Health Care Analysis Using Random Forest Algorithm

Deepa, Karthik Kumar, Dharneshawar, Rohith, Bharath
School of Information Technology and Engineering
VIT University, Vellore, Tamil Nadu
***Corresponding author: E-Mail: Karthikkumar732@gmail.com**

ABSTRACT

Data Mining is the most inspiring area of research that become most popular in health organization. It also plays an important part to uncover new patterns in medicinal services association which thusly accommodating for all the parties associated with this field. This project intend to form a diagnostic model of the various diseases mainly liver cancer based on the symptoms by using data mining technique such as classification in health domain. In this project, we are going to use algorithms like Random forest, Naive Bayes which can be utilized for health care diagnosis. Performances of the classifiers are compared to each other to find out highest accuracy. This also helps us to find out persons who are affected by the infection. The test information is completed utilizing some benchmark information.

KEY WORDS: Naive Bayes, J48, Random Forest, Decision Tree, Classification Algorithms.

1. INTRODUCTION

Liver Disease has become the major reason for many deaths in developed Countries. The effective way to decrease liver disease deaths is to detect it earlier. Early diagnosis requires an accurate and reliable diagnosis technique that can be used by physicians to separate liver tumors. In this project data mining methods that have been used in liver disease dataset to know people who are affected by disease and problems are surveyed. We will represent the analysis of data mining methods applied on data sets. i.e., accuracy of those dataset.

With this project, we will apply classification algorithms for the given data to determine which algorithm is best and suitable to classify. It is also used to find out whether the person is having the disease or not. Data Mining also inspires wide range of practical solutions that make the most of resource utilization and prolong the lifetime of the system.

Naive Bayes: In data mining, naive Bayes classifiers are the family of simple probabilistic classifiers based on applying Bayes' hypothesis with solid (credulous) autonomy presumptions between the components. Innocent Bayes considered extensively since the 1950's. It was introduced with a different name into the text rescue community in the early 1960's, and remains a standard method for text categorization, the problem of judging documents belonging to only one category or the other one with word frequencies of the elements.

With suitable pre-processing, it became competitive in this domain with highly advanced methods including bolster vector machines. It likewise discovers application in programmed medicinal analysis. Credulous Bayes classifiers are highly scalable, requiring a number of parameters direct in the quantity of factors in a learning issue.Greatest probability preparing should be possible by assessing a shut frame expression, which takes straight time, instead of by costly iterative guess as utilized for some different sorts of classifiers. In the measurements and software engineering writing, Naive Bayes models are known under an assortment of names, including straightforward Bayes and autonomy Bayes. Every one of these names reference the utilization of Bayes' hypothesis in the classifier's choice administer, yet innocent Bayes is not (really) a Bayesian strategy.

Decision Tree: A decision tree is a flowchart structure in which each internal node denotes a test on a characteristic, where each branch signifies the result of the test, and each leaf node denotes a class label. The paths from the root to leaf denotes classification rules. In tree the interrelated diagram are used as the analytical, visual and decision support tool, where the apparent values are calculated. A decision tree contains three types of nodes: one is decision nodes – denoted by squares, Chance nodes denoted by circles, and the other is End nodes –denoted by triangles you begin a decision tree with a conclusion that need to make. Draw a small square to represent the tree towards the left of a great piece of paper. From this draw out lines in the direction of the right for each likely solution, and write the solution along the line. At the end of each line, reflect the results. If the result of decision is undefined, then draw a small circle. If the result is alternate decision then you need to make, draw a different square. Squares represent decisions, circles that denotes uncertain result. Starting from the new decision squares from the diagram, draw the lines denoting options that could be select. From the circles we can even draw lines that represent possible outcomes. Finally make a short-term note on the line by saying what it says.

Random Forest: Random forest is a bootstrapping algorithm with the cart model. It built multiple trees with different initial variables and consider a sample of 100 observation and 5 random samples chosen initial variable to build a cart model. It will repeat the same process 10 times and they make a final prediction for each observation. Final prediction is a function of each prediction. This final one can simply be a mean of each prediction. Basically this process is done in the Weka tool. Weka tool is the machine learning tool which contains a large number of data science algorithms which can be used for classification, prediction and to find the missing values.

It is a collective learning method for organization and other tasks that operate by building a gathering of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the single trees. The process for the random forests was shaped by Tin Kam Ho used the random subspace method, in Ho's formulation, it is a process to implement the "stochastic discrimination" method to classification planned by Eugene Kleinberg. The above process defines the original bagging algorithm for trees. Random forests change in one from this general arrangement: they had used a modified tree learning algorithm that selects, at each person in learning process, a random subset of the features. This process is sometimes called "feature bagging". The importance of this correlation is the multiple trees in a normal bootstrap sample: if one or more additional features are correct predictors for the response variable, these features will be sensibly chosen in many of the *B* trees, making them to be correlated.

Tree bagging: The algorithm for random forests carries the method of bootstrap bagging, to tree learners. Given a set $X = x_1 ... x_n$ with responses $Y = y_1 ... y_n$, bagging regularly selects a sample with additional of the training set and fits the trees to these samples:

For $b = 1... B$:

- Sample, with replacement, *B* training examples from *X*, *Y*; call these X_b, Y_b.
- Training a decision or a regression tree f_b on X_b, Y_b.

