International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 16 (2018) pp. 12732-12737

Application of Machine Learning Classification Algorithms on Hepatitis Dataset

K. Santosh Bhargav
Department of Computer Science and Engineering
GITAM Institute of Technology, GITAM
Visakhapatnam, India.

T. Divya Kumari
Department of Computer Science and Engineering
G. Narayanamma Institute of Technology & Science
Hyderabad, India.

Dola Sai Siva Bhaskar Thota.
Department of Computer Science and Engineering
GITAM Institute of Technology, GITAM
Visakhapatnam, India.

Vikas B.
Department of Computer Science and Engineering
GITAM Institute of Technology, GITAM
Visakhapatnam, India.

Abstract

Machine learning is a method or a field used to conceive complicated models and algorithms for predictive analysis. The different analysis models help researchers, data scientists, engineers, and analysts to produce proper results and decisions , has numerous applications such as building predictive models which can be extremely beneficial in the healthcare industry. Hepatitis C is a liver disease caused by the hepatitis C virus and some cancers such as liver cancer and lymphomas in humans. In order to help in the diagnosis of hepatitis C, classification techniques such as Logistic Regression, Decision Tree, Linear Support Vector and Naive Bayes can be applied to classify real time hepatitis C data based on an established training set. In this paper, an attempt has been made to classify weather the person will live or die using the accuracies and other performance measures of the different machine learning algorithms.

Keywords: Machine Learning, Hepatitis C, Logistic Regression, Decision Trees, Linear Support Vector, Naive Bayes, Accuracy, Recall, Support, Precision.

INTRODUCTION

Hepatitis C is an infection that is caused due to the Hepatitis C virus attacking the liver leading to its inflammation. The virus can lead to both acute and chronic hepatitis. Hepatitis C can last either for a few weeks or for a lifetime [1]. The incubation period of hepatitis C is 2 weeks to 6 months. According to statistics from the World Health Organisation, "Following the initial infection, approximately 80% of people do not exhibit any symptoms." Few symptoms of Hepatitis C are Jaundice, Stomach pain, dark urine, grey-coloured faeces, joint pain, Loss of appetite, Nausea, Fatigue. The hepatitis C virus is a bloodborne virus, which means that it majorly spreads through direct contact of small quantities of blood [2]. Such contact could be due to various reasons such as unsafe injection practices, unsafe health care, and the transfusion of unscreened blood and blood products and sometimes from birth.

According to World Health Organisation globally there are 71 million people estimated suffering from this infection and nearly 399,000 people die each year from hepatitis C. At present there is no vaccine for hepatitis C, however continuous research is going on to find the cure. The prevention of HCV infection can be done by maintaining proper hygiene, safe and appropriate of health care injections, safe disposal sharps and wastes, testing the donators blood, educating about the virus and seasonal healthcare check-up.

Machine learning [3] is a field that automatically learns and makes accurate prediction based on past observations. The goal of machine learning is to give highly accurate predictions on given data. The advantages of using machine learning is that it is often much more accurate than human rules. It doesn't require a human expert or a programmer. It is also cheap and flexible- It can be applied to any learning task. However, it sometimes requires a lot of labelled data and it is hard to get perfect accuracy. Machine learning [4] plays a very important role in health care reform. For earlier prevention one can estimate the possibility of HCV of one-self by using various machine learning techniques such as logistical regression, Naive Bayes, linear SVM, decision tree etc. This will help the patient to take treatment in the earlier phase of virus and stopped from being amplified. In this paper, a comparative study on the performance measures of different classification techniques in machine learning applied of Hepatitis C data has been presented.

DATA COLLECTION

The hepatitis C data set has been referenced from UCI Repository [5]. The database consists of 155 samples. It has 20 attributes including the Class label attribute. The attribute values have been derived after carrying out numerous medical tests. Machine Learning Algorithms have been applied on this dataset to classify the records into two categories: Live or Die which are the values of the Class Label. There are 14 binary attributes and 6 numerical attributes in the data set, as shown below:

Number of Instances: 155
Number of Attributes: 20 (including the class attribute)

Attribute information is depicted in Table 1.

