International Journal of Research in Advent Technology, Vol.7, No.4, April 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Diagnosis of Indian Patients Liver Disease Using Machine Learning Techniques

B. Muni Lavanya[1], B. Hari Kishore[2], R. Sreekala[3], S. Sneha[4]
[1]*Assistant Professor(Adhoc), Dept of CSE, JNTUCEP,Pulivendula,AP,India,*
[2,3,4] *Student, Dept of CSE, JNTUCEP,Pulivendula,AP,India,*
[1]*munilavanya45@gmail.com,* [2]*kishorehari9464@gmail.com,* [3]*ragalasreekala@gmail.com,*
[4]*sankesneha25@gmail.com*

Abstract-Diagnosis of liver disease at a initial stage is important for better treatment. It is very difficult task for medical researchers to predict the liver disease in the early stages owning the delicate symptoms. If the symptoms become apparent when it is too late. To overcome this issue, the aim of this project is to improve liver disease diagnosis using machine learning approaches. The main objective of this research is to use different classification algorithms to identify the patients have liver disease or not. This project also aims to compare the classification algorithms based on their evaluation metrics.

Index Terms- Machine Learning, Liver Patients, Classification algorithms, SVM, Decision tree, Ada boost, accuracy.

1. INTRODUCTION

Problems with liver patients are not easily discovered in An early stage as it will be functioning normally even when It is partially damaged. An early diagnosis of liver problems will increase patient's survivalrate. The widespread occurrence of Liver infection in India is contributed due to deskbound Lifestyle, increased alcohol consumption and smoking. Given a dataset containing various attributes of 583 Indian patients, define classification algorithms. To apply different classification algorithms on the Indian patient liver disease dataset than choose the best algorithms based on the accuracy which can identify whether a person is suffering from liver disease or not. The remainder of the paper is presented as follows: Related work in section II, Proposed Work in section III, Metrics in section IV, Implementation in section V, Dataset Description in section VI, Data Preprocessing in section VII, Classification Techniques in section VIII, Results in section IX and conclusion in section X.

2. RELATED WORK

In recent research works, several data mining models have been developed to aid in diagnosis of liver diseases in the medical field by the physicians such as diagnosis support system , expert system, intelligent diagnosis system, and hybrid intelligent system. The liver disorder data warehouse contains the screening the data of liver disorder patients. Previously, the data warehouse is pre-processed to increase efficiency. Later on as compared to Machine learning algorithms techniques data mining techniques are lagging behind in performance and accuracy.

Different supervised machine algorithms learning algorithms derived from the WEKA data mining tool. Which include: Naive Bayes, Kstar, FT Tree.

3. PROPOSED WORK

Using data mining techniques predicting patients liver disease is a time consuming task which degrades patients survival rate, By Appling different machine learning techniques An early diagnosis of liver problems will increase patients' survival rate based on accuracy and F score to find the best suitable algorithm for diagnosis of liver disease which gives best performance. It added a greater advantage to medical field.

Some of the classification algorithms used are:

1. Naïve predictor Bench Mark
2. Decision trees
3. Support Vector Machine
4. Ada boost

4. METRICS

In this work accuracy score is used as evaluation metric for prediction of liver disease. The performance of a model cannot be assessed by considering only the accuracy, because there is a possibility for misleading. Therefore this experiment considers the F1 score along with the accuracy for evaluation.. This is because depending on the context like severity of disease, sometimes it is more important that an algorithm does not wrongly predict a disease as a no disease.

Thus, here we will use F-beta score as a performance metric, which is basically the weighted harmonic mean of precision and recall. Precision and Recall are defined as:

Precision=TP/(TP+FP),

Recall=TP/ (TP+FN), where

TP=True Positive

FP=False Positive

FN=False Negative

In the same vein, F-beta score is:

F-beta score = (1+β^2)*precision*recall/((β^ 2*precision)+recall)

We can use F-beta score as a metric that considers both precision and recall

Additionally, one more metric called as Receiver Operating Characteristics (ROC) curve will be used. It plots the curve of True Positive Rate and the False Positive Rate for a given algorithm, with a greater area

under the curve indicating a better True Positive Rate for the same False Positive Rate, indicating the usefulness of the classifier.

