Predicting Human Eye Disease by Ensembled Data Mining Classification Models

Dr. Pankaj Saxena

Asso. Professor, R.B.S. Management Technical Campus, Agra, India

Abstract

The inclusion of Data Mining Algorithms in various fields like that of health care of human has proven to enhance the existing progressions at all times. The Data Mining algorithms automate the manual effort of people in diagnosing the human disease. With the increasing human disease and the need to find proper tools for predicting the human disease is increasing. The data mining algorithms have been providing the patient with a head start to proper treatment and diagnosis. A significant part of the human population suffers from eye diseases. Glaucoma is a human eye disease that causes Irish-eye injury and eventually can lead to full blindness in patients. The detection of Glaucoma at a primary stage can avoid full blindness. Glaucoma is the disease that is focused on in this research paper, by using two data mining classification algorithms to conclude with the better performing classifier while predicting if a person might have Glaucoma by a given set of data along with the deliberations and significance of the two classifiers namely Gradient Boosting classifier and Extra Trees Classifier on the main characteristics Classifiers. The latter section will explain why Gradient Boosting Classifier surpasses Extra Trees Classifier when it comes to predicting glaucoma in a patient. As Accuracy percentage of Gradient Boosting Classifier is 93%.

Keywords: Data Mining Algorithms, Gradient Boosting Classifier, Extra Trees Classifier, Glucoma (Eye Disease),Weka Data Mining tool.

Introduction

Glaucoma is a medical disorder in which the Irish has been injured by blood vessel fluid leaking into the Irish. It is one of the most prevalent disorders of the diabetic eye and a major cause of blindness. Close to 415 million people with diabetes are in danger of blindness. This develops in the light-sensitive tissue at the rear of the eye when diabetes affects the little blood vessels. This mini scale blood vessel leaks blood and fluid in the Irish forms of micro-aneurysms, blood bleeding, hard exudates, spots of cotton wool or vein loops. The stages of DR can be defined depending on the presence of Irish characteristics. At NPDR, the condition may progress from mild, moderate to serious, with different levels of characteristics except less blood vessel formation. PDR is the advanced phase where new blood vessels are triggered by fluids supplied for food by the Irish. They develop along the ocular surface and over the transparent glass gel that fills the eye. A serious loss of vision and perhaps blindness may follow if they leak blood. DI detection is now a time-consuming and tedious method which requires a professional clinician to analyse and evaluate pictures of the Irish digital fundus. When people offer their opinions, frequently a day or two later, the retarded results lead to lost monitoring, misunderstandings and delayed treatment. The research focuses mostly on glaucoma prediction and analysis of several classification algorithms.

When diagnosed beforehand, it will help the patient to avoid long term misery. Data Mining algorithms, in numerous ways, has been aiding the health care sector for years. As the understanding of precautions and diseases increases over time, technology becomes more sought after. Like the method of prediction of diseases from a list of attributes, has been playing a major role with the accurate prediction of diseases like kidney diseases, heart attacks, risk of chronic diseases, cancers etc. Data Mining algorithms in particular, will help come up with classifiers that will be able to predict the status of the disease in a patient in near future with the highest accuracy. It has also been contemplated for machine learning to be a constituent of health-care with organizations insisting to exert efforts further in this sphere to serve and diminish human indulgence by several contributors and researchers in this delicate yet significant part indifferent of the field of medicine. The applications of such algorithms have even assisted people to examine the elucidation of conditions or ailments from the pathological reports and data to understand the necessity of patients and support when medical specialists are unavailable. The point of convergence of the paper is to figure out the better classifier among the two, namely Gradient Boosting and Extra Trees Classifiers in the diagnosis of eye disease i.e Glucoma in an individual. The data-set fed to the classifiers has a various number of attributes that might play a role in contributing to the growth and development of the ailment in a person. The paper even helps with the understanding of the two algorithms and draws a clear conclusion by using figure and numbers in the end.

Literature Review

V. Jackins, S. Vimal, M. Kaliappan and Mi Young Lee (2020) used Artificial Intelligence with classification algorithms like Naïve Bayes and Random Forest classifier to analyse and predict so many human diseases like diabetes, coronary heart disease, breast cancer. They compared both algorithms in their study and found both algorithms are dominant to check the patient is affected by the disease or not affected by the disease.

GowthamBathulaPavan, ChandanaYendluriHari, YeruvaSagar, Varalakshmi M. Sharada, Prasad P. E. S. N. Krishna, Jain Suman, Reddy Kumar Allam Ravi, KondaveetiSaroja and Gunda Padma (2020) used classification methods of data mining. Then they applied support vector machine, K- Nearest Neighbor, Logistic Regression, Decision Tree and Random Forest. From performance of experiments, precision, recall, F1- score, and support were obtained. Random Forest Algorithm showed the highest accuracy and detected Anemia Disease at the early stage.