TABLE I. *HEPATITIS C DATASET*

S.NO.	ATTRIBUTE	VALUE
1	CLASS	DIE, LIVE
2	AGE	10, 20, 30, 40, 50, 60, 70, 80
3	SEX	male, female
4	STEROID	no, yes
5	ANTIVIRALS	no, yes
6	FATIGUE	no, yes
7	MALAISE	no, yes
8	ANOREXIA	no, yes
9	LIVER BIG	no, yes
10	LIVER FIRM	no, yes
11	SPLEEN PALPABLE	no, yes
12	SPIDERS	no, yes
13	ASCITES	no, yes
14	VARICES	no, yes
15	BILIRUBIN	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16	ALK PHOSPHATE	33, 80, 120, 160, 200, 250
17	SGOT	13, 100, 200, 300, 400, 500
18	ALBUMIN	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19	PROTIME	10, 20, 30, 40, 50, 60, 70, 80, 90
20	HISTOLOGY	no, yes

CLASSIFICATION

In machine learning and statistics, classification algorithms classify tuples into a set of categories [6]. It is a supervised learning approach in which the computer program learns from the data input (i.e. trained example) given to it and then uses this learning to classify new observation based on classification rules. The input data set may be either a bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Classification algorithms have various practical applications in different fields such as bioinformatics, natural language processing, market segmentation and text categorization. It is used for speech recognition, facial detection, filtering spam messages, handwriting recognition, understanding spoken language, bio metric identification, document classification etc.
There are many classification techniques and algorithms, out of which some have been listed below:-

1. Linear Classifiers: Logistic Regression, Naive Bayes Classifier
2. Support Vector Machines
3. Decision Trees
4. Random Forest
5. Neural Networks
6. Boosted Trees
7. Nearest Neighbor

A. Decision Trees

Decision tree [7] builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

B. Naive Bayes Classifier

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors [8]. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of 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. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

C. Logistic Regression

It is a statistical method [9] for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.

D. Linear Support Vector Machine

A Support Vector Machine (SVM) [10] is a classifier generally represented by separating the hyperplane. The algorithm classifies the output from the given set of training data into an hyperplane categorizing the data. This hyperplane is a line which divides the plane into two parts where each class lay on either side of the plane. Linear algebra can be used for transforming the given problem into the hyperplane of linear SVM. Kernel plays a very important role here. For Linear Kernel new input can be predicted using dot product between the input (x) and each support vector (xi),which can be calculated as follows

$$f(x) = B(0) + sum(ai * (x,xi))$$

CLASSIFICATION PERFORMANCE MEASURES

The performance of the above-mentioned classification technique can be calculated by the following metrics [11]:

A. Confusion Matrix:

The Confusion matrix is one of the easiest metrics used for finding the correctness and accuracy of the model. It is used for Classification problem where the output can be of two or more types of classes. The Confusion matrix in itself is not a performance measure as such, but almost all of the performance metrics are based on Confusion Matrix and the numbers inside it. The following terms are associated with the confusion matrix.

1. **True Positives (TP):** True positives are the cases when the actual class of the data point was 1(True) and the predicted is also 1(True)

2. True Negatives (TN): True negatives are the cases when the actual class of the data point was 0(False) and the predicted is also 0(False)

3. False Positives (FP): False positives are the cases when the actual class of the data point was 0(False) and the predicted is 1(True). False is because the model has predicted incorrectly and positive because the class predicted was a positive one. (1)

4. False Negatives (FN): False negatives are the cases when the actual class of the data point was 1(True) and the predicted is 0(False). False is because the model has predicted incorrectly and negative because the class predicted was a negative one. (0)

B. Precision

It is the measure in which the fraction of true positives in contrary to all positive results is calculated.

$$Precision = \frac{frequency\ of\ true\ positives}{frequency\ of\ true\ positives + frequency\ of\ false\ positives}$$

C. Accuracy

The percentage of the test tuples that are properly classified by the classifiers is nothing but the accuracy of the particular algorithm in hand.

