Diagnosing Liver Disorder by Using Machine Learning Techniques

P.TAMIJE SELVY,
Dept. of CSE, Sri Krishna College of Technology,
Kovaipudur, INDIA

B.GOUTHAM SABARIES
Scholar Dept. of CSE, Sri Krishna College of Technology,
Kovaipudur, INDIA

Abstract — The liver is an important part in the body producing proteins and blood condensing to soaked fat, sugar, and all other digestion activities. Liver has many purposes, counting and removing toxics from the body, and is essential to happen. It has an extensive diversity of purposes which comprises earlier estimate of any sickness is very noteworthy to support human natural life and yield appropriate stages to eradicate the sickness. AI approaches actively busy in foremost categories of therapeutic information. This proposed project work has initial enquiry and forecast of liver disorder by means of various machine learning and deep learning methods and the project model is examined by various performance measures such as recall, precision and accuracy. A tabular analysis is carried out by using some methods such as naive bayes and logistic regression. We have optimised the project model by using ensemble methods such as XG boost and random forest classifiers. At last, we have applied neural networks for getting maximum accuracy.

Keywords — Liver disorder, Therapeutic information, AI, Prediction, Comparison.

1. Introduction

Artificial intelligence (AI) is a booming technology that demonstrates the human intelligence or knowledge in machines that are programmed & tuned to think like humans and perform their actions to the maximum extent. The AI term may also be applied to any machine that produce results associated with a human mind such as learning ability and problem-solving. The main characteristics of artificial intelligence is its ability to observer and take actions that have the best chance of achieving any of the specific goal or tasks. When most people hear the term artificial intelligence, the first thing they usually think of is robots. It's because high budget movies and novels create stories about human-like machines that wreak havoc on Earth. But nothing could be developed from the truth. Artificial intelligence is purely based on the principle that human intelligence can be defined in a path that a machine can easily observe, perform action and execute tasks, from the easiest to those that are even more complex. The goals of artificial intelligence include perception, learning and reasoning. As the technology becomes advanced, earlier proved benchmarks that defined artificial intelligence become outdated.

2. Related Work

Early prediction of liver disease is very important to save human life and take proper steps to control the disease. Decision Tree algorithms have been successfully applied in various fields especially in medical science. This research work explores the early prediction of liver disease using various decision tree techniques. The liver disease dataset which is select for this study is consisting of attributes like total bilirubin, direct bilirubin, age, gender, total proteins, albumin and globulin ratio. The main purpose of this work is to calculate the performance of various decision tree techniques and compare their performance. The decision tree techniques used in the study are J48, LMT, Random Forest, Random tree, REP Tree, Decision Stump, and Hoeffding Tree. The analysis proves that Decision Stump provides the highest accuracy than other techniques.

3. Existing Methodology

With the development of high-through sequencing technology and microbiology, many studies have evidenced that microbes are associated with human diseases, such as obesity, liver cancer and so on. Therefore, identifying the association between microbes and diseases has become an important study topic in current bioinformatics. The emergence of microbe-disease association database has provided an unprecedented opportunity to develop computational method for predicting microbe-disease associations. In the study, we propose a low-rank matrix completion method (called MCHMDA) to predict microbe-disease associations by integrating similarities of microbes and diseases and known microbe-disease associations into a heterogeneous network. The microbe similarity is computed from Gaussian Interaction Profile (GIP) kernel similarity based on the known microbe-disease associations. Then we further improve the microbe similarity by taking into account the inhabiting organs of these microbes in human body. The disease similarity is computed by the average of disease GIP similarity, disease symptom-based similarity and disease functional similarity. Then we construct a heterogeneous microbe-disease association network by integrating the microbe similarity network, disease similarity network and known microbe-disease association network. Finally, a matrix completion method is used to calculate the association scores of unknown microbe-disease pairs by the fast-Singular Value Thresholding (SVT) algorithm. Via 5-fold Cross Validation (5CV) and Leave-One-Out Cross Validation (LOOCV), we evaluate the prediction performance of MCHMDA and other state-of-the-art methods which include BRWMDA, NGRHMDA, LRLSHMDA and KATZHMDA. The experimental results show that MCHMDA outperforms other methods in terms of area under the receiver operating characteristic curve (AUC). MCHMDA achieves the AUC values of 0.9251 and 0.9495 in 5CV and LOOCV, respectively, which are the highest values among the competing methods. In addition, case studies also further prove the prediction ability in practical applications.

