Prediction of Liver Disease using Regression Tree

https://doi.org/10.3991/ijoe.v17i02.19287

Vinutha M.R. (✉), Chandrika J.
Malnad College of Engineering, Hassan, India
mrv@mcehassan.ac.in

Abstract—Data Mining plays a decisive role especially in medical domain. Decision trees are predominant model in machine learning. Decision trees are simple and very effective classification approach. The decision tree identifies the utmost prime features of a given problem. One of the most common disease in India is Liver Cirrhosis. It is distinctly difficult to uncover Liver Cirrhosis in its initial stage. However early diagnosis of Liver Cirrhosis is highly important. The liver disease data set has a collection of distinguishing features that affect the healthy state of a patient. Machine Learning methods enable knowledge acquisition in early stages and use of this acquired knowledge plays an important role in solving problems like suppose if we want to predict whether the patient with the Liver Cirrhosis has also been suffering from Hepatitis C or not. In order to easily arrive at this knowledge certainly there is a need for fully integrated system. In this paper the collected Liver disease data set is analyzed and prognosticated whether the patient is suffering from liver cirrhosis or not.

Keywords—Decision Tree, Regression Tree, Liver Cirrhosis, Machine Learning

1 Introduction

Machine Learning [1] is an approach to enhance the performance of the machines. This is done by developing efficient algorithms, which makes the system to learn by experience for a given task. Classification is one such method that makes the machines to learn. The well-known procedure for classification is decision trees which has the capability to recognize and split the data into separate classes. Some of the other classification learning techniques are C4.5, ID3, boosted decision trees. The liver of a human is amidst one of the largest organs. It weighs around 1500 grams. At the right side of abdominal cavity just below the diaphragm the liver is located. There are two large veins. One is Hepatic Artery and the other is Portal vein. These two veins are in charge for transporting blood to the liver. Blood from arteria, which is rich in oxygen is provided by artery. Hepatic Artery and Portal vein blood vessels branch within the liver in a continuous manner and ends with extremely tiny capillaries. Every single disorganization with respect to liver may steer to weakness of liver, an associated serious or chronic inflammation and sometimes there are chances of harming other organs inside the body. Cirrhosis is a slowly progressive disease where in healthy liver tissue is replaced with scar tissue. It prevents the liver from functioning accurately. The damaged

tissue blocks the flow of blood through the liver and this slows the processing of nutrients, hormones, drugs and naturally produced nutrients. Liver cirrhosis is the 12th leading cause of death by disease according to the National Institute of Health. One of the main reasons for liver cirrhosis is over consumption of alcohol for a long duration oftime. Hence it is required to predict any illness related to this organ very effectively. So is the development of this proposed work

2 Related Work

Manish Varma Datla et.al. [2] has done the comparison of two machine learning algorithms, Decision tree and Regression tree. After carrying out the study arrived at the conclusion that decision tree works well with a small data sets whereas the regression tree gives the better result for the huge data set.

For the prediction of liver disease researchers have used unsupervised machine learning algorithms [3]. The prediction is grounded onrecital of the implementation of different techniques. The multitude of factors like Adjusted Mutual Information, Homogeneity, Completeness, Vmeasure, Adjusted Rand Index were used for measuring the performance

Varun Vats et.al [4] used three machine learning techniques such as K-Means, DBscan and affinity propagation. These three algorithms were used in order to compare the complexity of computation and accuracy prediction on the liver disease data set. In order to predict the accuracy Silhouette coefficient was used. Out of three techniques K-means was found to be optimal method.

Amodel was proposed by Kanza Hamid et.al [5] which abstain from engender the label of a test example when the prediction is not correct. A novel stochastic gradient descent-based solver has been proposed by the researcher for learning with abstention paradigm and this is been used in order to construct practical up to the minute model for performing classification of liver disease data set.

A model was developed by R. H. Lin et.al [6] that performs the task in two stages. In the First stage the task of identifying the presence of disease is carried out and in the second stage recognizing the type of liver disease is done. CART and CBR techniques were integrated in the proposed intelligent model, which was used for the prognosis of liver disease. CART was used to diagnose whether a patient is suffering from liver disease or not and CBR is used to identify liver disease type.

