HOMOGENOUS ENSEMBLES OF DATA MINING ALGORITHMS IN PREDICTING LIVER DISEASE

Samuel Omokanye, Taye Aro

University of Ilorin-Ilorin, Department of Computer Science

Corresponding author: Samuel Omokanye, oladejiomokanye@yahoo.com

ABSTRACT: Application of data mining algorithms to medical fields have been of interest as it helps patients get access to a better and faster healthcare. In this study, the effect of homogenous ensemble methods of bagging and boosting has been investigated as related to the prediction of the presence or absence of liver diseases. Experimental results show that while bagging and boosting did not improve the accuracy and sensitivity of algorithms in predicting liver disease, Boosting increased the specificity of algorithms.

KEYWORDS: Homogenous, Bagging, Boosting, algorithms, data mining, classification.

1. INTRODUCTION

The liver is a large, complex and important organ in the body, performing a variety of crucial functions in the body which includes producing proteins and enzymes, detoxification, activities related to metabolism, regulation of cholesterol and blood clotting ([M+03]). The liver is located between the portal and the general circulation, between the organs of the gastrointestinal tract and the heart. its main function is to take up nutrients, store them, and make the nutrients available to other organs, but at the same time, the liver can take up potentially damaging substances like bacterial products or drugs delivered by the portal blood or microorganisms, which reach the circulation ([R+08]).

The liver can be diseased due to many factors such as alcohol consumption, nutrition, diabetes, obesity amongst others ([Dha14]) as these things makes potentially damaging substances available to it.

Computer science and medical fields are nested to provide diagnosis of different human diseases. Information given by patients in biomedical diagnosis may include redundant, interrelated symptoms and signs especially when a patient suffers from than one type of disease of the same category. It becomes a serious issue for physicians to diagnose perfectly. Data mining with computational intelligent algorithms can be used to handle prediction in clinical datasets with multiple inputs ([Bin16]). The techniques in data mining have contributed immensely in transforming large data into specific and more relevant information for knowledge discovery and prediction purpose ([Cha14]). Data mining approach in healthcare is a significant component of knowledge discovery in database that is used for the extraction of data associated with several diseases from dataset in order to facilitate easier prognosis of diseases ([AA15]). Applying computational methods in the medical field is well known and gaining ascendency as researches are being geared towards it ([AV15]). Data mining approaches in medical domains is increasing rapidly due to the improvement effectiveness of these approaches for classification and prediction ([L+14]) and a number of researches is on the application of data mining to the diagnosis, prognosis and classification of liver diseases as a way of making the process faster, cheaper and easier ([AV12, JKK14, KK15, M+03, PSR17]).

Ensemble methods have been used to improve the accuracy of data mining algorithms [Zho12] and this study is investigating the effects of two homogenous ensemble methods of bagging and boosting on data mining algorithms which have been considered to perform well in previous works in order to understand how such ensembles contribute to the classification accuracy of liver diseases. The data mining algorithms considered are: C4.5 decision trees, support vector machines (RBF kernel), Random forest, FT trees and logistic regression.

2. REVIEW OF RELATED WORKS

[PSR17] conducted a survey on classifying liver diseases using image processing and data mining techniques. From the different modalities of imaging considered, it is observed that classification of liver diseases is more accurate in computed tomography imaging than the ultrasound imaging. As computed tomography imaging provides a good basis for analyzing the texture of the liver where ultrasound images impose some difficulties in analyzing the liver structure thus making the texture analysis a challenge, though ultrasound imaging is cheaper. The need for a hybrid computer aided diagnostic system that will result in a higher classification accuracy was highlighted.

[SR16] performed a survey of classification techniques in data mining for analyzing liver diseases, the techniques considered are C4.5, Naïve Bayes, Support vector machine, Back propagation neural network and Classification and Regression tree (CART). The algorithms gave various results based on speed, accuracy, performance and cost, and C4.5 was said to give better results in comparison with the other algorithms.

