IJIRMET
Issn (online): 2456-0448
International journal of innovative research in management, engineering and technology
vol. 4, issue 2, February 2019

Liver Disorderprognosis With Apache Spark Random Forest And Gradient Booster Algorithms

[1] Thari Krishna, [2] Dr C Rajabhushanam

[1] Dept. Of Computer Science And Engineering, Bharath University, Chennai, India
[2] Dept. Of Computer Science And Engineering, Bharath University, Chennai, India

Abstract: computerbecome an essential component in all the domains including health care. Liver disorder is one of the extremelife-threatening medical conditionthatcompete with cancer and leading death cause in us. More than 10 percent of the american population are affected by liver disorders due to heavy alcohol consumption and unhealthy food habits. Prediction of liver disorders helps in patient diagnosis to increase the survival. In this paper, we analyze the liver disorder datasetgradient boosting and random forest algorithm and compare their performance in terms of accuracy and error.

Keywords—big data,apache spark, machine learning , random forest and gradient booster algorithms.

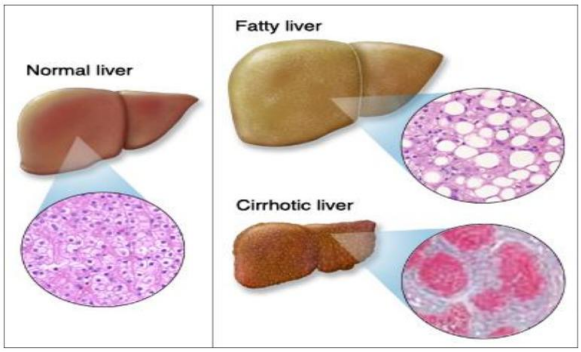
I. INTRODUCTION

Liver is the vital organ which filters the digestive region blood and passed to the entire body. It is in reddish-brown color with the weight of 3 lbs. It resides right side of abdomen and secured by rib cage. It works with gallbladder, pancreas and intestines in digesting, processing and absorbing the nutrition. It produces and store several chemicals required of normal body functioning by cleansing chemicals and drugs. This is one of the leading death cause in us and 10 percent of the american population are affected by liver disorders. Huge consumption of alcohol and unhealthy food habits are the main causes for liver disorder.prediction of liver disorders helps in patient diagnosis to increase the survival time.computers become essential part of all the domains including health care [1]. Using computers lab technicians and pathologists can be able to produce their reports fast and accurate in this paper, we analyze the liver disorder dataset using gradient boosting and random forest machine learning algorithm. And compare their performance in terms of accuracy and error the paper is organized as follows.in section ii we discuss about the of liver disorder overview, types and implications along with schema attributes. In section iii we brief about big data, apache spark concepts. In section iv, we presented the basics of machinelearning, spark machine learning code work flow, random forests algorithm and gradient boosting algorithms. In section v, we present the evaluation parameters of machine learning algorithms. In section vi presented the experiments with results. Finally concluded based on the experiment results.

II. Liver Disorder Overviw

Liver is the major and essential organ in the body thatfights with infections and illness. It cleanses toxic components from the body and make the bodyto function normally. The core parenchymal tissue of the liver is called as hepatocyte. It makes several proteins in the body that needs for different tasks contains blood clotting and preserve liquid within the conduction system. It manufactures cholesterol and triglycerides. It makes glycogen for liver and muscle cell by processing the glucose. It breaks down medicines, drugs, insulin and other hormones in the body. It converts ammonia into urea which is filtered by kidneys as urine. With critical liver failure, patient go to death within days. In chronic liver failure, people health condition is gradually decreased thatleads to death by vomiting blood or taking bloody stools, caused by bleeding from varicose veins in the esophagus and

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 4, Issue 2, February 2019



Stomach. Normal hepatocyte and hepatocyte with liver disorders are as shown in fig 1.

Fig 1. Hepatocyte of normal and disordered livers

A. Types Of Liver Discorders

There are several types of liver disorders [2].some of them include: (a) *liver inflammation*is called hepatitis that is caused by hepatitis a, b, c, d and e viruses. Massive alcohol consumption, drugs, allergic effects and obesity are also caused for hepatitis. Poor sanity habits cause hepatitis a and spread through food handlers. Hepatitis b and c spread through infected body fluids. Hepatitis d uses hepatitis b in spreading and caused liver damage. Hepatitis e is cause by food and water borne toxicity (b) *cirrhosis*is a long-term liver damage that not allows the liver to function normally (c) *hepatocellular carcinoma* is a common type of liver cancer, occurs after cirrhosis (d)*liver failure* occurs due to infection, alcohol consumption and genetic diseases. (e) in *ascites*, liver leaks ascites into stomach and make it weighty and swollen (f) *gallstones* causes hepatitis and cholangitis (g) *hemochromatosis* allows iron to deposit in entire body along with liver, causes various health issues (f) *primary sclerosing cholangitis* is a rear issue causes swelling and damaging the bile ducts in the liver (h) *primary biliary cirrhosis* is also an uncommon disease slowly damages the liver and causes cirrhosis. In next section, we discuss about the implications of each disease.