SinaBahramiradet.al [7] applied eleven data mining classification algorithms on the data set containing four hundred and sixteen liver patient's record and one hundred and sixty-seven non liver patient records. Out of six hundred twenty-seven, four hundred and forty-one male patient records and one hundred and forty-two female patient records were taken. The measures such as Precision, Recall and Accuracy were used to measure the performance.

P. Rajeswariet.al [8] carried out the data classification on the liver disordered data set collected from UCI repository. A total of three hundred forty-five records with seven different attributes were taken. To classify the data WEKA tool was used and tenfold

cross validation was done in order to assess the data. In this paper regression tree learning is used to inspect the collected liver disease data set.

From literature survey it is evident that one should make the good choice of features which play an important role in making the decision whether the patient suffers from liver cirrhosis or not. The paper is collocated as follows, section 3 confers about the classification methods, section 4 briefs feature engineering, section 5 deals with empirical analysis of prediction, section 6 summarizes and discusses future work.

3 Classification Methods

Classification [9][10][11] is an important procedure for machine learning. It has three forms 1. Supervised learning 2. Unsupervised learning and 3. Semi-supervised learning. In supervised learning process the procedure works with the group of examples whose labels are known. The classification learning approach considers categorical values but the regression procedure takes numerical values. In the unsupervised learning method, the class labels are un-known in advance but are grouped into clusters as per their attribute characteristics. Semi-supervised learning utilizes both labeled and unlabeled class data. The classification learning is normally a supervised procedure that takes an example in the data set and identifies to it to a class attribute. An example has two parts the predictor attribute values and target attribute values respectively. The predictor attribute values are used to predict the values of target attribute value. It is also used to predict the class of an example. In the classification learning process the collected dataset is split up into two sets, the training data set and the test data set. The classification process consists of two stages. The model is obtained by using training data set at the training stage. The testing stage uses the model on the test data set to predict the target attribute value. When classifying examples in the test set are unseen during training, the classifier maximizes the predictive accuracy. The knowledge learnt by the classification procedure can be constituted in different manner such as the association rule learning, decision tree learning and artificial neural network learning.

3.1 Decision tree

Decision tree [12][13][14][15][16] is utilized for classification in the decision-making process. It consists of two distinct nodes, the internal node and the leaf node. One of the internal nodes is designated as the root node. The internal nodes are related to attributes, whereas the leaf nodes represent the class name. Every non-leaf node has an outgoing branch. To find the class name for the new record in the data set the search process starts at the root node. The subsequent internal nodes are covered till the leaf node is arrived. To find the right class for a leaf node, testing is done for every internal node from a given root node. Starting from the root node move down by visiting each and every internal node between them and assign the class of the new record same as the class of the leaf node.

3.2 Regression tree

A Regression tree [13] may be observed as a variant of decision tree. It is depicted to approximate real-valued function instead of being used for classification methods. Regression trees used especially for prediction type problems but for the classification types of problems classification trees are used. However, classification tree are used where there is a need for the dataset to be fractionated into classes that belongs to a response variable. The construction of regression tree is carried out with a binary recursive partitioning process. This process is an iterative splitting method. In this each partition is split into smaller groups and this method of splitting keeps on moving up for each branch.

4 Feature Engineering

The collected liver data set is taken for the purpose of studying the classification process. Methods used for Data collection are: (1) By having direct interaction with the patients (2) Recording the outcomes of blood tests and (3) Recording the outcome of the scanning. A total of four hundred and thirty-five records were collected. This collected data set is fractionated into 2 sets, training data set and testing data set. The procedure and the associated feature engineering are performed on the training data set and this results in building a classification model. Then obtained model is applied on the test data set in order to predict whether or not the patients suffering from liver disease. In this study, we have used three measures of performance for the purpose of analysis. They are the Root Mean Squared Error (RMSE), Mean of Squared Error (MSE) and Mean Absolute Error (MAE). RMSE is the square root of the average of squared errors. It is given by equation 1,

$$\text{RMSE}=\sqrt{1/n\sum_{i=1}^{n}(f_i - o_i)^2} \quad (1)$$

Where RMSE is Root Mean Square Error, n is the number of samples, f_i is the i^{th} predicted value and o_i is the i^{th}actual value in the data set.