Architecture Diagram:

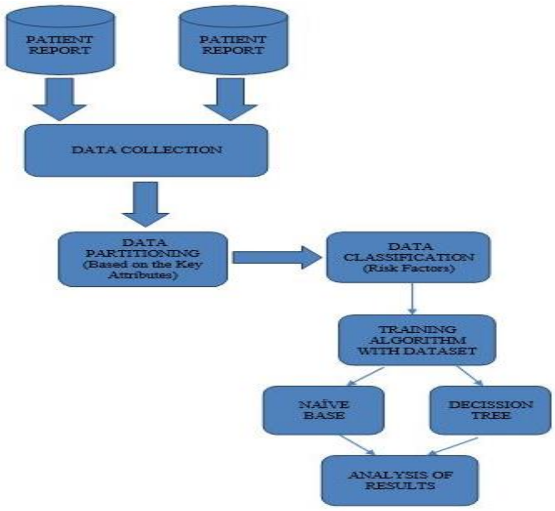
PATIENT REPORT
PATIENT REPORT
DATA COLLECTION
DATA PARTITIONING (Based on the Key Attributes)
DATA CLASSIFICATION (Risk Factors)
TRAINING ALGORITHM WITH DATASET
NAÏVE BASE
DECISSION TREE
ANALYSIS OF RESULTS

Figure.1. System Model

From the figure 1 our system model shows the data is initially collected from the sources. The data should be distributed uniformly where it should not contain any missing values. Out of many attributes one attribute contains all the classification values where the patient is suffering from the certain disease or not. Data needs to be trained initially and after training the data later our algorithm can efficiently can classify or can be predicted the disease. Algorithm over fit problem can be reduced by pruning the dataset. Unnecessary attribute values needs to be removed so that our algorithm efficiency can be removed. By using naïve Bayes we predict all health data and alter us predict using random forest these concludes which model can shows the highest accuracy.

3. EXPERIMENTAL RESULTS

A confusion matrix is used to describe the classification model performance for the test data which the true values are known. The confusion matrix itself comparatively simple to recognize, but the related terminology might be confusing.

The most basic concepts need to be known:

True positives (TP): We predict yes if they have disease and yes they have disease.

True negatives (TN): If they don't have the disease then we predict no.

False positives (FP): We predicted yes but actually they do not have disease. ("Type I error.")

False negatives (FN): Prediction can be no but actually they are suffering from the disease. ("Type II error.")

Accuracy: Overall, how frequently is the classifier correct?

Accuracy = (TP+TN)/total

Misclassification Rate: How frequently is it wrong, overall?

Misclassification Rate = (FP+FN)/total

True Positive Rate: It is actually yes, how frequently it was predict yes?

It is also called as "Recall" or "Sensitivity".

True Positive Rate = TP/actual yes

False Positive Rate: When it is truly no, how frequently does it was predict yes?

True Positive Rate = FP/actual no

Specificity: When it is really no, how frequently does it predict no?

Specificity = TN/actual no

Precision: When it was predicted yes, how frequently is it correct?

Precision = TP/predicted yes

Prevalence: How frequently does the yes condition truly occur in our sample?

Prevalence = actual yes/total

Cohen's Kappa: This basically shows how well the classifier performed after comparing to how well it would have been performed simply by chance. In other words, if our model will have maximum Kappa score only if there is a large difference between the accuracy and the null error rate.

F Score:F score is the average of the between precision true positive rate and

ROC Curve:

Table.1. Accuracy percentage of the individual algorithms

Algorithm	Kappa statistic	Mean Absolute error	Root mean squared error	Relative absolute error	Roc area	Accuracy Percentage
Naïve Bayes	0.244	0.44	0.654	107.69%	0.72	55.74%
J48	0.177	0.329	0.481	80.46%	0.67	68.78%
Random Forest	0.194	0.339	0.422	83.07%	0.75	70.50%

This is normally used in graph that summarizes the performance of the classifier over all possible thresholds. It is formed by scheming the TP Rate (y-axis) in contrast to the FP Rate (x-axis) as you differ the threshold for assigning the observations to a given class.

All the performance measurements are calculated and tabulated in the table 1. From the table we can predict and can finalize random forest algorithm has the highest accuracy. Random forest has the greatest accuracy when compared to other algorithms. Highest accuracy means highest classification rate ie.., it has the highest prediction rate. We can use random forest algorithm for any data classification in the medical growth to find and predict any disease. In our research we used liver dataset to find accuracy and other performance measurements.

4. CONCLUSION

Here we will end up our paper by implementing naïve Bayes and decision tree algorithm in Anaconda navigator by using data mining model created by us. This model can be created for any algorithm that are related to data sciences. We finally made visualizations of data collected by health care analysis through data mining. Our final predictions shows that accuracy of Random forest is greater than the naïve Bayes. Accuracy is the one if the performance evaluation criteria to accept the model and to be used for thee further predictions. So we can conclude that prediction any disease mainly classification can be done accurately and can be classified using random forest algorithm.

REFERENCES

Lin R.H, an Intelligent Model for Liver Disease Diagnosis. Artificial Intelligence in Medicine, 47 (1), 2009, 53-62.

Onisko A, Druzdzel M.J and Wasyluk H, A Bayesian Network Model for Diagnosis of Liver Disorders. In Proceedings of the Eleventh Conference on Biocybernetics and Biomedical Engineering, 2, 1999, 842-846.

Rajeswari P and Reena G, Analysis of Liver Disorder using Data mining Algorithm. Global Journal of Computer Science and Technology, 10 (14), 2010, 48-52.

Ramana B.V, Babu M.S.P and Venkateswarlu N.B, A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. Global Journal of Database Management Systems, 3 (2), 2011, 101-114.