$$Accuracy = \frac{frequency\ of\ true\ positives + \ frequency\ of\ true\ negatives}{frequency\ of\ true\ positives + \ false\ negatives + \ false\ positives + true\ negatives}$$

D. Recall

It is the ratio of the number of relevant tuples obtained to the total number of relevant tuples in the data set. It is usually expressed as a percentage.

$$\text{Recall} = \frac{X}{X+Y}(100)$$

X= number of relevant tuples obtained

Y=number of relevant tuples not obtained

E. F1 Score

It is the weighted average of precision and recall. In other words, it conveys the balance between precision and recall.

$$\text{F1 Score} = \frac{2*Precision*Recall}{Precision+Recall}$$

EXPERIMENTAL RESULTS

The table 2 and table 3 depict the confusion matrix and the other classification performance measures of the Support Vector Machine on the hepatitis dataset respectively.

TABLE II. *CONFUSION MATRIX FOR SUPPORT VECTOR MACHINE*

	Predicted = YES	Predicted = NO
Actual = YES	3	7
Actual = NO	2	27

TABLE III. *SUPPORT VECTOR MACHINE*

	Precision	Recall	F1 Score	Support
1.	0.60	0.30	0.40	10
2.	0.79	0.93	0.86	29
Average/Total	0.74	0.77	0.74	39

The table 4 and table 5 depict the confusion matrix and the other classification performance measures of the Naïve Bayes Algorithm on the hepatitis dataset respectively.

TABLE IV. *CONFUSION MATRIX FOR NAÏVE BAYES*

	Predicted = YES	Predicted = NO
Actual = YES	10	0
Actual = NO	12	17

TABLE V. *NAÏVE BAYES*

	Precision	Recall	F1 Score	Support
1.	0.45	1.00	0.62	10
2.	1.00	0.59	0.74	29
Average/Total	0.86	0.69	0.71	39

The table 6 and table 7 depict the confusion matrix and the other classification performance measures of the Decision Trees Algorithm on the hepatitis dataset respectively.

TABLE VI. *CONFUSION MATRIX FOR DECISION TREES*

	Predicted = YES	Predicted = NO
Actual = YES	6	4
Actual = NO	3	26

TABLE VII. *DECISION TREES*

	Precision	Recall	F1 Score	Support
1.	0.67	0.60	0.63	10
2.	0.87	0.90	0.88	29
Average/Total	0.82	0.82	0.82	39

The table 8 and table 9 depict the confusion matrix and the other classification performance measures of the Logistic Regression on the hepatitis dataset respectively.

TABLE VIII. *CONFUSION MATRIX FOR LOGISTIC REGRESSION*

	Predicted = YES	Predicted = NO
Actual = YES	5	5
Actual = NO	0	29

TABLE IX. *LOGISTIC REGRESSION*

	Precision	Recall	F1 Score	Support
1.	1.00	0.50	0.67	10
2.	0.81	1.00	0.90	29
Average/Total	0.89	0.87	0.86	39

RESULT ANALYSIS

The paper deals with the application of four classification algorithms on the acquired data set and then drawing out a comparison of the results to one another and also predicting the outcome whether to live or die from the given data. The results of the four selected data algorithms namely Decision Trees, Naïve Bayes Classifier, Logistic Regression, Linear Support Vector Machine were compared and tabulated. According to the outputs derived with the help of python, implementing SciPy Libraries. In order to calculate the accuracy and find out the performance, a confusion matrix was constructed at first. From that matrix, the true positives, true negatives along with the false positives and false negatives were used to calculate the support, recall, f1-score and precision were calculated by implementing specified modules. Final accuracy was calculated using these parameters. From the results, the conclusion obtained is as follows: the Logistic Regression algorithm gives with optimum accuracy of 87.17% which is closely followed by Decision Tree Algorithm with the optimal accuracy of 82.05%. Following the Decision Tree Algorithm is the Linear Support Vector Machine with an optimal accuracy of 76.92%, and lastly the Naïve Bayes Algorithm which has the optimal accuracy of 76.92%. Finally, these algorithms can help in classifying whether a person lives or dies.

Table 10 displays the accuracies of the various classification algorithms when applied on the Hepatitis dataset.

TABLE X. *ACCURACY TABLE*

Algorithm	SVM	Naïve Bayes	Decision Tree	Logistic Regression
Accuracy (%)	76.92	69.23	82.05	87.17

Figure 1 graphically represents the performance of the classification techniques based on their accuracy measures.

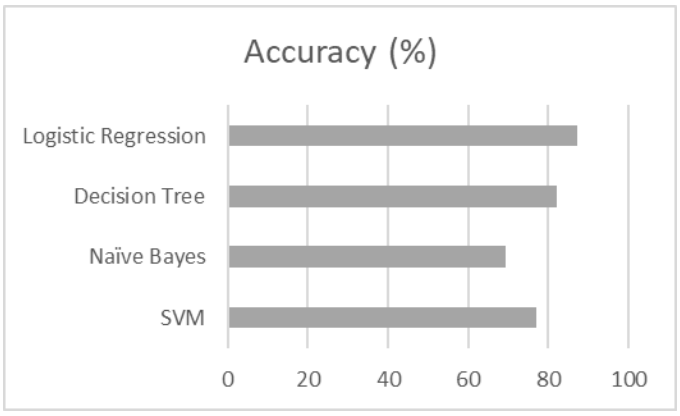

Figure 1. Graphical Representation of Accuracy measure

CONCLUSION

In this paper, popular Machine Learning classification algorithms were chosen to evaluate their performance in the terms of classification performance measures which are accuracy, precision, recall, f1 score and support to classify if the particular person lives or dies based on the various medical tests input as the independent attributes. It has been noticed in the result analysis that the opted algorithms under the classification technique show some well-to-do accuracy percentages, especially the Logistic regression and Decision Tree Algorithm. These algorithms can be preferred over the others to classify the dependent variable LIVE or DIE. However, in further studies the accuracy of all these classification algorithms can be further enhanced by increasing the size of the dataset as well as by using ensemble methods such as bagging and boosting. Furthermore, it is also possible to automate the process of classifying the dataset by creating a simple interface.

ACKNOWLEDGMENT

We would like to thank Dr. K. Thammi Reddy, Professor and Head, Dept. of CSE, GITAM Institute of Technology, GITAM (Deemed to be University) for giving us an opportunity to explore the vast field of Machine Learning through this research.

REFERENCES

[1] Vikas B, Yaswanth D.V.S., Vinay W., Sridhar Reddy B., Saranyu A.V.H. (2017). Classification of Hepatitis C Virus Using Case-Based Reasoning (CBR) with Correlation Lift Metric. *4th International Conference on Information System Design And Intelligent Applications, 2017.*

[2] Christian Brechot .Hepatitis C virus. *Kluwer Academic Publishers-Plenum Publishers, 0163-2116.*

[3] Alex Smola., S.V.N. Vishwanathan. An introduction to machinelearning.

[4] By George D. Magoulas., Andriana Prentza. Machine Learning in Medical Applications. *Lecture Notes in Computer Science, LNCS, volume 2049.*

[5] WHO, Hepatititis C (Fact Sheet No. 164), World Health Organisation, Geneva, 2000.

[6] R.S. Michalski., J.G. Carbonell., T.M. Mitchell. Machine Learning: An Artificial Intelligence Approach.

[7] S. B. Kotsiantis.(2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica 31*, 249-268.

[8] Igor Kononenko. Machine Learning for Medical Diagnnosis: History, State of the Art and Perspective.

[9] Jason Brownlee. How to Implement Logistic Regression with Stochastic Gradient Descent from Scratch with Python - Machine Learning Mastery.

[10] Lee, M., & To, C. (2010). Comparison of Support Vector Machine and Back Propagation Neural Network in

Evaluating the Enterprise Financial Distress. *International Journal Of Artificial Intelligence & Applications*, *1*(3), 31-43.

[11] Vikas B, B.S.Anuhya, K Santosh Bhargav, Sipra Sarangi, Manaswini Chilla. (2017, June). Application of the Apriori Algorithm for Prediction of Polycystic Ovarian Syndrome (PCOS). *4th International Conference on Information System Design And Intelligent Applications, 2017.*