[AV15] studied the relevance of data mining for identifying negatively influenced factors in sick groups, various symptoms of liver disorders in alcoholic patients were analyzed and negative influence factors were identified especially excessive alcoholic consumption.

[KK15] deployed random tree algorithm to classify liver based diseases. The liver disease type being classified into are fatty liver disease, Wilson disease, Inherited disease, autoimmune disease and Cholestatic disease, while the dataset used contains neurological, psychiatric, pathological, physical and cognitive features all of which are categorized into either high, medium, low or medium/low categories. The paper showed that decision trees are used to model actual diagnosis of liver cancer for surgical and non-surgical treatment.

[KB15] did a survey on the trends of data mining in predicting various diseases in the healthcare system. The use of data mining in discovering relationships between health conditions and a disease is presented, how data mining is being used to discover hidden healthcare patterns from related databases, the diagnosis, prognosis, and classification of diseases while focusing on the current techniques used and the future trends are all discussed.

[JKK14] investigated various algorithms which are Naïve Bayes, Decision Tree, Multi-Layer perceptron, KNN, random forest and Logistic on Indian lover patient datasets containing 414 liver patients and 165 non-liver patients, thus the classification was to either identify one with liver disease or one not having liver disease. From experimental results, in terms of precision, Naïve Bayes was preferable while in terms of recall and sensitivity, Logistic and random forest were preferable.

[AV12] used regression analysis to predict the chances of people with ectopic pregnancies to have liver disease by finding the correlation between them and results show that there is an increasing relationship between them.

[RR10] analysed the classification accuracy on liver disorder of three classification algorithms, Naïve Bayes, FT tree and Kstar, and presented results showing that FT tree has the best classification accuracy and closely followed by Naïve Bayes, and in terms of time taken to build classification model, Naïve Bayes took the fastest time.

The algorithms considered in this research were chosen based on the algorithms considered as best in these reviewed literatures and SVM was added based on its wide use as a classification algorithm.

3. METHODOLOGY

Dataset: The dataset to used is gotten from the University of California, Irvine (UCI) repository (http://kdd.ics.uci.edu). The dataset contains 416 persons who have liver disorders and 167 persons who have no liver disease making a total of 583 instances. Based on gender, it contains 441 male patient records and 142 female patient records.

The dataset contains 10 attributes and a selector field which shows whether a person has liver disease or not. The attributes are age, gender, total Bilirubin, direct Bilirubin, Alkpos Alkaline Phosphotase, SGPT Alamine Aminotransferase, total proteins, albumin, A/G ratio Albumin and Globulin Ratio.

Method: Five data mining algorithms are used in this research, they are C4.5 decision trees, Support Vector Machines (SVM), Random forest, Functional trees (FT trees) and logistic regression. The algorithms are applied on the liver diseases data, Ensembles of the algorithms are then implemented, the boosted version of the algorithms using Adaboost applied and then bagging is also done on the algorithms, The individual application of the algorithms were taken as the base results and compared with each other, the boosted version and bagged version are also compared to the base results so as to study their influence on the accuracy of classification and the effect on time taken to build their classification models. All classifications were evaluated using 10-fold cross validation. All implementations were done on using WEKA toolkit, a tool used for data mining related implementations.

Parameters used for evaluation

A. Accuracy: The percentage of correctly classified instances is the accuracy

If TP = True positive
FP = False positive
TN = True negative
FN = False negative

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100\% \quad (1)$$

B. Sensitivity: This is the proportion of people who have the disease and was rightly classified as having the disease. It is also known as recall or true positive rate.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \quad (2)$$

