International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 03 Issue: 09 | Sept -2019 ISSN: 2590-1892

DETECTION OF CARDIAC ARRHYTHMIA DISEASE USING MORE EFFICIENT MACHINE LEARNING ALGORITHM

SIDDAMMA D. K[1] , DR. SAI MADHAVI .D[2]

PG, DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,RYMEC,BELLARY ,INDIA[1]
ASSOCIATE PROFESSOR ,COMPUTER SCIENCE & ENGINEERING,RYMEC,BELLARY ,INDIA[2]

Abstract-Cardiac Arrhythmia refers to a medical condition in which heart beats irregularly. This paper aims to detect and classify arrhythmia into various classes based on the Electrocardiogram(ECG) readings and also other attributes. A some popular techniques were implemented namely Naive Bayes, SVM, Random Forests and Neural Networks. In new approach I have implemented k-nearest algorithm for prediction of cardiac arrhythmia. The new implemented method achieves an overall accuracy when compared with various other existing approaches.

Keywords- Arrythemia, KNN, LR, Naïve bayes,SVM

1. INTRODUCTION

Millions of people are getting some sort of heart disease every year and heart disease is the biggest killer of both men and women around the world. Medical diagnosis plays vital role and yet complicated task that needs to be executed efficiently and accurately. The prediction of cardiac arrhythmia disease is one of the areas where machine learning can be implemented. Machine learning algorithms have the advantage of dealing with complex non-linear problems with a good flexibility and adaptability. The undertaking goes for utilizing distinctive machine learning calculations like Naive Bayes, SVM, Logical regression & KNN.

2. DATASET

The dataset for the venture is taken from the UCI AI Repository https://chronicle.ics.uci.edu/ml/datasets/Arrhythmia (1 csv record, 1 data document).There are (452) columns, each speaking to medical record of an alternate patient. There are 279 characteristics like age, weight and patient's ECG related information. The informational index is marked with 16 unique classes. Classes 2 to 15 compare to various kinds of arrhythmia. Class 1 relates having a place with class 1 and 185 in- positions being part among the 14 arrhythmia classes and the rest 22 are unclassified, 3 of the classes identified with the level of AV square don't show up in the informational collection. The marks for this informational index are acquired from cardiologists

Heart disease dataset contains data from four institutions.

1. Cleveland Clinic Foundation.

2. Hungarian Institute of Cardiology, Budapest.

3.V.A. Medical Centre, Long Beach, CA.

4. Hospital, Zurich, Switzerland.

For the purpose of this study, the data set provided by the Cleveland Clinic Foundation is used. This dataset was provided by Robert Detrano, M.D, Ph.D. Reason to choose this dataset is, it has less missing values and is also widely used by the research community.

3. DATA PROCESSING

The dataset has features with large numeric values which may directly affect the accuracy of prediction when compared to the features with small numeric values.

3.1 Classification Task

From the perspective of automatic learning, heart disease detection can be seen as a classification or clustering problem. On the other hand, we formed a model on the vast set of presence and absence file data; we can reduce this problem to classification. For known families, this problem can be reduced to one classification only - having a limited set of classes, including the heart disease sample, it is easier to identify the right class, and the result would be more accurate than with clustering algorithms. For the purpose of comparative analysis, four Machine Learning algorithms are used. The different Machine Learning (ML) algorithms are K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes and Logistic Regression. The reason to choose these algorithms is based on their popularity .

4. EXISTING SYSTEM

In existing system we use SVM algorithm, Naïve-Bayes algorithm and Logistic Regression algorithm.

4.1 NAÏVE-BAYES

Naive Bayes is a surprisingly powerful algorithm for predictive modeling. It is a statistical classifier which assumes no dependency between attributes attempting to maximize the posterior probability in determining the class. Theoretically, this classifier has the minimum error rate, but may not be the case always. Inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data. This model is associated with two types of probabilities which can be calculated from the training dataset directly:

a) The probability of every class.

b) The conditional probability of each class with each x value.

According to Bayesian theorem
$P(A|B)=P(A)*P(B/A)/P(B)$,

where

$P(B|A)=P(A\cap B)/P(A)$.