MSE is stated as the mean of the squares of the actual and predicted values of the instances in the data set. It is given by equation 2,

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i - o_i)^2 \quad (2)$$

Where MSE is the mean square error, n is the number of samples, $y_i$is the i^{th} predicted value and o_i is the i^{th}actual value in the data set.

MAE is defined as the mean of the absolute difference between the actual and predicted values of the records in the data set. It is given by equation 3,

$$\text{MAE}= \frac{1}{n}\sum_{i=1}^{n}|y_i - o_i| \quad (3)$$

Where MAE is the mean absolute error, $y_i$is the i^{th} predicted value and o_i is the i^{th}-actual value in the data set.

The features that are considered for our study are shown below:
lc_age: attribute age of patient expressed in terms of number of years
lc_gen: attribute gender of patient expressed in number as male (1) or female (0)
lc_dalc:attribute duration of alcohol consumptionexpressed in years
lc_qalc:attributequantity of alcohol consumptionexpressed in quarters per day
lc_mcv:attribute mean corpuscular volume is expressed as femtoliters per cell
lc_plcnt:attribute total platelet count expressed in lakhs per mm.
lc_alb: attribute albuminexpressed in gm per dl
lc_tpn: attribute total protein expressed in gm per dl
lc_gln:attribute globulinexpressed in gm per dl
lc_sgotast:attribute SGOT/AST expressed in (U/L)
lc_agratio:attribute albumin/ globulin ratio
lc_sgptatl:attribute SGPT/ALT.
lc_dia:attribute patient suffering from diabetes expressed with values, yes or no.
lc_obe:attribute for patient suffering from obesity expressed with values ,yes or no.
lc_class: attribute for patient suffering from liver cirrhosis or not, expressed as yes or no.

The liver data set containing 435 records is taken, the data set is divided into training data set of 348 records and remaining 87 records into test data that is 80% training data and 20% testing data. For fractionating the dataset into training data set and testing data set we have used the measures mean absolute error, mean squared error and root mean squared error. Observing the Table 1, we find that MAE, MSE and RMSE is high when 70% of data set is taken for training and 30% of data set is taken for testing.MAE, MSE and RMSE is moderate when 60% of data set is taken for training and 40% data set is taken for testing.MAE, MSE and RMSE is comparatively low when 80% of the data set is taken for training and 20% of data is taken for testing. Hence it is appropriate to take 80% of the data set for training the model and remaining 20% is taken as test data set. The regression decision tree is then constructed and used for prediction of class for test data set.

Table 1. Table for type of error with range of training and testing data set

Type of Error	60%training& 40% testing data	70%training & 30%testing data	80% training & 20%testing data
MAE	0.85	0.95	0.83
MSE	0.85	0.90	0.68
RMSE	0.92	0.95	0.83

5 Prediction Analysis

Before the prediction analysis is carried out the liver disease dataset has been pre-processed. The missing data are filled by taking the mean of the attribute. For the collected data set MAE, MSE, RMSE is calculated. Obtained results are recorded in the tables 2, 3, 4 and 5. Now analyzing Table 2, we could observe from the table 2 that

MAE has the value 0.29 for female and 0.54 for male. MSE has the value 0.44 for female and 0.68 for male.MAE and MSE is less when the gender attribute lc_gen is 0 that is for female. Also, RMSE is lower for female that is when lc_gen is 0. This supports the fact that the model prediction with higher accuracy when the gender attribute lc_gen is female.

Table 2. MAE, MSE & RMSE for Male and Female w.r.t tolc_dia

Type of Error	lc_gen=0	lc_gen=1
MAE	0.29	0.54
MSE	0.44	0.68
RMSE	0.25	0.44

Now going through the Table 3 we observe there are equal values for MAE, MSE and RMSE for both male and female. This occurs when the total platelet count attribute lc_plcnt is less than 1.5 for female, and less than 1.25 for male. Hence male and female have equal chances for liver disease for the corresponding values of total platelet count attribute lc_plcnt.