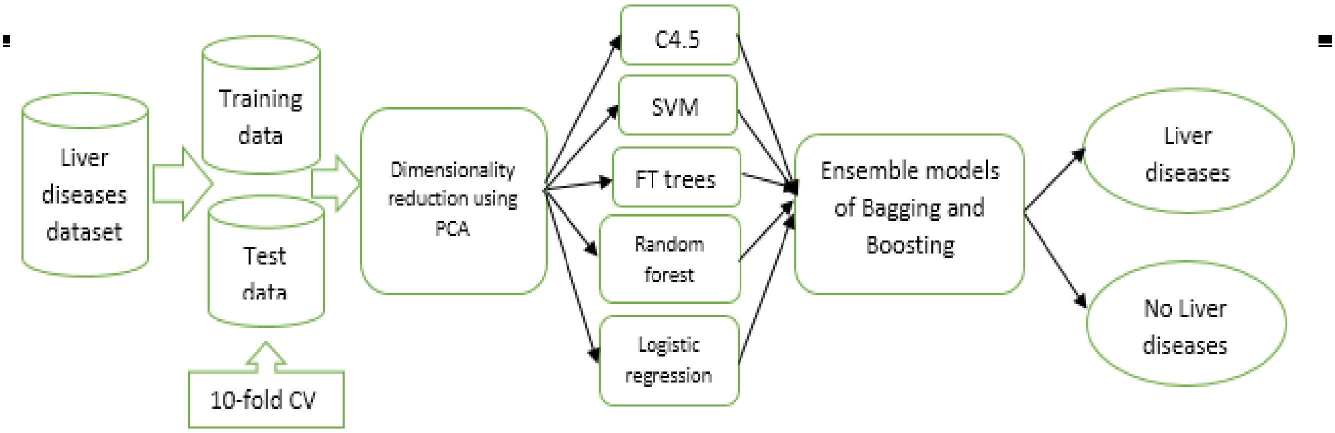

Fig. 1: System Architecture

C. Specificity: This is the proportion of the people who don't have the disease and are rightly classified as not having the disease. It is also known as true negative rate.

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \times 100\% \quad (3)$$

4. RESULTS AND DISCUSSION

Comparing the accuracy of the algorithms, they all performed similarly with FT performing the least and Random forest performing the best, but not statistically significant at a 5% level of statistical significance. The accuracy of all the algorithms were reduced by the ensemble methods of bagging and boosting with the exception of bagged SVM which only increased the accuracy slightly. Boosting the algorithms seem to reduce the accuracy more in comparison with bagging the algorithms, the only exception to this is in Random forest where boosting has a slightly higher accuracy though still lower than random forest without an ensemble.

Support vector machine has the highest sensitivity at a 99% rate, implying that it classified those having the disease correctly almost all the time, followed by C4.5 decision tree. Both bagging and boosting also reduced sensitivity with boosting reducing sensitivity more than bagging. In a situation where we cannot afford to misclassify a person having liver disease as not having it, SVM is the go to algorithm to use for such classification.

In contrast to SVM's sensitivity, it has a very low specificity, meaning that as it is predicting accurately those having liver disease, it is also wrongly predicting a lot of people not having the disease as having the disease. Except for Random Forest where specificity reduced slightly, Bagging and Boosting increased the specificity of algorithms, though generally the specificity is low. Boosted FT tree has the highest specificity.

Table 1: Accuracy of algorithms

Algorithms	Single	Bagged	Boosted
C4.5	70.84	69.30	67.58
SVM	71.01	71.70	68.61
FT	69.13	68.95	67.41
Random Forest	73.07	71.70	72.04
Logistic regression	72.04	71.70	71.36

Table 2: Sensitivity of algorithms

Algorithms	Single	Bagged	Boosted
C4.5	96.63	87.26	87.26
SVM	99.04	95.67	83.89
FT	89.18	88.94	78.37
Random Forest	90.14	89.66	88.94
Logistic regression	91.11	90.62	88.94

Table 3: Specificity of algorithms

Algorithms	Single	Bagged	Boosted
C4.5	6.59	24.55	18.56
SVM	1.19	11.98	30.54
FT	19.16	19.16	40.12
Random Forest	30.54	29.95	29.94
Logistic regression	24.55	24.55	27.54

5. CONCLUSION

Homogenous ensemble methods of bagging and boosting did not cause any significant increase in the degree of accuracy and sensitivity in the prediction of liver diseases, in fact they reduced the accuracy and sensitivity in more cases with boosting causing a higher reduction. In contrast bagging and boosting were discovered to increase specificity in most cases with boosting increasing the specificity more. Thus in cases where cost of wrongly predicting a person who does not have the disease as having such, boosting algorithms is a method to consider.