Bayesian classifier calculates conditional probability of an instance belonging to each class, based on the above formula, and based on such conditional probability

data, the instance is classified as the class with the highest conditional probability. If these probabilities are calculated, then the probabilistic model can be implemented to make predictions with new data using Naïve Bayes Theorem. When the data is real-valued it is likely to assume a Gaussian distribution (bell curve). Thus, these probabilities can easily be estimated. Naive Bayes is called naive because of assuming each input variable independent.

4.2. Support Vector Machine:

SVM is a technique for ramification of both linear and non-linear data. It applies a non-linear mapping method so that it can transform the training data into a higher dimension. A hyperplane is a kind of line which separates the input variable space in SVM. The hyperplane can separate the points in the input variable space containing their class that is either 0 or 1

4.3 Logistic Regression

A logistic regression is a classification algorithm. For binary classification problem, in order to predict the value of predictive variable y when y ∈ [0, 1], 0 is negative class and 1 is positive class. It also uses multiclassification to predict the value of y when y ∈ [0, 1, 2, 3].

5. PROPSED SYSTEM

5.1 KNN

KNN is a non-parametric machine learning algorithm. The KNN algorithm is a supervised learning method. This means that all the data is labeled and the algorithm learns to predict the output from the input data. It performs well even if the training data is large and contains noisy values. The data is divided into training and test sets. The train set is used for model building and training. A k-value is decided which is often the square root of the number of observations. Now the test data is predicted on the model built.

6. METHODOLOGY

In this paper, comparison of various machine learning methods is done for predicting the 10 years risk of coronary heart disease of the patients from their medical data. The following fig1 is the data flow diagram for proposed methodology.

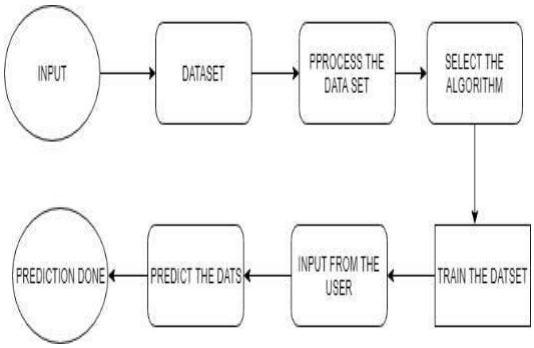

Fig 1: Data Flow diagram

Here we give dataset and process the given dataset, select the algorithm, train the dataset , take input parameters from users and predict the data by using dataset, and it will predict a patient has cardiac arrthymia disease or not.

6.1 PREDICTION ACCURACY METRICS

Prediction of cardiac arrhythmia disease can be determined using accuracy metrics as:

Accuracy = (TP+TN)/ (TP+TN+FP+FN)

Where,

TP – Number of true positives (number of instances which are unhealthy and predicted correctly)

TN – Number of true negatives (number of instances which are healthy and predicted correctly)

FP – Number of false positives (number of instances which are healthy but predicted wrongly)

FN – Number of false negatives (number of instances which are unhealthy but predicted correctly)

The obtained result is multiplied by 100 to get the value in percentage.

6.2 COMPREHENSIVE ANALYSIS OF PREDICTION ACCURACY

The different types of machine learning techniques used for comparison includes Naive Bayes (NB), K Nearest Neighbor (K-NN), Support Vector Machine (SVM),Logistic Regression(LR). In this experiment, 220 instances were selected as training data and 83 were selected as testing data. Since the total number of attributes was 13, the input layer was designed with 13 nodes. The experiment resulted with 87% accuracy in predicting the cardiac arrhythmia disease, which out performed KNN technique.

Table 1.Analysis of Prediction Accuracy Rates.

S No	Algorithm	Prediction accuracy	No of Attributes
1	SVM	83%	14
2	LR	79%	14
3	Naïve Bayes	84%	14
4	KNN	87%	13

Table 1 Analysis of Prediction Accuracy.

6.2 Performance

A learning curve is a plot of model learning performance over experience or time. Learning curves are a widely used diagnostic tool in machine learning for algorithms that learn from a training dataset incrementally. The model can be evaluated on the training dataset and on a hold out validation dataset after each update during training and plots of the measured performance can created to show learning curves.

