RATE
k-NN	64.43	35.57
DECISION TREE	96.67	3.33
NAIVE BAYES	75.83	24.17

. Figure 1 and 2 displays the performance analysis of classification techniques using WEKA and Rapidminer respectively. From figure 3, it is clearly evident that Decision tree using Rapid miner is best classifier for this particular data set.

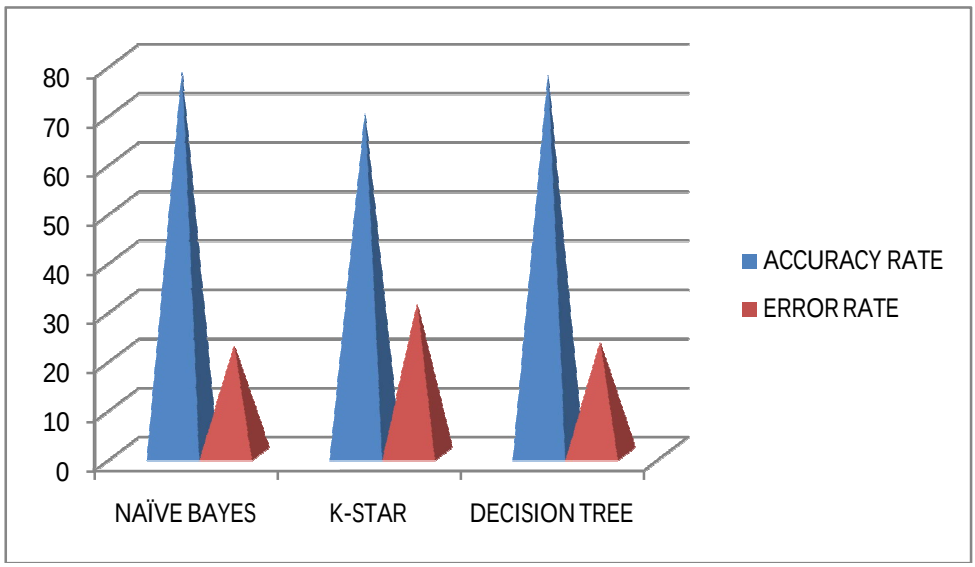

Fig 1. Performance analysis of classification techniques using WEKA

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

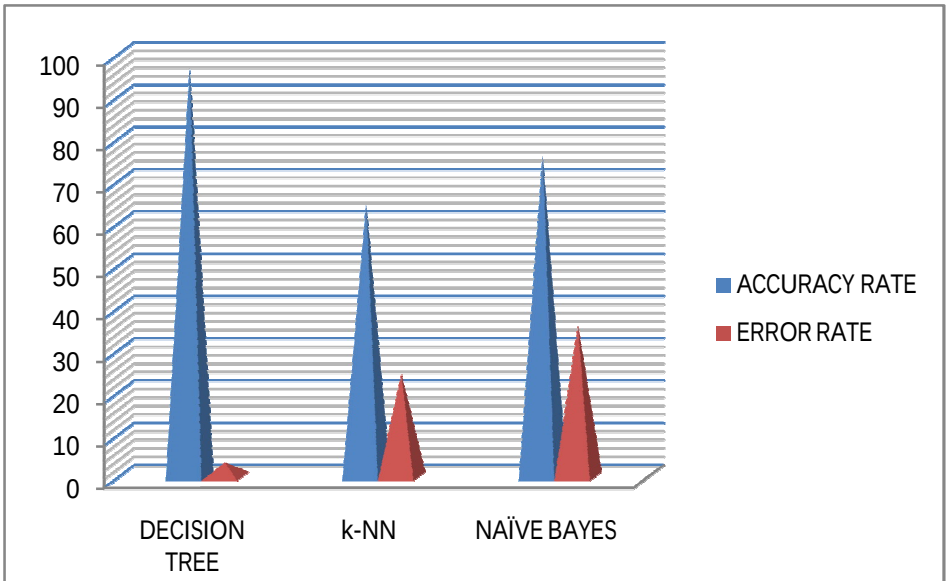

Fig 2. Performance analysis of classification techniques using Rapidminer

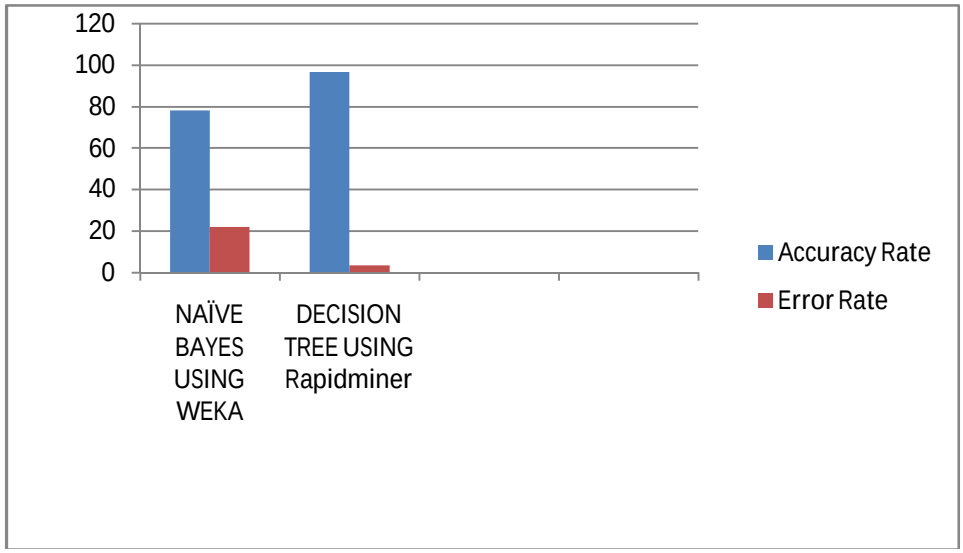

Fig 3: Comparison between the best classification techniques applied on Public healthcare dataset

VI. CONCLUSION AND FUTURE WORK

In this study, different data mining classification techniques are applied on the specific dataset using different data mining toolkits such as WEKA and Rapidminer. This study concluded that the Decision Tree algorithm using Rapidminer data mining tool is the best classification method for this particular dataset. This study will help researchers to get efficient results after knowing the best classification method for this particular dataset.

.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

REFERENCES

1. Abdullah H. Wahbeh, et al "A Comparison Study between Data Mining Tools over some Classification Methods" (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence
2. Al-Radaideh et al "Using data mining techniques to build a classification model for predicting employee's performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012
3. Dipti Patil et al, "An adaptive parameter for data mining approach for healthcare applications" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 1, 2012
4. D.K, "Classification of women health disease (Fibroid) using decision tree algorithm", International Journal of Computer Applications in Engineering Science Vol.2, Issue 3, September2012,
5. Fartash. Haghanikhameneh "A Comparison Study between Data Mining Algorithms over Classification Techniques in Squid Dataset" International Journal of Artificial Intelligence, Autumn (October) 2012, Vol. 9
6. Garchchopogh et al, "Application of decision tree algorithm for data mining in healthcare operations: A case study", International Journal of Computer Applications Vol 52 – No. 6, August 2012
7. Hearty et al, "Analysis of meal patterns with the use of supervised data mining techniques-Artificial Neural Network and Decision Tree", The American Journal of Clinical Nutrition
