Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients

Maria Alex Kuzhippallil
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, India

Carolyn Joseph
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, India
josephcarolyn7@gmail.com

Kannan A
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, India

Abstract: **Machine Learning has a strong potential in automated diagnosis of various diseases. With the recent upscale in various liver diseases, it is necessary to identify the liver disease at a preliminary stage. In this paper, we propose a new classifier by extending the XGBoost classifier with genetic algorithm. This paper compares various classification models and visualization techniques used to predict liver disease with feature selection. Outlier detection is used to find out the extreme deviating values and they are eliminated using isolation forest. The performance is measured in terms of accuracy, precision, recall f-measure and time complexity. The results of various classifiers are obtained by using proposed feature selection algorithm. From the experiments and comparative analysis, it increases classification accuracy and also leads to reduction in classification time and hence aids in the prediction of the disease more efficiently.**

Keywords – Data mining, liver disease, genetic algorithm, outlier elimination, feature selection, classification models.

I. INTRODUCTION

Machine Learning (ML) a part of Artificial Intelligence (AI) allows the system to obtain knowledge with no explicit knowledge. Supervised algorithms make use of human inputs and outputs for training process and prediction accuracy, and thus used for different classification applications [1].

Therefore, the application of ML has extended to healthcare as well. One of the major problems in healthcare is the rising number of liver disease patients. Liver is vital organ with functionalities like production of bile, detoxification of chemicals and production of important proteins for blood clotting [2]. Long term drinking habits has been directly linked to the increased risk of having different liver diseases which may further lead to death which can be prevented if the disease is detected early. Fatty liver infiltration is the initial stage and Cirrhosis is the final stage in most chronic liver diseases which may further lead to liver cancer. Many data mining techniques or Medical Data Mining (MDM) techniques help in the detection and predict the presence of liver diseases in the early stage itself and reduce work of doctors to some extent [3][5].

The data set that is taken into consideration for the following paper is Indian Patient Data set from UCI [4]. Table 1 explains the liver disease data set that contains 11 attributes with 583 instances.

TABLE 1
Dataset

Index	Attribute	Range
1	Age	4-90
2	Total_bilirubin	0.4-75
3	Direct_Bilirubin	0.1-19.7
4	Total_Proteins	63-2110
5	Albumin	10-2000
6	A/G ratio	10-4929
7	SGPT	2.7-9.6
8	SGOT	0.9-5.5
9	Alkphos	0.3-2.8
10	Gender	Female/Male
11	Selector Field	0/1

II. LITERATURE SURVEY

Data mining techniques have widely been used for the prediction of various diseases. In [6], the authors describe the use of classification algorithms like Decision tree algorithm, Bayes Algorithm and Rule based Algorithm for diabetes disease prediction and they are considered popular classification algorithms at the time. To improve the effectiveness of classification algorithm feature extraction is

978-1-7281-5197-7/20/$31.00 ©2020 IEEE

used.

In [7], various classification algorithms namely Support Vector Machine, K – Nearest Neighbour and Logistic Regression have been used for the prediction of liver disease. On the bases of sensitivity the latter classification algorithm has proved to be more appropriate for the prediction of the disease. Also in [8], the authors explain various unsupervised classification algorithms for the classification of the dataset. The three techniques used in the paper are K-means, DBSCAN and Affinity Propagation. The proposed technique is divided into three stages namely analysis, prediction and comparison. The analysis is being done using K-means, Affinity Propagation and DBSCAN. The data set used contains various levels of enzymes present in the liver system. In order to find the best, performance of the techniques is implemented and is calculated using mainly Silhouette Coefficient. This factor determines accuracy and number of cluster which determines complexity. Finally K-means is found to be the optimal method in comparison to other. The paper further points to future work in determination of other diseases like heart, lung, brain etc.

In [9], the authors implemented genetic algorithm for feature extraction and compared against classification algorithms like Naïve Bayes, Nearest Neighbours and Support Vector Machines. The approach of using feature extraction successfully improved classification and achiever a higher accuracy in classification. Another way to increase the performance and provide better result is the use of hybrid algorithms.