The learning curves of different algorithms are shown below:

a.NaïveBayes

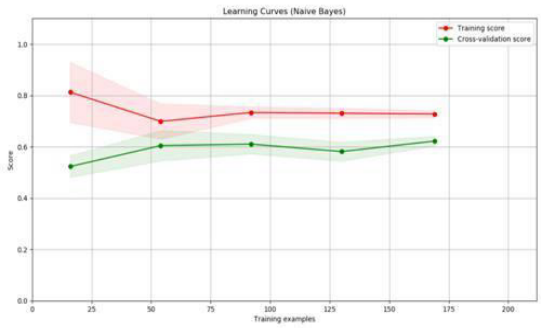

Fig 2: Learning score for the Naïve Bayes Algorithm.

b. SVM

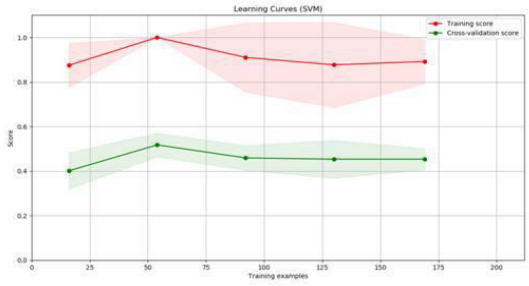

Fig3: Learning score for the SVM Algorithm.

c. Logistic Regression

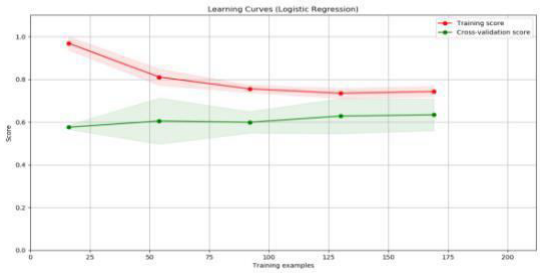

Fig 4: Learning score for the LR Algorithm

d. KNN

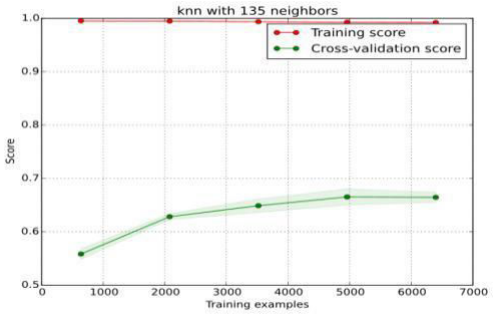

Fig 5: Learning score for the KNN Algorithm

7. CONCLUSION

The purpose of this paper was to compare algorithms with different performance measures using machine learning. All data were pre-processed and used for test prediction. Each algorithm worked better in some situations and worse in others. KNearest Neighbour K-NN, and SVM, are the models likely to work best in the data set used in this paper.

The paper involved analysis of the cardiac arrhythmia disease patient dataset with proper data processing. Then, 4 models were trained and tested with maximum scores as follows:

1. K Neighbors Classifier: 87%
2. Support Vector Classifier: 83%
3. Logistic regression: 79%
4. Naïve Bayes: 84%

K Neighbors Classifier scored the best score of 87% .

REFERENCES

[1] H. Altay Guvenir, Burak Acar, Gulsen Demiroz, Ayhan Cekin "A Supervised MachineLearningAlgorithmfor Arrhythmia Analysis." Proceedings of the Computers in Cardiology Conference, Lund, Sweden, 1997

[2] zift, Akin."Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis." Computers in Biology and Medicine 41.5 (2011): 265-271

[3] Hall, Mark A., and Lloyd A. Smith. "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper." FLAIRS conference. 1999.

[4] Uyar, Asl, and Fikret Gurgen. "Arrhythmia classification using serial fusion of support vector machines and logistic regression." Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2007. IDAACS 2007. 4th IEEE Workshop on. IEEE, 2007.

[5] Polat, Kemal, Seral ahan, and Salih Gne. "A new method to medical diagnosis: Artificial immune recognition system (AIRS) with fuzzy weighted pre-processing and application to ECG arrhythmia." Expert Systems with Applications 31.2 (2006): 264- 269. [6] Rudokait-Margeleviien, Dovil, Henrikas Praneviius, and Mindaugas Margeleviius. "Data classification using Dirichlet mixtures." Information Technology and Control 35