8. Jabbar et al "Heart disease prediction system using associative classification and Genetic Algorithm", International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012
9. Keunho Choi et al. "Classification and Sequential Pattern Analysis for Improving Managerial Efficiency and Providing Better Medical Service in Public Healthcare Centers" health inform res, pp.67-76, June 2010
10. Koç et al, "A comparative study of artificial neural network and logistic regression for classification of marketing campaign results", Mathematical and Computational Applications, Vol. 18, No. 3, 2013, pp. 392-398
11. Nakul Soni, Chirag Gandhi, "Application of data mining to health care", International Journal of Computer Science and its Applications,
12. Pushpalata Pujari " Classification and comparative study of data mining classifiers with feature selection on binomial data set" Journal of Global Research in Computer Science, Vol. 3, No. 5, May 2012
13. S.Asha Rani and Dr.S.Hari Ganesh, " A comparative study of classification algorithm on blood transfusion" International Journal of Advancements in Research & Technology, Volume 3, Issue 6, June-2014
14. Saichanma et al. "The Observation Report of Red Blood Cell Morphology in Thailand Teenager by Using Data Mining Technique." Advances in hematology, 2014
15. Sakshi and Prof.Sunil Khare "A Comparative Analysis of Classification Techniques on Categorical Data in Data Mining" International Journal on Recent and Innovation Trends in Computing and Communication Vol. 3 Issue: 8,pp.5142 – 5147
16. Shelly Gupta et al. "Performance Analysis of Various Data Mining Classification Techniques on Healthcare Data" International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 4, August 2011
17. Sundar et al. "Performance analysis of classification data mining techniques over heart disease database", [IJESAT] International Journal of Engineering Science and Advanced Technology, Volume-2, Issue-3,pp. 470 – 478
18. Tariq O. Fadl Elsid and Mergani. A. Eltahir "An Empirical Study of the Applications of Classification Techniques in Students Database" Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 4, Issue 10(Part - 6), pp.01-10, October 2014
19. Zurada et al, "Comparison of the performance of several data mining methods for bad debt recovery in the healthcare industry", The Journal of Applied Business Research – Springer Vol.21, Number 2, 2005