5. IMPLEMENTATION

The Indian Liver Patient Dataset comprised of 10 different attributes of 583 patients. The patients are classified as either 1 or 2 on the basis of disease data set. The table describes description of the dataset. Attribute and attribute type are presented as columns. As clearly visible from the table, all the features are real valued integers but not sex. The feature Sex is converted to numeric value (0 and 1) in the data pre-processing step.

	age	gender	tot_bilirubin	direct_bilirubin	tot_proteins	albumin	ag_ratio	sgpt	sgot	alkphos	is_patient
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.90	1
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	0.89	1
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	1.00	1
4	72	Male	3.9	2.0	195	27	59	7.3	2.4	0.40	1

6. DATASET DESCRIPTION

It is a multivariate data set, contain 10 variables that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. All values are real integers. This data set contains records of Liver patirntsupto 416 and non liver patient records of upto 167. The data set was collected from Andhra Pradesh, India.

7. DATA PREPROCESSING

Some datasets contain irrelevant information, noise, missing values, and so on. These datasets should be handled for better results of the data mining process. Data preprocessing includes data cleaning, preparation, analysis transformation, and dimensionality reduction, which convert the raw data into a form that is suitable for further processing.

As explained in the section 'Exploring the data', rows having 'Null' values were removed from the dataset. Thereafter, log transformation was applied to features which were showing a skewed pattern (Albumin, Direct Bilirubin, A/G ratio, Total Bilirubin, Total Protein).

Thereafter, all columns in the dataset except 'Gender' are normalized. We use MinMaxScaler here as StandardScaler gives very low values here, with some in the order of 10^-16, which might be difficult to relate to and visualize.

Then we use pd.get_dummies() method to one-hot encode the feature 'gender' as well as the label 'is_patient' with the integer '1' representing the presence of disease and '2' representing the not presence of disease.

8. CLASSIFICATION TECHNIQUES

Three supervised learning approaches are selected for this problem. Care is taken that all these approaches are fundamentally different from each other, so that we can cover as wide an umbrella as possible in term of possible approaches. For example- We will not select Random Forest and Ada Boost together as they come from the same family of 'ensemble' approaches. The choice of algorithms was influenced from these source:

https://stackoverflow.com/questions/2595176/which-machine-learning-classifier-to-choose-ingeneral

For each algorithm, we will try out different values of a few hyper parameters to arrive at the best possible classifier. This will be carried out with the help of grid search cross validation technique. For these dataset we apply the different supervised learning algorithms there are:

1. Ada Boost Classifier:

Adaboost is a boosting type ensemble learner. Several weak learning hypothesis are combined to form strong model. Random chance is not considered much than weal hypothesis. However, it's the combination of all of these independent weak learning hypotheses what makes the model more capable of predicting accurately on unseen data than each of the individual hypothesis would. n_estimators: The maximum number of estimators at which boosting is terminated..

learning_rate: Learning rate reduces the contribution of each classifier by learning_rate. There is a trade-off between learning_rate and n_estimators.

Advantages:

- AdaBoost is the of the ensemble method. The AdaBoost is more robust than singl estimators, have improved generalizability.
- Simple models can be combined to build a complex model, which is computationally fast Disadvantages:
- If we have a biased underlying classifier, it will lead to a biased boosted model.
- AdaBoost is sensitive to noisy data and outliers. In some cases,it can be less susceptible to the over fitting problem than most learning algorithms.

2. Support Vector Machine:

SVM aims to find an optimal hyper plane that separates the data into different classes, using a method called as kernel to project data points belonging to a particular class into different dimensions, so that a hyperplane can easily pass through and maintain the largest possible distance between itself and these data points.