AlghurairIbrahem Nora and Mezher Dr. Mohammad A. (2020) proposed a very interesting task and compared Support Vector Machine (SVM) classification algorithm and K-means clustering algorithm to predict and diagnose diabetes mellitus. Their experiments had three phases. At the end phase they compared their result to other previous results from their results they obtained the highest accuracy from SVM up to 83% and integrated technology obtained 82%.

Wu Han, Yang Shengqi, Huang Zhangqin, He Jian and Wang Xiaoyi(2018)tried to improve the accuracy of the prediction model and to make an adaptive to more than one data set. This model is based on series of pre-processing procedures, this model is comprised of two parts, the improved K-means algorithm and the logistic regression. Researchers verified the performance of this model from K-fold cross-validation. Researchers found this experiment attained a 3.04% higher accuracy of prediction other than researchers study.

Saxena Pankaj, Singh Vineeta and Lehri Sushma (2013) presented their study on clustering algorithms of Data Mining. In their study they used K-means and K-medoids algorithms of clustering over liver patient data set and proposed an improved K-means medoids clustering algorithm. In their study, they used primary data from Bisariya Pathological Lab, Etah, India and secondary data from UCI repository. They compared their result to previous results and found better accuracy of their results. In their experiment, they found K- medoids more accurate and easily found with less computation of time.

Supervised Machine Learning

Supervised learning is a master learning job that uses supervised training data to determine a function. The supervised learning training data provides a number of instances with matched input topics and desires. A supervised learning algorithm analyses the training information and creates a function known as a classification. For any valid input item, the function should predict the proper output value. The learning algorithm demands that the training data be generated reasonably in unseen condition. Here two of the classification algorithms are used to predict if the person is suffering from eye disease or not. The algorithms are Gradient Boosting Classifier and Extra Tree Classifier.

Gradient Boosting Classifier

Gradient Boosting Classifiers are specific types of algorithms that are used for classification purpose. Features are the inputs that are given to machine learning algorithm, the inputs that are used to calculate an output value. In a mathematical sense, the features of the dataset are the variables used to solve the equation. The other part of the equation is the target or label, which are the classes the instances are categorized into. Because the labels contain the target values for the machine learning classifier, when training a classifier, the data is splitted into training and testing sets. The training set will have labels/targets.

Extra Trees Classifier

Extra Trees Classifier is an ensemble learning method fundamentally based on decision trees. Extra Trees Classifier, like Random Forest, randomizes certain decisions and subsets of data to minimize over-learning from the data and overfitting. The Extra Trees has an optional parameter for selection of subsamples that is bootstrap. In this classification the dataset split is done based on random cut points for each sub trees.

Experimental Analysis

Data Description

The dataset of eye disease i.e glaucoma is collected from UCI Repository. It includes many feature categories which are extracted from the image collection of Messidor. It used to predict if an image has or does not have symptoms of glaucoma. The first 19 attributes of the dataset are separate variables, while the last column is the column for classification. Binary numbers represent outputs. “1” implies glaucoma patients, and “0” signifies the glaucoma is not present.

The study has a total of 1151 data points from different people. The dataset consists of 1151 rows and 20 fields. The study now consists of 920 rows to train data and 231 rows for test data after the division of the data into two classes. The train data was trained for the analysis of different algorithms during performing the investigations.

In this paper, two classification methods used are Gradient Boosting Classifiers and Extra Tree Classifier. The suggested model accept the data of each patient and give us the result put together from the training dataset that is, whether the person has a chance of Glaucoma or not. The training dataset will be brought into play to make the model acquainted with all the attributes and their values. Then the testing dataset will verify the accuracy of the output given by the model using different data value and their actual outcomes. The available data is splited into two parts. The 80% of this dataset is for training and rest 20% of this dataset is for testing. K-Fold Cross Validation technique is used in training the model. The original sample is divided into sub-samples for k-fold cross-validation. During the experiment, the validation used is 10-fold.To settle upon the better classifier in eye disease, some factors are evaluated. The one of the main factors to evaluate the models is Accuracy.

Accuracy = Number of correct predictions/Total number of Predictions

Table:1 Accuracy Score of Classification Models

Classification Algorithm	Accuracy
Extra Tree Classifier	0.85
Gradient Boosting Classifier	0.93

As per the result shown above in Table 1, we can deduce that the accuracy value of Gradient Boosting Classifier and Extra Trees Classifier is almost same. From above result as shown in Table 1, we cannot deduct which model is better as the value of each model is almost same. So further evaluation techniques are required to find the best model.

Table: 2 Average Precision, Precision and Recall of Classification Models

Classification Algorithm	Average Precision	Precision Recall	Recall
Extra Tree Classifier	0.695	0.653	0.596
GrB Classifier	0.654	0.698	0.576

From table 2 we can say the Average Precision value of Extra Tree Classifier is slightly is higher than of Gradient Boosting Classifier. But the difference of the Average Precision of both the classifiers is minimal (0.695 - 0.654 = 0.041) which is approx to 0.04.