In [10], the authors provide a hybrid algorithm which makes uses of clustering and decision tree induction for the classification where the dataset is first divided into clusters and classified using decision tree. This approach when tested on real life dataset proved to show better result and improved accuracy. Another approach is shown in [11], where the authors propose the use of genetic algorithm along with K-means algorithm for the classification of the dataset where genetic algorithm is used to clean the data to improve the initial cluster centre for the K-means algorithm. And the results proved to be more accurate than just K-means algorithm. Some research paper like in [12], the authors made use of an open source tool called WEKA. WEKA is a tool used to perform data mining work. It is developed at the university in New Zealand. The dataset of sample of 20 patients is collected from BUPA research lab. The dataset has seven attributes which are taken into account. The main criteria set as a parameter is the alcohol consumption. After the use of Bayes theorem into the final result it is found that the alcohol consumption causes more likeliness for liver cancer. In the paper the author establishes the fact that the Bayes model gives the exact prediction. Sometimes the outliers can affect the performance of the algorithms. Therefore it is necessary to detect them and discard if possible.

In [13], the authors explain about outliers detection is essential component of data mining. Usually algorithms used for outlier detection includes depth based, statistical based and cluster based algorithms. This paper clearly highlights recall for various algorithms and the best performance algorithms include Pruning based KNN, KNN and Local Outlier Factor Method.

III. PROPOSED METHODOLGY

The main aim of this research is to propose a method for building predictive model for liver disease using various supervised machine learning algorithms. The comparative analysis of the proposed method has been done and performance is measured using various classification metrics. The brief details of each steps involved for diseased prediction are described as follows –

A. Data Selection

Data selection process involves the need for selecting appropriate data for analysis and obtaining effective knowledge by performing diverse data mining techniques. The data used for research is Indian Liver Disease Patients (ILDP) from UCI repository.

B. Data Exploration

Initial step of data analysis which inculcates summarizing the data and observing initial patterns in the data and attributes is known as Data Exploration. Various visualization techniques such as histogram and boxplot to identify the extreme and outlier values. Feature correlation of values is assessed in order to identify highly linearly dependant features.

C. Data Pre-processing

- *Imputation of Missing Values* - It refers to identifying missing values in the data and imputing the empty values with median values. For Indian Liver Disease Patients data, Albumin and Globulin ratio has four missing values which is replaced by median values. If there is many missing values then imputation can be implemented using KNN imputation.
- *Label Encoding* - Another data pre-processing technique includes label encoding data which focuses on converting the data into machine readable form. Label encoding converts the labels into numeric forms. In the data used, Gender attribute has labelled data which is converted in to values 1 and 0 for better analysis.

- *Elimination of Duplicate Values* – In order to improve the efficiency and quality of data it's very necessary to eliminate redundant values.
- *Resampling* - Due to the presence of imbalanced data which has majority liver disease and minority non liver disease patients, Synthetic minority over sampling technique used. SMOTE is used to synthesize new samples for the minority. This technique involves blindness problem which is overcome by combining Genetic algorithm with SMOTE known as GASMOTE.
- *Outlier Detection and Elimination* – Outliers are extreme values that significantly deviate from the rest of the values which is caused due to inappropriate measurement or experimental error. Different types of outliers include univariate and multivariate outliers. Univariate outliers considers a single feature whereas multivariate outliers looks at n-dimensional space consisting of features or attributes of ILPD data. For univariate outlier detection, skewness of attribute is observed and extreme value is replaced. For multivariate outlier detection, isolation forest algorithm is used to identify the contaminated data and it's deleted.

D. Feature Selection

Feature selection is the process of finding input features for a predictive model which involves removing irrelevant features that don't contribute towards the model.

Genetic algorithm is one of the most advanced method for feature selection and optimization method. Which mimics Darwin method of natural selection. Genetic algorithm follows initialization, fitness assignment, selection, crossover and mutation.

Initialization – Population is initialized with the individuals

Fitness Assignment – Fitness is evaluated by training the predictive model and selection error is found out.

Selection – Selection operator finds out the individuals that will recombine for the next generation.

Crossover – Crossover creates a new population by recombining the individuals.

Mutation – Mutation operation is carried out to create a generation varied diversity compared to the previous generation.

Pseudocode for Feature Selection:

Input: Training Feature Data (X), Target Data (Y), Classifier used XGBoost algorithm (model)

Start

Read input data X, Y, XGBoost model.