Advantages:

- Performs well with high dimensional data. SVM's are very good when we have no idea on the data.
- Works well with even unstructured and semi structure data like text, Images.
- The kernel trick is strength of SVM. By using the kernel function to solve the complex problem.
- The SVM model have generalization in practice, the risk of over fitting is less in SVM.

Disadvantages:

- Choosing the good kernel function is not easy and it take long training time for large datasets.
- Final model, variable weights and individual impact are difficult to interpret.

3. Decision Tree Classifier:

The decision tree can be used to visually and explicitly represent decisions and decision making.. Though a commonly used tool in data mining for deriving a strategy

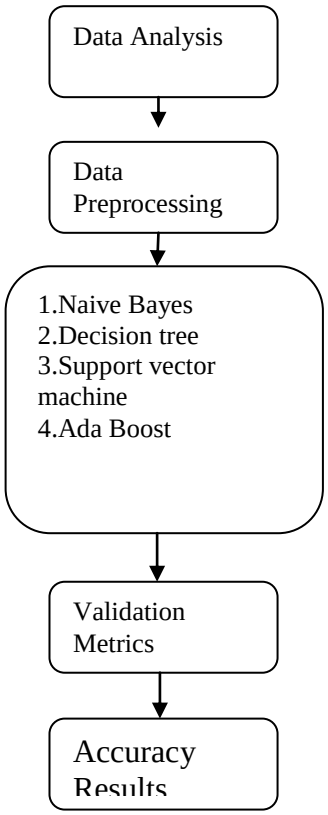
to reach a particular goal, its also widely used in machine learning.

Advantages:

- Doesn't require much data pre-processing, and can handle data which hasn't been normalized, or encoded for Machine Learning Suitability.
- Simple to understand ,visualize and interpret.

Disadvantages:

- Complex Decision Trees do not generalize well to the data and can result in over fitting.



9. RESULTS

Since the size of dataset is small at present , there is not much difference between training and testing times of different algorithms. However, for the sake of comparison, these times have been displayed in the 'Implementation' sub-heading of 'Analysis' section. Adaboost Classifier consumes maximum time during training and good accuracy score, ,F_score. From this dataset use three models based on the accuracy_score,,F_score we decide Adaboost is best suitable for this dataset.

Accuracy: The accuracy of a classifier is the percentage of the test set tuples that are correctly classified by the classifier.

$$\text{Accuracy} = \frac{TN+TN}{TP+TN+FN+FP}$$

Recall

Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples.

Recall is given by the relation:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Precision

To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples.

Precision is given by the relation:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Classification Algorithm	F_score Test	F_score Train	Accuracy Test	Accuracy Train
Adaboost	0.714286	0.845865	0.655172	0.805000
SVC	0.679702	0.771670	0.629310	0.730000
Decision Tree	0.659472	1.000000	0.577586	1.00000

10. CONCLUSION

In this paper, we have proposed methods for diagnosing liver disease in patients using machine learning techniques. The four machine learning techniques that were used include SVM,Ada boost, Decision tree, Naïve Bayes. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metric.Ada boost techniques resulted highest accuracy of 75.6%. From the above results Adaboost plays a key role in shaping improved classification accuracy of a dataset.

REFERENCES

[1] http://ijsetr.org/wpcontent/uploads/2015/04/IJSET R-VOL-4-ISSUE-4-816-820.pdf
[2] https://globaljournals.org/GJCST_Volume10/9-Analysis-of-Liver-Disorder-Using-Datamining-Algorithm.pdf
[3] http://article.sapub.org/10.5923.j.ajis.20140401.02.html
[4] https://www.ripublication.com/ijaer17/ijaerv12n17_03.pdf
[5] Paul R. Harper, A review and comparison of classification algorithms for medical decision making
[6] https://www.tutorialspoint.com/mahout/mahout_machine_learning.htm
[7] https://epdf.tips/queue/schiffs-diseases-of-the-liver-2-volume-set-10th-ed.html
[8] https://www.kaggle.com/uciml/indian-liver-patient-records