B. Desease Implecations

Implications of liver disorder [9] include: (a) with nonalcoholic fatty liver disease, people may experience fatigue, pain or weight loss. Over time, inflammation and scarring of the liver (cirrhosis) can occur (b) hepatitis c symptoms remain fever, depression, weight loss, fatigue, nausea and appetite. Skin and eyes of the patient turns into yellow color by swollen blood vessels (c) hepatitis b symptoms include yellowing of eye color, abdominal pain and dark urine. In children these symptoms will not be visible (d) hepatitis a symptoms are fatigue, nausea, abdominal pain, loss of appetite and low-grade fever. Patient gets pain in abdomen, joint and muscles along with diarrhea, nausea, or vomiting (e) cirrhosis of liver [8] symptoms exist fatigue, weakness, reduced hormone production and weight loss. In advanced stages patients may get jaundice, gastrointestinal bleeding, abdominal swelling and confusion. The common symptoms are hemorrhage, breast expansion, bruising, dark urine, inflamed veins around belly button, itching, mental confusion, muscle faintness, speed of breath, inflammation, swelling in extremities, or swollen veins in the lower esophagus (f) alcoholic hepatitis symptoms remain loss of appetite, weight loss, or yellow skin and eyes. Increase in stomach size of patient due to fluid growth (g) hemochromatosis symptoms include joint pain, fatigue, weakness, stomach pain along with diabetes and cirrhosis.

C. Liver Disorder Dataset

Blood tests referred to identify the liver disorderare: (a)*gamma-glutamyltransferase test (ggt):* gamma-glutamyltransferase is an enzyme test that transfers gamma-glutamyl functional groups from molecules including glutathione to amino acid, water or peptide. Ggt is raised by consumption of huge quantities of alcohol. Reference range of normal ggt levels for the individual tests are 15-85 iu/l for men, and 5-55 iu/l for women (b) *aspartate aminotransferase (ast):* this is also called as serum glutamic-oxaloacetic transaminase (sgot). This is an enzyme produced by liver mainly and also by other body organs including heart, kidney, brain and muscles. Usually, ast levels in the blood is low. It will be raised, if the liver hot damaged. High level ast signifies that other organs may be damaged along with liver. (c) *alkaline phosphatase test (alpt):* it measures the alkaline phosphatase enzyme level in the blood sample. The unusual levels of alp refer the problem with liver, gallbladder and bones. This

test also signifies the kidney cancer tumors, pancreas issues, serious infections and malnutrition. Normal range of alp depends on age, gender, blood group and pregnancy status. University of california san francisco (ucsf) is following the normal range for serum alp level as 20–140 iu/l (d)*alanine aminotransferase test (alt):* this is also called as serum glutamic-pyruvic transaminase (sgpt). Increase alt levels in blood represents liver damage. (e) *mean corpuscular volume (mcv):* it depicts red blood cell status. It is also known as mean cell volume of erythrocytes and is calculated by the division of hematocrit with total number of red blood cell count i.e. Erythrocytes. The normal range of mcv is 95fl.an abnormal mcv indicates anemia, liver disease, and etc. Hematocrit count includes hemoglobin concentration with the sum of red blood cell count, white blood cell count and platelets.

The ration of serum levels of ast and alt was described in 1957 by fernando de ritis [13] and it is known as de ritis ratio.in bupa set ast and alt are represented by sgot and sgpt respectively. Most of the authors and researchers identified that this ratio helps in alcoholic hepatitis diagnosis. Normally ast is higher than alt where alt is higher than ast in alcoholic hepatitis cases.

III. Big Data And Apache Spark

A. Big Data

Big datais complex and massive data which cannot be processed by conventional systems [11]. Capture, store, analyze, search, share, transfer, visualize, query, update and privacy of data are the challenges of big data. The main use cases of big data are prediction analysis and behavior analysis. All most all industries producing more data on daily basis which can be treated as big data. Big data has three dimensions:

Volume signifies the massive amount of generated data which includes emails, sensex data, videos and sensor data. It is in the size brontobytes or zettabytes. As per 2008 statistics, facebook got received 10 billion messages with 4.5 billion “like” response and 350 million image uploads per day. The conventional databases are not capable to handle this huge data.