Table 3. MAE, MSE& RMSE for Male and Female w.r.t to lc_plcnt

Type of Error	lc_gen=0&lc_plcnt<1.5	lc_gen=1 &lc_plcnt<1.25
MAE	0.1724	0.1724
MSE	0.029	0.029
RMSE	0.1724	0.1724

Now going through the Table 4 we observe that the MAE and MSE are 0.17 and 0.20 which is lowest value for female. This turns out when the albumin attribute lc_alb is less than 4 and lc_gen is 0. We also observe that RMSE is lower whenlc_gen is 0, which is female. This affirms that the model predicts with higher accuracy when the attribute lc_gen is female.

Table 4. MAE, MSE& RMSE for Male and Female w.r.t tolc_alb<=4

Type of Error	lc_gen=0	lc_gen=1
MAE	0.17	0.20
MSE	0.03	0.06
RMSE	0.17	0.23

We define a measure p, to estimate the accuracy of alcohol consumption for male and female. The measure p is stated as the absolute difference between duration of alcohol consumption and quantity of alcohol consumption. We consider 70% of alcohol consumption by female affects the liver in comparison with male. For a given value of p with 6 for male and 4.2 for female, the value for the performance measures MSE, RMSE of MAE are calculated and tabulated in Table 5.

Table 5. MAE,MSE&RMSE for Male and Female w.r.t to p

Type of Error	lc_gen=1 and p>4.2	lc_gen=0 and p>6
MAE	0.23	0.17
MSE	0.09	0.03
RMSE	0.3	0.17

Now interposing the Table 5 we notice that the MAE and MSE has lowest value that is 0.23 and 0.17for the gender attribute lc_gen is 0. We also observe that RMSE is lower when the attribute lc_gen is 0, which is female. The outcomes of tables 2 to 5 supports the affirmation that the model predicts with higher accuracy when lc_gen is 0. That is women has the higher chances of getting liver cirrhosis.

6 Conclusion and Future Work

In our work, prediction is carried out using regression tree on the liver disease data set. The collected liver cirrhosis data set has the attributes such as gender, obesity, age, quantity and quality of alcohol consumption, platelet count, albumin, globulin etc. The MAE, MSE and RMSE are calculated. It is found that MAE for the male is more than the female for the attributes such as diabetes, albumin, platelet count and duration of alcohol consumption in the data set. From the analyses of regression tree, we find that the prediction model performs better for lc_gen=0 attribute in the data set in terms of MAE and MSE that is for female attribute. The model predicts that female have higher chances of being affected with liver disease than male. By observing the results in section 5 it is clear that female are more prone to liver cirrhosis than male. In our future work we are planning to apply various other machine learning techniques such as Support Vector Machine, Artificial Neural Networks and Genetic algorithms for analyses and also, we are planning to take more number medical attributes such as mean corpuscular volume (mcv), globulin, albumin/globulin ratio (a/g ratio), obesity, that have the direct impact on the liver disease.

7 References

[1] Qinghong Yang, Pengfei Feng, Zhichao Cheng, "Clothing Product Reviews Mining based on Machine Learning ", International Journal of online and Biomedical Engineering, iJOE – Volume 11, Issue 9, 2015, https://doi.org/10.3991/ijoe.v11i9.5069

[2] Manish Varma Datla. “Bench Marking of Classification Algorithms: Decision Trees and Random Forests using R –A Case Study”, International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15), Bangalore, Dec 21-22, 2015, pp.1-7. https://doi.org/10.1109/itact.2015.7492647

[3] J. R. Quinlan, “Learning decision tree classifiers,” ACM Computing Surveys, 28(1), Volume 28, Issue 1, March 1996, NY, USA. pp. 71-72. https://doi.org/10.1145/234313.234346

[4] Varun Vats, Lining Zhang, Sreejit Chatterjee, Sabbir Ahmed, Elvin Enziama and Kemal Tepe Dept of Electrical and Computer Engineering, University of Windsor, Windsor ON N9B 3P4, A Comparative Analysis of Unsupervised Machine Techniques for Liver Disease

Prediction Louisville KY, USA, 6-8 Dec, pp. 486-489. https://doi.org/10.1109/isspit.2018.8642650

[5] Kanza Hamid, Amina Asif, Wajid Arshad Abassi, DureSabih, "Machine Learning with Abstention for Automated Liver Disease Diagnosis, 2017 International Conference on Frontiers of Information Technology, pp.356-361., https://doi.org/10.1109/fit.2017.00070

[6] R. H. Lin, "An Intelligent Model for Liver Disease Diagnosis. Artificial Intelligence in Medicine", 47(1):53-62, 2009.