REFERENCES

[AA15] **O. O. Adeyemo, T. O. Adeyeye** - *Comparative Study of ID3 / C4 . 5 Decision tree and Multilayer Perceptron Algorithms for the Prediction of Typhoid Fever*, African J. Comput. ICT, vol. 8, no. 1, pp. 103–112, 2015.

[AV12] **A. S. Aneeshkumar, C. Venkateswaran** - *An Approach of Data Mining for Predicting the Chances of Liver Disease in Ectopic Pregnant Groups*, in The International Conference on Communication, Computing and Information technology, 2012, pp. 19–22.

[AV15] **A. S. Aneeshkumar, C. J. Venkateswaran** - *Relevance study of Data mining for the identification of negatively influenced factors in sick groups*, Procedia - Procedia Comput. Sci., vol. 47, pp. 101–108, 2015.

[Bin16] **D. C. Bindushree** - *Prediction of Cardiovascular Risk Analysis and Performance Evaluation Using Various Data Mining Technioques: A Review*, Int. J. Enginnering Res., vol. 5013, no. 5, pp. 796–800, 2016.

[Cha14] **D. Chandna** - *Diagnosis of Heart Disease Using Data Mining Algorithm*, Inetrnational Comput. Sci. Inf. Technol., vol. 5, no. 2, pp. 1678–1680, 2014.

[Dha14] **S. Dhamodharan** - *Liver Disease Prediction Using Bayesian Classification*, in 4th National Conference on Advanced computing, applications & Technologies, 2014.

[JKK14] **H. Jin, S. Kim, J. Kim** - *Decision Factors on Effective Liver Patient Data Prediction*, Int. J. Bio-science Bio-Technology, vol. 6, no. 4, pp. 167–177, 2014.

[KB15] **S. Kaur, R. K. Bawa** -*Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System*, Int. J. energy, Inf. Commun., vol. 6, no. 4, pp. 17–34, 2015.

[KK15] **P. Kaur, A. Khamparia** - *Classification of Liver based Diseases Using Random Tree*, Int. J. Adv. Eng. Technol., vol. 8, no. 3, pp. 306–313, 2015.

[L+14] **A. Lebbe, S. Saabith, E. Sundararajan, A. A. Bakar** - *Comparative Study on Different Classification Techniques for Breast Cancer Datset*, Int. J. Comput. Sci. Mob. Comput., vol. 3, no. 10, pp. 185–191, 2014.

[M+03] **L. S. Marsano, C. Mendez, D. Hill, S. Barve, C. J. Mcclain** -*Diagnosis and Treatment of Alcoholic Liver Disease and Its Complications*, Alcohol. Res. Heal., vol. 27, no. 3, pp. 247–256, 2003.

[PSR17] **R. V. Patil, S. S. Sannakki, V. S. Rajpurohit** - *A Survey on Classification of Liver Diseases using Image Processing and Data Mining Techniques*, Int. J. Comput. Sci. Eng., vol. 5, no. 3, pp. 29–34, 2017.

[RR10] **P. Rajeswari, G. S. Reena** - *Analysis of Liver Disorder Using Data mining Algorithm*, Glob. J. Comput. Sci. Technol., vol. 10, no. 14, pp. 48–52, 2010.

[R+08] **G. Ramadori, F. Moriconi, I. Malik, J. Dudas** - *Physiology and pathophysiology of liver inflammation, damage and repair*, J. Physiol. Pharmacol., pp. 107–117, 2008.

[SR16] **D. Sindhuja, R. Priyadarsini** - *A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder*, Int. J. Comput. Sci. Mob. Comput., vol. 5, no. 5, pp. 483–488, 2016.

[Zho12] **Z. - H. Zhou** - *Ensemble Methods: Foundations and algorithms*. Taylor & Francis Group, LLC, 2012.