International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 8 Issue 08 Series. I || Aug 2019 || PP 45-48

Performance Comparison of K-Nearest Neighbors and Gaussian Naïve Bayes algorithms for Heart Disease Prediction

TemesgenAbera Asfaw[1]
Department of Computer Science & Engineering,College of Engineering, Jawaharlal Technological University, Telangana–500085, India

Abstract: *Most of people are dying due to Heart disease around the biosphere nowadays. Often, the large amount of information is gathered to detect diseases in medical science. All of the information is not useful but vital in taking the correct decision. Thus, it is not always easy to prevent the heart disease because it requires experiences about heart failure symptoms for an early prediction. Most of the therapeutic dataset are discrete, general and miscellaneous. However, data mining is an enthusiastic technique for mining unseen, prophetic and actionable information from the widespread catalogs. In this paper, two classification techniques of Data mining are used to evaluate the performance of algorithms such that KNN, Gaussian Naïve Bayes are used on heart disease dataset for better prediction. Several performance measurement features such as accuracy, ROC curve, precision, recall, and F1-score are dignified to control the routine of the classification. Among both Naïve Bayes and KNN, Naïve Bayes Classification performs better accuracy and it's 77.05%.*
Keywords: *Data mining; KNN; Gaussian Naïve Bayes, Heart Disease*

Date of Submission: 03-08-2019 Date of acceptance: 14-08-2019

I. Introduction

Data mining is speedily expanding in extensive range of applications. One of the important data mining fields is therapeutic data mining. There is a wealth of data available in healthcare but there is no real analysis tool to discover hidden relations in data. While millions of people die of heart disease yearly, application of data mining methods in heart disease diagnosis appears to be vital. Exposed Knowledge can help physicians in diagnosis of heart disease. Data mining is the fundamental which results in the discovery of roundabout but potentially valuable knowledge from vast amount of data. Data mining technology provides the user with the approaches to find novel and understood designs from enormous data. In the healthcare area, exposed knowledge can be castoff by the healthcare administrators and medical physicians to advance the accuracy of diagnosis, to improve the golly of surgical operations and to diminish the harmful effects of drug It aims also to suggest less expensive therapeutic. Heart disease is the foremost cause of death in the ecosphere over the past 10 years. Investigators have been using some data mining techniques to help health care specialists in the diagnosis of heart disease. K-Nearest-Neighbors (KNN) is one of the fruitful data mining techniques used in classification problems. However, it is fewer used in the diagnosis of heart sickness patients. Lately, researchers are screening that merging diverse classifiers through voting is outperforming other single classifiers.

The knowledge data is classified by using different classification algorithms such as Naive Bayes, K-Nearest Neighbor the accuracy of both classification algorithms is well-known. From these Naive Bayes algorithm does better than K-Nearest Neighbor for heart disease classification. Medical decision support methods are designed to provision clinicians in their diagnosis for heart disease.

II. Literature Review

HardikManiya et al made an effort to predict the most widely spread disease in India named tuberculosis. Using data collected from various TB centers, we made an effort to fetch out hidden patterns and by learning this pattern through the collected data for tuberculosis we can diagnose and predict the disease. In the research work we are comparing naïve Bayes classifier and KNN, two the most effective techniques for data classification (especially for medical diagnoses), implemented using C language and using Weka tool respectively and classify the patient affected by tuberculosis into two categories (least probable and most probable). We have used 19 symptoms of tuberculosis and collect 154 cases. We have achieved nearly 78% accuracy with low false negative. This algorithm extracts hidden patterns from available TB database. Naïve Bayes could identify all the significant medical predictors. The prototype can further be improved by incorporating various other attributes and increasing the number of cases for training and testing. The efficiency of results using KNN can be further improved by increasing the number of data sets and for Naïve Bayesian classifier by increasing attributes or by selecting weighted features. **M.Akhil jabber et al** Nearest neighbor

(KNN) is very simple, most popular, highly efficient and effective algorithm for pattern recognition. KNN is a straight forward classifier, where samples are classified based on the class of their nearest neighbor. Medical data bases are high volume in nature. If the data set contains redundant and irrelevant attributes, classification may produce less accurate result. Heart disease is the leading cause of death in INDIA. In Andhra Pradesh heart disease was the leading cause of mortality accounting for 32%of all deaths, a rate as high as Canada (35%) and USA. Hence there is a need to define a decision support system that helps clinicians decide to take precautionary steps. In this paper we propose a new algorithm which combines KNN with genetic algorithm for effective classification. Genetic algorithms perform global search in complex large and multimodal landscapes and provide optimal solution. Experimental results shows that our algorithm enhance the accuracy in diagnosis of heart disease. In this paper we have presented a novel approach for classifying heart disease. As a way to validate the proposed method, we have tested with emphasis on heart disease on A.P besides other machine learning data sets taken from UCI repository.

III. Methods for Prediction and Classification

K-Nearest Neighbors (KNN):classifies the test data using the training set directly. To classify any test data; it first calculates K value, whichdenotes the number of K-Nearest Neighbors. For all testdata, it calculates the distance between all the training data and then sorts the distance. Then by using majority voting, classlabel will be allotted to the test data.

Gaussian Naïve Bayes: When all the data values of any particular dataset arenumeric, then Gaussian Naïve Byes is used. It follows a normal distribution. Mean and standard deviation are used to definethe likelihood density function. It calculates the mean andstandard deviation for all attribute of the dataset. Aftercalculating this, when any test data pattern comes, then byusing the mean and standard deviation calculate theprobabilities for each test data. It assigns a class label to the test.Bayes' Theorem is stated as:

$$P(h|d) = (P(d|h) * P(h)) / P(d) \quad (1)$$

Where P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability.P(d|h) is the probability of data d given that the hypothesis h was true(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.P(d) is the probability of the data (regardless of the hypothesis).You can see that we are interested in calculating the posterior probability of P(h|d) from the prior probability p(h) with P(D) and P(d|h).

IV. Result

To conduct the experiment, two classification algorithms were used. Classification algorithms were implemented in pycharm 3.6.

K-Nearest Neighbors: In this method, K- Nearest Neighbors showed poor performance because KNN classifies test data directly from the dataset, no training was performed before testing. **Gaussian Naïve Bayes**: At the training stage, it calculated the mean and standard deviation of each attribute. This mean and standard deviation were used to calculate the probabilities for the test data. For this reason, some attributes values are too big or too small from the mean. When testing data pattern contains those attributes values, it affects the classifier performance and sometimes gives wrong output label.

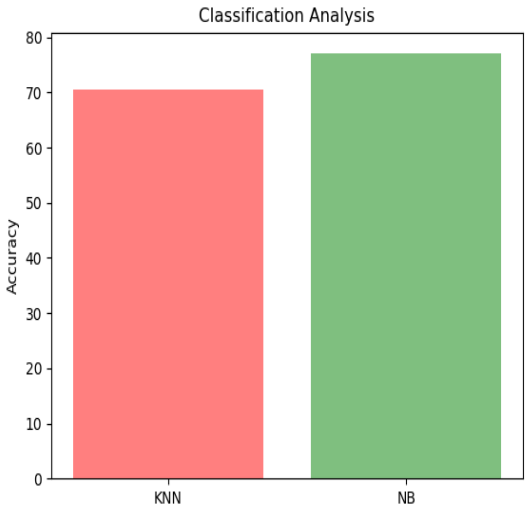

Figure 1.Bar chart based result of comparison

TABLE ICONFUSIONMATRICESOFCLASSIFIERSUSING14ATTRIBUTES

Classification	Confusion Matrix		Accuracy
KNN	TP=61	FN=21	66.73%
	FP=25	TN=45	
Gaussian Naïve Bayes	TP=75	FN=11	74.15%
	FP=54	FN=16	

TABLE II.CONFUSION MATRICES OF CLASSIFIERS USING 10 ATTRIBUTES

Classification	Confusion Matrix		Accuracy
KNN	TP=63	FN=23	71.05%
	FP=25	TN=51	
Gaussian Naïve Bayes	TP=70	FN=30	77.05%
	FP=13	TN=40	

TABLE III.CLASSIFICATION REPORT OF CLASSIFIERS

Classification		Precision	Recall	Specificity	F1-Score	Roc Area
KNN(using attributes)	14	0.66	0.68	0.64	0.67	0.65
KNN(using attributes)	10	0.70	0.71	0.69	0.71	0.71
Gaussian NB(using attributes)	14	0.73	0.74	0.73	0.72	0.70
Gaussian NB(using attributes)	10	0.77	0.76	0.76	0.77	0.77

V. Conclusion and Future work

1.1. Conclusion

This paper compares the performances of the classification algorithms in the prediction of heart disease. It tries to find out the best classifier for this task. In the experimental dataset, 10 attributes were used. Among the studied classifiers, Naïve Bayes classifier performs better than KNN classification algorithms. Based on experiment obtained in these studies, designing heart disease prediction system using Gaussian Naïve Bayes classifiers have high accuracy than KNN classifiers. Binary class problem is solved to identify whether the patient has heart disease or not. It is suggested to elucidate the multiclass problem for distinguishing heart disease by isolating heart disease patients into various classes.

1.2. Future Work

This system will be customized to predict not only the presence or absence of heart disease but also to predict the risk factor of heart failure to take extra care of those patients at an early stage and avoid heart failure. Real-time data from different hospitals may be collected for detecting heart disease patients and compute the effectiveness of classifiers for more consistent diagnosis of heart disease patients.

References

[1]. M. Kamber and P. J. Han, Data Mining Concepts, and Techniques, 3rd ed.,2012.

[2]. S. Palaniappan, R. Awang, " Intelligent Heart Disease Prediction System Using Data Mining Techniques," IJCSNS International Journal of Computer Science and Network Security, vol. 8, no. 8, August2008.

[3]. A. Khemphila, V. Boonjing "Comparing Performances of Logistic Regression, Decision trees, and Neural Networks for Classifying Heart Disease Patients," 2010 IEEE International Conference on Computer Information Systems and Industrial Management Systems(CISIM), pp. 193-199,2010.

[4]. M. Sultana, A. Haider and M. S. Uddin, "Analysis of Data Mining Techniques for Heart Disease Prediction," 3rdInternational Conference on Electrical Engineering and Information Communication Technology (ICEEICT),2016.

[5]. S. Xu, Z. Zhang, D. Wang, J. Hu, X. Duan and T. Zhu, "Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework," 2017 IEEE 2ndInternational Conference on Big Data Analysis,2017.

[6]. S. Pouriyeh, S. Vahid, G. Sannino, G. D. Pietro and H. Arabnia, J. Gutierrez, "A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease," 22nd IEEE Symposium on Computers and Communication (ISCC 2017):Workshops- ICTS4eHealth, 2017.

[7]. P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, D. S. Lee, "Using Methods from Data Mining and Machine Learning Literature for Disease Classification and Prediction: a Case Study Examining Classification of Heart Failure Subtypes," Journal of Clinical Epidemiology 66 (2013) pp. 398-407,2013.

[8]. H. M. Islam, Y. Elgendy, R. Segal, A. A. Bavry and J. Bian, "Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: A machine learning approach," Journal of Heart & Lung, pp. 1-7,2017.

[9]. M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Computational Intelligence Technique for Early Diagnosis of Heart Disease," 2015 IEEE International Conference on Engineering and Technology (ICETECH), 20thMarch2015.
[10]. UCI Machine Learning Repository. [Online]. Available: http://archive.ics.edu/ml/datasets/heart+disease.

Temesgen Abera Asfaw" Performance Comparison of K-Nearest Neighbors and Gaussian Naïve Bayes algorithms for Heart Disease Prediction" " International Journal of Engineering Science Invention (IJESI), Vol. 08, No. 08, 2019, PP 45-48