[7] Sina Bahramirad, Aida Mustapha and Maryam EshraghiEvari: Classification of liver disease diagnosis: A comparative study. 2nd International Conference on Informatics and Applications, ICIA 2013, 10.1109/ICoIA.2013.6650227, 31 October 2013, pp.42-46. https://doi.org/10.1109/icoia.2013.6650227

[8] P. Rajeswari and G. Reena, Analysis of Liver Disorder using Data mining Algorithm. Global Journal of Computer Science and Technology, 10(14):2010, pp.48-52.

[9] Dr. Purushottam, Dr. KanakSaxena, Richa Sharma, "Efficient Heart Disease Prediction System using Decision Tree ", 978-1-4799-8890-7/15/2015 IEEE

[10] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining -concepts and Techniques, Third Edition

[11] T. John Peter, K.Somasundaram, "An Empirical Study on Prediction of Heart Disease using Classification Data Mining Techniques", International Conference on advances in Engineering, Science and Management, ISBN:978-81-909042-2-3/2012IEEE.

[12] HanenBouali, JalelAkaichi, "Comparative Study of Different Classification Techniques", 13th International conference on Machine Learning and Applications", 978- 1-4799-7415-3/14/2014 IEEE.

[13] G. A Melnikov,V.V.Gubarev " Comparative Study of Different Classification Techniques", 13th International conference on Machine Learning and Applications", 978- 1-4799-7415-3/14/2014 IEEE.

[14] M. Deepika and Kalaiselvi, "An Empirical study on disease diagnosis using Data Mining Techniques", second international conference on inventive communication and computational Technologies, 10.1109/ ICICCT2018.8473185/ 2018/IEEE, https://doi.org/10.1109/icicct.2018.8473185

[15] Vyshali J.Gogi and VijayalakshmiM.N, "Prognosis of Liver Disease: Using Machine Learning Algorithms", International conference on Recent Innovations in Electrical Electronics and Communication Engineering ICRIEECE44171.2018.9008482/2018/IEEE. https://doi.org/10.1109/icrieece44171.2018.9008482

[16]Worawut Yimyam, Mahasak Ketcham," The System for Driver Fatigue Monitoring Using Decision Tree via Wireless Sensor Network for Intelligent Transport System", International Journal of online and Biomedical Engineering, iJOE – Vol. 14, No. 10, 2018, https://doi.org/10.3991/ijoe.v14i10.7507

8 Authors

Ms. Vinutha M.R., is working as Assistant Professor at Information Science and Engineering, Malnad College of Engineering, Hassan and has total teaching Experience of 17 years. Her research areas of interest are Data Mining, Soft Computing and Machine Learning. Currently pursuing Ph.D in the field of Data Mining and Machine Learning.

Dr. Chandrika. J, is working as Professor at CS&E, MCE, Hassan since 2015. Received Ph.D from VTU in the year 2014 and has total teaching experience of 28 years.

Her research areas of interest are Data Mining, Big Data Analytics, Soft Computing & Machine Learning. She has more than 45 publications in peer reviewed Journals and national & international conferences. She is also a Member of technical program committee to peer review the papers for IEEE conference on big data and smart city ICBDSC Muscat, Oman in March 2016 Is a peer reviewer of international journals on data mining like IJDKP. Attended as Session chair and TPC member for many national & international conferences held in India. jc@mcehassan.ac.in

Article submitted 2020-10-18. Resubmitted 2020-11-17. Final acceptance 2020-11-18. Final version published as submitted by the authors.