Initialising the population size, generation size and cross validation split.

Evaluating fitness of all Features by averaging cross validation score using XGBoost.

While (termination criteria) do:

Fetching subset of features.

Select best individuals for reproduction.

Generate new individuals through crossover and mutation.

Evaluate the fitness of new individuals by averaging cross validation split using XGBoost.

Replacing least fit population with new individuals.

Return best individual

End

Output: Best Features

E. Classification using machine learning algorithms

Classification is performed using various machine learning algorithms which includes –

- *Logistic Regression* – Used specifically when the target variable which is dependant features is categorical data. Different types logistic regression includes binary logistic regression, multinomial logistic regression and ordinal logistic regression. For liver disease prediction the target variables are presence or absence of liver disease which follows binary logistic regression.
- *KNN* – K-Nearest Neighbours is an algorithm that works based on the close proximity of similar data points.
- *Decision Tree* – Tree based learning algorithm which is a good predictive model which produces better accuracy and ease of interpretation. We use categorical variable decision tree that creates splits based on gini method. Higher gini value higher homogeneity.
- *Random Forest Tree* – Random forest tree is a classification algorithm that is made of many decision trees. Each decision tree trains on different observation. Final outcome prediction is obtained by averaging the predictions of individual decision tree.
- *AdaBoost Classifier* – Adaptive Boosting algorithm which majorly focuses on converting weak classifiers into stronger one. The performance of decision trees is boosted by using the instance of previously trained tree. Additive logistic regression concept is applied in a stage wise fitting.
- *XGBoost Classifier* – Optimised gradient boosting algorithm which improves performance by processing in a parallel manner and uses regularization concept to avoid overfitting issue.

- *LightGBM Classifier* – is one of the high performance gradient boosting algorithm which works based on decision tree algorithm which splits the tree leaf wise rather than depth wise.
- *Multilayer Perceptron* – Feedforwards for artificial neural network which is a deep learning method which is essential in classifying data which are not linearly separable. Perceptron consist of input layer and output layer consisting of multiple hidden layers.

F. Performance metrics analysis

Performance of different machine learning models in analysed by using metrics such as –

- *Confusion Metric* – is a table that is used for performance analysis which allows easy visualization. Confusion matrix allows distinguish between true positives, false negatives, false positives and true negative.
- *Accuracy* – This performance measure is calculated by performing ratio of correctly predicted observation to the total number of observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- *Precision* – This is the ratio of relevant instances to total retrieved instances.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

- *Recall* – This is a ratio of relevant instances retrieved over the total amount of relevant instances.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

- *F-measure* – this measure performs weighted average of precision and recall.

$$\text{F} - \text{measure} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

- *Time Complexity* – Time complexity is a measure to identify total time taken for predictions.

IV. RESULTS

The performance of various machine learning algorithms have been evaluated for liver data. The data is obtained from UCI machine learning repository and over sampling is performed using a variant of SMOTE using Genetic algorithm.

Following describes various machine learning models performance analysed by different metrics –

TABLE 2
Before feature selection and Outlier elimination

Algorithm	Accuracy	Precision	Recall	F-measure	Time Complexity
Multilayer Perceptron	0.71	0.50	0.71	0.59	2.243
KNN	0.72	0.69	0.72	0.69	0.0079
Logistic Regression	0.74	0.71	0.74	0.70	0.184
Decision Tree	0.67	0.68	0.67	0.67	0.0079
Random Forest Tree	0.74	0.72	0.74	0.72	0.2503
Gradient Boosting	0.66	0.67	0.66	0.66	0.0568
AdaBoost	0.68	0.64	0.68	0.65	0.390
XGBoost	0.70	0.68	0.70	0.69	0.191
Light GBM	0.70	0.70	0.70	0.70	0.063
Stacking Estimator	0.83	0.83	0.83	0.83	0.388

TABLE 3
After feature selection and outlier elimination

Algorithm	Accuracy	Precision	Recall	F-measure	Time Complexity
Multilayer Perceptron	0.82	0.81	0.82	0.80	0.00099
KNN	0.79	0.77	0.79	0.74	0.0069
Logistic Regression	0.76	0.72	0.76	0.72	0.00099
Decision Tree	0.84	0.84	0.84	0.84	0.00099
Random Forest Tree	0.88	0.88	0.88	0.88	0.0109
Gradient Boosting	0.84	0.84	0.84	0.84	0.0019
AdaBoost	0.83	0.83	0.83	0.83	0.0289
XGBoost	0.86	0.86	0.86	0.86	0.191
Light GBM	0.86	0.85	0.86	0.85	0.0059
Stacking Estimator	0.85	0.85	0.85	0.85	0.364

Table 2 contains the performance metrics of various classification algorithms before applying feature selection and outlier deletion.