Table:3 Representation of F1, Log Loss, ROC AUC of Classification Models

Classification Algorithm	**F1**	**Log loss**	**ROC AUC**	**Build Time (s)**
Extra Tree Classifier	0.562	0.823	0.784	56
Grient Boosting Classifier	0.612	0.653	0.865	49

The **F1-score** is used when we seek a between the balance in value of precision and recall.

Log Loss: It is the most essential factor for classification based on the probabilities.

ROC AUC: It is a curve which shows the performance of the classification problems at different threshold settings.

ROC AUC are two distinguished parts, ROC is the probability curve and AUC is used to measure the separability or it represent the degree of separability.

Build Time: It is the time taken to create the prediction model.

From all the Table.3, it is evident that Gradient Boosting classifier is better than Extra Trees Classifiers . As it clearly seen in Table.3, the Log loss value of Gradient B is less as compared to that of extra tress. And the time consumed by the classifier to generating the model for prediction is less of Gradient Boosting Classifier than that of Extra Trees Classifier.

Conclusion

After observations of all the results and comparing all the factors of the classifiers it is concluded that the Gradient Boosting Classifier is better than Extra Trees Classifier in predicting if a human may have Glaucoma or not. After performing experiments it is observed that the accuracy of the Gradient Boosting Classifier is 93% making it greater than the Extra Trees Classifier's accuracy value. From Table: 3 it is observed that the log loss value of Gradient Boosting Classifier is 0.653 which is comparatively less than that of Extra Trees Classifier i.e 0.823.

References

[1] T. Zheng, W. Xie and L. Xu, L., "A Machine Learning-based Framework to Identify Type 2 Diabetes through Electronic Health Records", International Journal of Medical Informatics, Vol. 97, pp. 120-127, 2017.

[2] K. Plis, R. Bunescu and C. Marling, "A Machine Learning Approach to Predicting Blood Glucose Levels for Diabetes Management", Proceedings of International Conference on Artificial Intelligence, pp. 1-12, 2014.

[3] L. Jena, S. Nayak and R. Swain, "Chronic Disease Risk (CDR) Prediction in Biomedical Data using Machine Learning Approach", Advances in Intelligent Computing and Communication, Vol. 109, pp. 1-13, 2021.

[4] Usma Niyaz et al, "Advances in Deep Learning Techniques for Medical Image Analysis" 5th IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC-2018), 978-1-5386-6026-3/18/$31©2018 IEEE, 20-22 Dec, 2018, Solan, India.

[5] M. S. Al-tarawneh, "Lung Cancer Detection Using Image Processing Techniques," Leonardo Electronic Journal of Practices and Technologies (LEJPT), no. 20, pp. 147–158, 2012.

[6] Abhishek S. Sambyal, Asha T., "Knowledge Abstraction from Textural Features of Brain MRI Images for Diagnosing Brain Tumor using Statistical Techniques and Associative Classification," 2016 International Conference on Systems in Medicine and Biology, IIT Kharagpur, 2016.

[7] Sudhakar, K. and M. Manimekalai, 2014. Study of Heart Disease Prediction using Data Mining, International journal of advanced research in computer science and software engineering, 4(1): 1157-1160.

[8] Pankaj saxena and Sushma lehri, 2013. Analysis of various clustering algorithms of data mining on health informatics, International Journal of Computer & Communication Technology, 4(2): 108-112.

[9] Aqueel Ahmed and Shaikh Abdul Hannan, 2012. Data Mining Techniques to Find Out Heart Diseases: An Overview, International journal of innovative technology and exploring engineering, 1(4): 18-23.

[10] H. Chougrad, H. Zouaki, O. Alheyane “Convolutional Neural Networks for Breast Cancer Screening: Transfer Learning with Exponential Decay,”- arXiv, 2017.

[11] Divya Tomar, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5.5 (2013): 241-266.

[12] D. S. Deulkar and R. R. Deshmukh. "Data Mining Classification."Imperial Journal of Interdisciplinary Research Vol. 2, No.4, 2016.

[13] Parvez Ahmed, Saqib Qamar, and Syed Qasim Afser Rizvi. "Techniques of Data Mining In Healthcare: A Review." International Journal of Computer Applications Vol. 120, No.15, 2015.

[14] Chang, Chun-Lang, and Chih-Hao Chen. "Applying decision tree and neural network to increase quality of dermatologic diagnosis." Expert Systems with Applications , Vol. 36, No. 2 , 2009, pp. 4035-4041.

[15] Kavitha, K. S., K. V. Ramakrishnan, and Manoj Kumar Singh. "Modeling and design of evolutionary neural network for heart disease detection."International Journal of Computer Science Issues Vol.7, No.5, 2010, pp. 272-283.