4. Proposed Methodology

Feature selection:

Pearson coefficient process is used for feature selection process. Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features. A Pearson's correlation is used when you want to find a linear relationship between two variables. A Pearson correlation is a number between -1 and 1. From the above joint plots and scatterplots, we find direct relationship between the following features: Direct_Bilirubinm and Total_Bilirubin Aspartate_Aminotransferase and Alamine_Aminotransferase, Total_Protiens and Albumin

Albumin_and_Globulin_Ratio and Albumin. Hence, we can very well find that we can omit one of the features. We kept the follwing features: Total_Bilirubin, Alamine_Aminotransferase, Total_Protiens, Albumin_and_Globulin_Ratio, Albumin.

5. Machine Learning Methods

Logistic regression:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logistic regression) is estimating the parameters of a logistic model (a form of binary regression). The dependent variable should be dichotomous in nature (e.g., presence vs. absent). There should be no outliers in the data, which can be assessed by converting the continuous predictors to standardized scores, and removing values below -3.29 or greater than 3.29. There should be no high correlations (multicollinearity) among the predictors. This can be assessed by a correlation matrix among the predictors. Tabachnick and Fidell (2013) suggest that as long correlation coefficients among independent variables are less than 0.90 the assumption is met. At the center of the logistic regression analysis is the task estimating the log odds of an event.

Naïve Bayes:

Naive Bayes (NB) is 'naive' because it makes the assumption that features of a measurement are independent of each other. ... We can simply take each feature separately and determine proportion of previous measurements that belong to class A that have the same value for this feature only. Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes. The Naive Bayes algorithm affords fast, highly scalable model building and scoring. It scales linearly with the number of predictors and rows. P(L|features) =P(L)P(features|L)

P(features)P(L|features)=P(L)P(features|L)P(features)

Here, $(L \mid features)$ is the posterior probability of class.

(L) is the prior probability of class.

$(features|L)$ is the likelihood which is the probability of predictor given class.

$(features)$ is the prior probability of predictor.

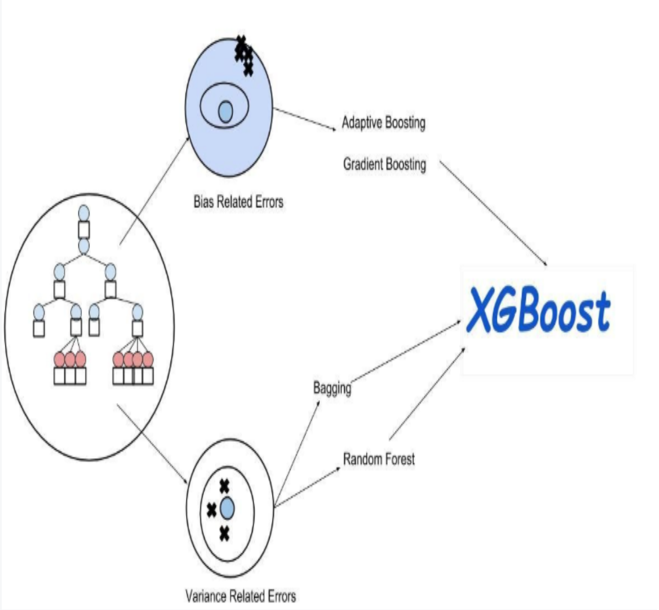
Ensemble methods:

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking). sequential ensemble methods where the base learners are generated sequentially (e.g. AdaBoost). The basic motivation of sequential methods is to exploit the dependence between the base learners. The overall performance can be boosted by weighing previously mislabelled examples with higher weight. parallel ensemble methods where the base learners are generated in parallel (e.g. Random Forest). The basic motivation of parallel methods is to exploit independence between the base learners since the error can be reduced dramatically by averaging.

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working Process Of Random Forest:
XG Boost:
This algorithm goes by lots of different names such as gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines. Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. A popular example is the AdaBoost algorithm that weights data points that are hard to predict. Gradient boosting

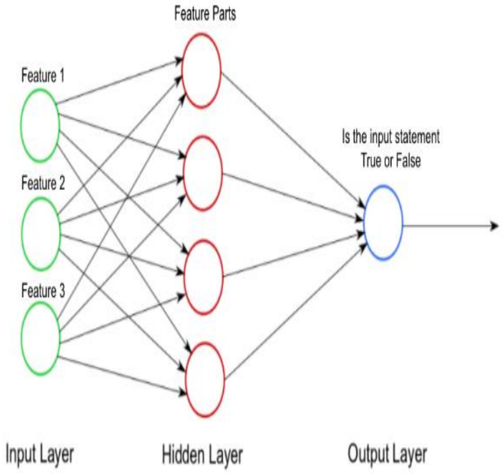
is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. This approach supports both regression and classification predictive modelling problems.



XG Boost Process

Neural networks:

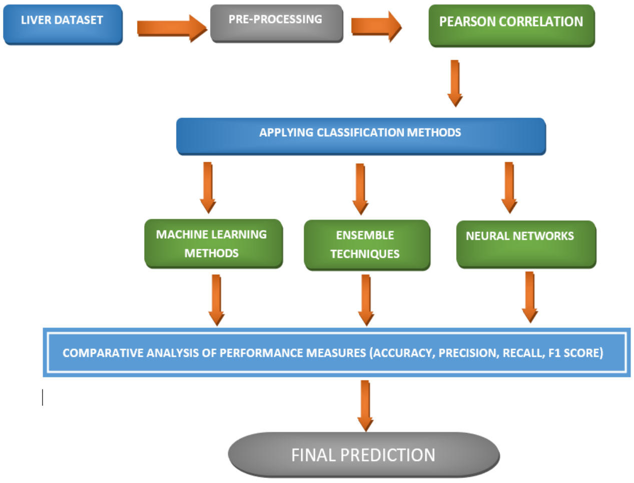
Neural networks are used for solving many business problems such as sales forecasting, customer research, data validation, and risk management. The basic idea behind a neural network is to simulate (copy in a simplified but reasonably faithful way) lots of densely interconnected brain cells inside a computer so you can get it to learn things, recognize patterns, and make decisions in a humanlike way. ANNS have the ability to learn and model non-linear and complex relationships, which is really important because in real-life, many of the relationships between inputs and outputs are non-linear as well as complex.



Process involved in neural network

Proposed work

The figure depicts the proposed model for liver disease prediction. the first phases employ pre-processing stage. The dataset is introduced to person correlation method for attribute selection.



Process involved in neural network

The next stages used classification method such as logistic regression and naive bays method. Then, we applied ensemble methods such as random forest and XG boost. Finally, we applied neural networks.

Results and discussion:

We examined the models designed using various performance measures such as precision, recall, F1-score and accuracy.

First three methods such as logistic regression, naive byes, random forest gives less accuracies of 66% ,52% and 66% respectively. XG boost classifier provides 88% of accuracy. Next, we applied, neural networks and provided maximum accuracy of 100%.

Classifier	**Precision**	**Recall**	**F1**	**OA**
Logistic regression	0.62	0.66	0.64	66%
Naive Bayes	0.79	0.53	0.52	52%
Random Forest	0.65	0.66	0.65	66%
XG Boost	0.81	0.76	0.78	88%
CNN	***0.89***	***0.83***	***0.85***	***91%***

6. Conclusion

Data Analytics is the procedure of retrieve a pattern from large data set in connection with machine learning, data base, and statistics. Machine learning can be a very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases. Our project can assist in proper treatment methods for a patient diagnosed with Liver disorders.

References

[1] Chen, Min, et al. "Disease prediction by machine learning over big data from healthcare communities." Ieee Access 5 (2017): 8869-8879.

[2] A.S.Aneeshkumar and C.J. Venkateswaran, “Estimating the Surveillance of Liver Disorder using Classification Algorithms”, International Journal of Computer Applications (0975 –8887) , Vol. 57,no. 6, (2012), pp. 39-42.

[3] P.Rajeswari and G.S. Reena, “Analysis of Liver Disorder Using data mining Algorithms”, Global Journal of Computer Science and Technology, Vol.10, no. 14 (2010), PP. 48- 52.

[4] B. V. Ramanaland and M.S. P. Babu, “Liver Classification Using Modified Rotation Forest”, International Journal of Engineering Research and Development ISSN: 2278-067X, Vol. 1, no. 6 (2012), PP.17-24.

[5] C.K. Ghosk, F. Islam, E. Ahmed, D.K. Ghosh, A. Haque and Q.K. Islam, “Etiological and clinical patterns of Isolated Hepatomegaly” Journal of Hepato-Gastroenterology, vol.2, no. 1, PP. 1-4.

[6] X. Lu, Q. Xie, Y. Zha, and D. Wang, "Fully automatic liver segmentation combining multi-dimensional graph cut with shape information in 3D CT images," Scientific reports, vol. 8,p. 10700, 2018.

[7] Q. Huang, H. Ding, X. Wang, and G. Wang, "Fully automatic liver segmentation in CT images using modified graph cuts and feature detection," Computers in biology and medicine, vol. 95, pp. 198-208, 2018.

[8] B. Ibragimov and L. Xing, "Segmentation of organs at-risks in head and neck CT images using convolutional neural networks," Medical physics, vol. 44, pp. 547-557, 2017.

[9] S. Nikolov, S. Blackwell, R. Mendes, J. De Fauw, C.Meyer, C. Hughes, et al., "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," arXiv preprint arXiv:1809.04430, 2018.

[10] Y. Guo, Y. Gao, and D. Shen, "Deformable MR prostate segmentation via deep feature learning and sparse patch matching," IEEE transactions on medical imaging, vol. 35, pp. 1077-1089, 2016.