Table 3 describes the performance metrics of various classification algorithms after applying feature selection and outlier deletion.

Feature selection is implemented using combination of Genetic Algorithm and XGBoost classifier and outliers are eliminated using isolation forest.

Maximum accuracy is obtained by using LightGBM and Stacking estimator algorithms. Stacking algorithm uses combination of SGD Classifier, Multinomial Naïve Bayes and finally gradient boosting algorithm for prediction.

Due to feature selection and outlier elimination there is considerable improvement of performance in most of the algorithms and considerable decrease in the time taken to train and test.

V. CONCLUSION

In this paper, liver disease prediction has been studied and analysed. The data is cleaned by performing various techniques such as imputation of missing values with median, label encoding to convert categorical into numerical data for easy analysis, duplicate value elimination and outliers are eliminated using Isolation forest in order to improve the performance. Genetic algorithm combined with XGBoost is used to fetch the best attributes required for prediction of liver disease. Different classification algorithms are used to predict the presence or absence of liver disease. Performance metrics such as accuracy, precision, recall, f-measure and time complexity is effectively utilized to analyse the performance of various classification algorithms.

References

[1] R. Saravanan and Pothula Sujatha , "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification" in Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018).

[2] Pushpendra Kumar and Ramjeevan Singh Thakur, " Diagnosis of Liver Disorder Using Fuzzy Adaptive and Neighbor Weighted K-NN Method for LFT Imbalanced Data" in IEEE 6th International Conference on smart structures and systems ICSSS 2019.

[3] Sina Bahramirad, Aida Mustapha and Maryam Eshranghi, "Classification of Liver Disease Diagnosis: A Comparative Study" published in 2013 Second International Conference on Informatics & Applications (ICIA).

[4] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[5] Insha Arshad, Chiranjit Dutta, Tanupriya Choudhary, Abha Thakral, "Liver Disease detection due to excessive alcoholism using Data Mining Techniques", in International Conference on Advances in Computing and Communication Engineering (ICACCE-2018).

[6] Panigrahi Srikanth, Dharmaiah Deverapalli, " A Critical Study of Classification Algorithms Using Diabetes Diagnosis" in IEEE 6th International Conference on Advanced Computing (2016).

[7] Thirunavukkarasu K., Ajay S. Singh, Md Irfan, Abhishek Chowdhury, "Prediction of Liver Disease using Classification Algorithms", 4th International Conference on Computing Communication and Automation (ICCCA)(2018).

[8] Disease Influence Measure Based Diabetic Prediction with Medical Data Set Using Data Mining B.V. Baiju Department of Information Technology Hindustan Institute of Technology and Science.

[9] Noria Bidi, Zakaria Elberrichi, " Feature Selection for Text Classification Using Genetic Algorithms" in 8th International Conference on Modelling, Identification and Control (ICMIC-2016).

[10] Akansha Ahlawat, Bharti Suri, "Improving Classification in Data mining using Hybrid algorithm" in IEEE 2016.

[11]Haobin Shi, Meng Xu, "A Data Classification Method Using Genetic Algorithm and K-means Algorithm with Optimizing Initial Cluster Center" in IEEE International Conference on Computr and Communication Engineering Technology (CCET)(2018).

[12] N.Ramkumar, S. Prakash, S. Ashok Kumar, K Sangeetha, "Prediction of liver cancer using Conditional probability Bayes theorem" published in 2017 International Conference on Computer Communication and Informatics (ICCCI).

[13] Ke Yan, Xiaoming You, Xiaobo Ji, Guangqiag Yin, Fan Yang, "A Hybrid Outlier Detection Method for Health Care Big Data" published in 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom).