An Intangible System to Augment the Prediction of Heart Diseases Using Machine Learning Techniques

Shaik Akbar, P. Sri Silpa, Anand Thota, K. Nageswara Rao

Abstract--- In Present day medical scenario, it is very difficult to find the Heart-attack, Blood Pressure of the patient. So we are introducing an idea to monitor and analyze the disease of the patients they are suffering from and will alert the particular or nearby hospitals about the patient if there is any danger. So that the patient's medical condition can be identified and can be cured.

Keywords--- Prediction, data analytics, machine learning, data mining, heart diseases.

I. INTRODUCTION

The Main object of this paper is to save the lives of the people who are suffering from Heart-strokes, Blood pressures & Some critical health problems by sensing their disease.

Our proposed system will show a solution to some of the problems in the present medical department. We are using sensors for monitoring patient health. These sensors will update the patient details frequently and check with the existing database values.

If there are any changes in the values, then the comparison will be started. Based on those comparisons the disease is identified. If the compared values are too different from the pre-defined database values, then the problem is identified. It will be informed to the patient relatives and to the nearby or particular hospital and patient is taken to the hospital.

II. PROPOSED SYSTEM

Here we consider the dataset from UCI repository and Divide them into two groups train dataset and test dataset. Now, the dataset is trained in such a way that it is divided into n models.

Now, we will combine all these models predictions and vote for the highest probability. Based on the voting we will predict the model.

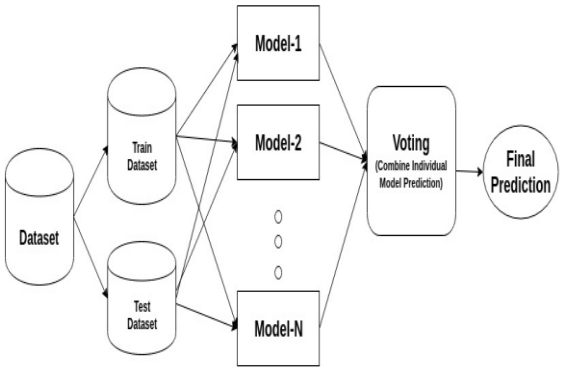

Fig. 1: Shows Dataset Classification

A. Cross Fold Validation

Cross-validation is a technique which is used to apply machine learning algorithms on the particular part of a dataset. There are various ways to do cross validation like validation set approach, LOOCV, etc. Out of these methods in this paper we are using K-Cross Validation Method as follows

1) Randomly choose K Value and divide the entire dataset into K folds
2) For each k-fold in our dataset, build your model on k – 1 folds of the dataset.
3) Find the error on each iteration
4) Repeat until all the K-folds are completed
5) The average of your k recorded errors is called the cross-validation error and will serve as your performance metric for the model.

It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

Accuracy

Accuracy is one metric for evaluating classification models.

Accuracy=Number of correct predictions/Total number of predictions

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

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

Where *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.

Revised Version Manuscript Received on 22 February, 2019

Shaik Akbar, Professor, PSCMRCET, Vijayawada. AP, India.(e-mail: dr.shaikakbar@gmail.com)

P. Sri Silpa, Assistant Professor, PSCMRCET, Vijayawada. AP, India. (e-mail: silpa@pscmr.ac.in)

Anand Thota, Associate Professor, PSCMRCET, Vijayawada. AP, India. (e-mail: anand4uammu@gmail.com)

K. Nageswara Rao, Principal, PSCMRCET, Vijayawada. AP, India. (e-mail: principal@pscmr.ac.in)

Let's try calculating accuracy for the following model that classified 303 Heart Patient Details NO (the positive class) or YES (the negative class):

True Positive (TP): • Reality: Patient Will not Suffer from Heart Attack • ML model predicted: Patient Will not Suffer from Heart Attack • Number of TP results: 143	False Positive (FP): • Reality: Patient Will Suffer from Heart Attack • ML model predicted: Patient Will not Suffer from Heart Attack • Number of FP results: 21
False Negative (FN): • Reality: Patient Will not Suffer from Heart Attack • ML model predicted: Patient Will Suffer from Heart Attack • Number of FN results: 27	True Negative (TN): • Reality: Patient Will Suffer from Heart Attack • ML model predicted: Patient Will Suffer from Heart Attack • Number of TN results: 112

Accuracy=TP+TN/TP+TN+FP+FN=143+112/143+112+21+27=0.84

Recall: The ability of a model to find all the relevant cases within a dataset.

Mathematically, recall is defined as follows:

Recall=TP/TP+FN =143/143+27=0.842

Precision: The ability of a classification model to identify only the relevant data points.

Precision is defined as follows:

Precision=TP/TP+FP=143/143+21=0.842

F-Measure

- A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure.
- F-measure = 2 * ((Precision*Recall) / Precision + Recall)) or = 2*TP / (2*TP) + FP + FN =2*((0.842*0.842)/(0.842+0.842))=0.842.

B. ROC curve

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate

True Positive Rate (**TPR**) is a synonym for recall and is therefore defined as follows:

TPR=TP/TP+FN

False Positive Rate (**FPR**) is defined as follows:

FPR=FP/FP+TN

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

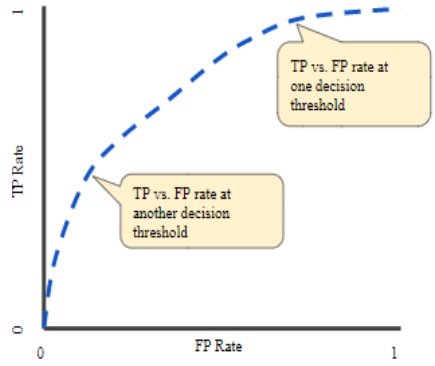

Fig. 2: Shows ROC Curve

C. Attributes

Table I: Attribute List

S.No	Name of the Attribute	Description	Type of Attribute
1	Age	age of the patient in the years will be stored	Real
2	Sex	Sex of the patient will be stored as 2 types of values. a. Male will be represented with 1 b. Female will be represented with 0	Binary
3	ChestPain	Chest Pain Type (There are 4 types of chest pain) 1: typical 2: asymptomatic 3: nonanginal 4: nontypical	Nominal
4	RestBP	resting blood pressure, It will store the bp level of patient	Real
5	Chol	serum cholestoral in mg/dl	Real
6	Fbs	fasting blood sugar will be stored as 2 types of values a. fbs >= 120 mg/dl will be represented as 1 b. fbs < 120 mg/dl will be represented as 0	Binary
7	RestECG	resting electrocardiographic results will be stored as 3 types of values a. normal condition will be represented as 0 b. ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) will be represented as 1 c. Probable or definite left ventricular hypertrophy by Estes' criteria will be represented as 2	Nominal
8	MaxHR	maximum heart rate achieved	Real
9	ExAng	exercise induced angina will be stored as 2 types of values a. yes will be represented as 1 b. no will be represented as 0	Binary
10	Oldpeak	oldpeak = ST depression induced by exercise relative to rest	Real
11	Slope	the slope of the peak exercise ST segment will be stored as 3 types of values a. unsloping will be represented as 1 b. flat slope will be represented as 2 c. down sloping will be represented as 3	Ordered
12	Ca	number of major vessels (0-3) colored by flourosopy	Real
13	Thal	Thal will be stored as 3 types of values 3 = normal; 6 = fixed defect; 7 = reversable defect	Nominal

D. Attribute Analysis

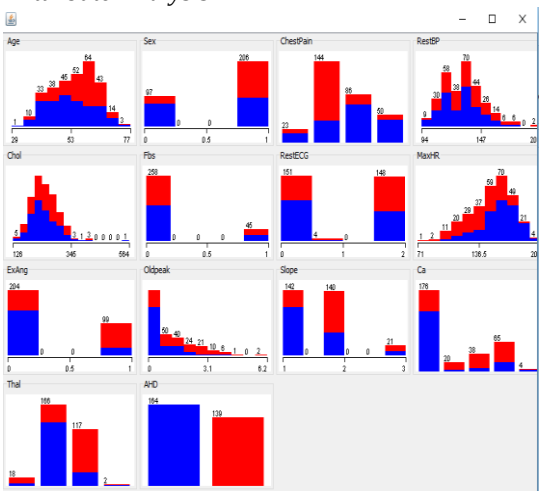

Fig. 3: Attribute Analysis graph

E. Naive Bayesian Algorithm

Naive Bayes is among one of the most simple and powerful algorithms for **classification** based on Bayes' Theorem with an assumption of independence among predictors. Naive Bayes model is easy to build and particularly useful for very large data sets. There are two parts to this algorithm:

- **Naive**
- **Bayes**

The Naive Bayes classifier assumes that the presence of a feature in a class is unrelated to any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that a particular model and that is why it is known as **"Naive".**

Given a Hypothesis **H** and evidence **E,** Bayes' Theorem states that the relationship between the probability of Hypothesis before getting the evidence **P(H)** and the probability of the hypothesis after getting the evidence **P(H|E)** is :

$$P(H|E) = P(E|H).P(H) / P(E)$$

This relates the probability of the hypothesis before getting the evidence **P(H)**, to the probability of the hypothesis after getting the evidence, **P(H|E)**. For this reason, is called the **prior probability**, while **P(H|E)** is called the **posterior probability**. The factor that relates the two, **P(H|E) / P(E)**, is called the **likelihood ratio**. Using these terms, Bayes' theorem can be rephrased as:

"The posterior probability equals the prior probability times the likelihood ratio."

Application in Medical Field

Nowadays modern hospitals are well equipped with monitoring and other data collection devices resulting in enormous data which are collected continuously through health examination and medical treatment. One of the main advantages of the Naive Bayes approach which is appealing to physicians is that **"all the available information is used to explain the decision"**. This explanation seems to be "natural" for medical diagnosis and prognosis i.e. is close to the way how physicians diagnose patients.

When dealing with medical data, Naïve Bayes classifier takes into account evidence from many attributes to make the final prediction and provides transparent explanations of its decisions and therefore it is considered as one of the most useful classifiers to support physicians' decisions.

F. Random Forest Algorithm

Random Forest is a supervised learning algorithm. Like you can already see from its name, it creates a forest and makes it somehow random. The „forest" it builds, is an ensemble of Decision Trees, most of the time trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Random Forest pseudo code

1) Randomly select **"k"** features from total **"m"** features.
    1. Where **k << m**
2) Among the **"k"** features, calculate the node **"d"** using the best split point.
3) Split the node into **daughter nodes** using the **best split**.
4) Repeat **1 to 3** steps until "l" number of nodes has been reached.
5) Build forest by repeating steps **1 to 4** for "n" number times to create **"n" number of trees**.

Random forest prediction pseudo code

To perform prediction using the trained random forest algorithm uses the below pseudo code.

1. Takes the **test features** and use the rules of each randomly created decision tree to predict the oucome and stores the predicted outcome (target)
2. Calculate the **votes** for each predicted target.
3. Consider the **high voted** predicted target as the **final prediction** from the random forest algorithm.

F. ADA Boost Algorithm

Ada-boost classifier combines weak classifier algorithm to form strong classifier. A single algorithm may classify the objects poorly. But if we combine multiple classifiers with selection of training set at every iteration and assigning right amount of weight in final voting, we can have good accuracy score for overall classifier.

AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem.

The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level. Because these trees are so short and only contain one decision for classification, they are often called decision stumps.

Each instance in the training dataset is weighted. The initial weight is set to:

$$weight(xi) = 1/n$$

Where xi is the i'th training instance and n is the number of training instances.

How does the AdaBoost algorithm work?

It works in the following steps:

1) Initially, Adaboost selects a training subset randomly.
2) It iteratively trains the AdaBoost machine learning model by selecting the training set based on the accurate prediction of the last training.
3) It assigns the higher weight to wrong classified observations so that in the next iteration these observations will get the high probability for classification.
4) Also, It assigns the weight to the trained classifier in each iteration according to the accuracy of the classifier. The more accurate classifier will get high weight.
5) This process iterate until the complete training data fits without any error or until reached to the specified maximum number of estimators.
6) To classify, perform a "vote" across all of the learning algorithms you built.

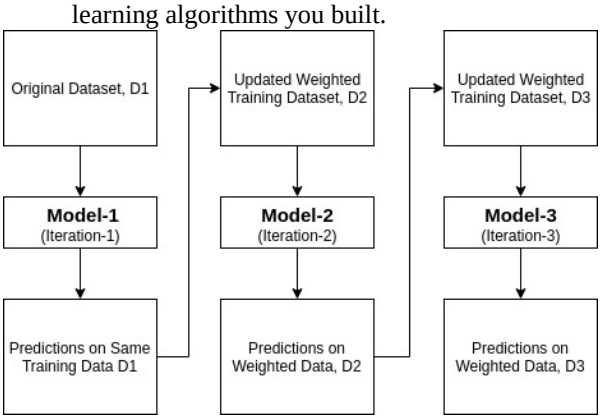

Fig. 4: Ada Boost working process

III. RESULTS

Table 2: Results Comparison

	Recall	Precision	Accuracy
Naive Bayesian Algorithm	0.842	0.842	84.1
Random Forest Algorithm	0.818	0.82	81.8
ADA Boost Algorithm	0.818	0.818	82.4

IV. CONCLUSION AND FUTURE WORK

In this paper we compared the heart attack dataset against 3algorithms. From the above comparison we found that naive bayesian algorithm has good results when compared to others but if the dataset size increases it becomes difficult for us to train the dataset with the traditional algorithms so in future work we can train the real time data by taking with the help of IoT Sensors and apply convolution neural networks to obtain the better results.

REFERENCES

1. Manikantan, V. and S. Latha. "Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods." (2013).
2. Shadab Adam Pattekari and Alma Parveen," Prediction system for heart disease using naïve bayes", International Journal of Advanced Computer and Mathematical Sciences, vol.3,pp 290- 294,2012.
3. Soni, Jyoti & Ansari, Ujma & Sharma, Dipesh & Soni, Sunita. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. International Journal of Computer Applications. 17. 43-48. 10.5120/2237-2860.
4. Agrawal, Rakesh & Imielinski, Tomasz & Swami, Arun. (1993). Mining Association Rules Between Sets of Items in Large Databases, SIGMOD Conference. 10.1145/170036.170072.
5. Hnin Wint Khaing, "Data mining based fragmentation and prediction of medical data," 2011 3rd International Conference on Computer Research and Development, Shanghai, 2011, pp. 480-485. doi: 10.1109/ICCRD.2011.5764179
6. Masilamani, Anbarasi & , ANUPRIYA & Iyenger, N Ch Sriman Narayana. (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology. 2.
7. D. Burdick, M. Calimlim and J. Gehrke, "MAFIA: a maximal frequent itemset algorithm for transactional databases," Proceedings 17th International Conference on Data Engineering, Heidelberg, Germany, 2001, pp. 443-452. doi: 10.1109/ICDE.2001.914857
8. Srinivas, Kavitha et al. "SURVEY ON PREDICTION OF HEART MORBIDITY USING DATA MINING TECHNIQUES." (2011). DOI:10.5121/ijdkp.2011.1302
9. S.Vijayarani, M. Divya, " An Efficient Algorithm for Generating Classification Rules", IJCST ,vol. 2, Issue 4, 2011.