Velocity denotes the data generation and distribution speed. For example, social media messages which passes viral in seconds and milliseconds and financial transactions including credit card transactions.

Variety indicates the different types of data categorized into:(a) structured data which fits into database or table (b) unstructured which cannot be fit into relational data bases includes photos, videos, status updates and etc. (c) semi structured data which is partially structured.

B. Apache Spark

Apache sparkis a wild in memory processing engine to process big data [12]. It contains several modules to process streaming, machine learning, graph processing and sql data.

It handles both sql and nosql databases. It can perform stream processing along with micro batch processing and code can be written in java, python, scala and r languages. Spark core handles input and output functionalities using distributed task dispatcher and job scheduler. Resilient distributed dataset is the base component of spark code. Spark sql provides effective library to convert input data into data frames. Data frame is dataset structured with named column. Spark mllib provides rich library for machine learning algorithms which are applied to data frames. Apache spark pipeline model include below steps:

- Spark read the input data set as sql data frames and convert them to ml datasets.
- Train the data frame with features using estimator.
- Transformer used to convert data frames into model by training i.e. Data frame with predictions.
- Creates machine learning workflow pipeline by chaining estimator and transformer.
- Selects the parameter grid to fit the model.
- Evaluate the model with test data
- Cross validates the model with different parameter map to identify the best fit.

IV. Model And Algorithgms

In this section we brief about the machine learning overview with different types, spark work flow random forest algorithm, gradient booster algorithm with apache spark machine learning algorithm work flow.

A. Machine Learning And Types

Machine learning is an artificial intelligent based application that makes machines learn themselves based on data [3]. Machine learning algorithms are classified into: (a) supervised learning algorithms (b) unsupervised learning algorithms (c) semi supervised learning algorithms (d) reinforcement algorithms.

Supervised learning algorithm helps to predict the output based on pervious data set by modelling the relationships and dependencies between input features and output predictions. These are classified into classification and regression algorithms. Nearest neighbor,naive bayes,decision trees,linear regression,support vector machines (svm),neural networks are the examples for supervised algorithms.

Unsupervised algorithms are trained with unlabeled data and categorized into clustering and association algorithms. K-means clustering and association rules are examples for unsupervised learning.

Semi-supervised learning algorithms are the algorithms fit into both supervised learning and unsupervised learning. These are classified into classification and clustering algorithms.

Reinforcement algorithms are design to attain maximum performance by allowing machines agents to determine the best performance in certain context automatically. They are categorized into classification and control algorithms. Queue learning and temporal difference and deep adversarial networks are the examples for reinforcement algorithms.

B. Random Forests Algorithm

Random forests are the decision tree model that supports both classification and regression. They can handle categorical features by extending multiclass classification setting and able to capture feature relations and non-linarites. They reduce over fitting risk by combining many decision trees. They train the decision trees independently in parallel. The algorithm combines the predictions from each decision tree reduces the prediction variance leads to performance improvement on test data.

The impacting parameters of random forest algorithm include: (a) number of trees in the forest (b) maximum depth of each tree (c) sub sampling rate and (d) number of features to use as candidates for splitting at each tree node.

C. Gradient Boosted Algorithm

Gradient boosting trains the decision trees sequence iteratively. On each iteration, the algorithm uses the current group to predict each training instance label and compares the real label with the prediction. To put more emphasis on training instances with meager predictions the dataset is re-labeled. Thus, in the next iteration it corrects the previous mistake by the decision tree.

The impacting parameters of gradient boosting algorithm include: (a) loss (b) number of iterations (c) learning rate and (d) algorithm i.e. Tree strategy

D. Spark Machine Learning Flow

The flow of apache spark machine learning program is as described in fig2
Step 1: read the input data set
Step 2: eliminate the rows those contain missing values.
Step 3: convert the resultant data into data frames.
Step 4: split the data frames into train and test data sets
Step 5: train the model using logical regression or random forests
Step 6: use the model for prediction
Step 7: calculate the accuracy using predicted data and test data

V. EVALUATION PARAMETERS

Evaluation parameters are defined as follows [4][5] :
Root mean square error: it is the most standard metric that is used for evaluation in regression. It is a reliable metric that prevent positive and negative error cancellation and impacted by outlier values. Rmse metric is given by:

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 4, Issue 2, February 2019

Rmse = $\sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$

Where p is predicted target, a is actual target and n is total number of observations.

Mean absolute error:the mean absolute error (mae) can be compared between models whose errors are measured in the same units. It is slightly smaller than rmse and similar in magnitude.

Mae= $\frac{\sum_{i=1}^{n}|p_i - a_i|}{n}$

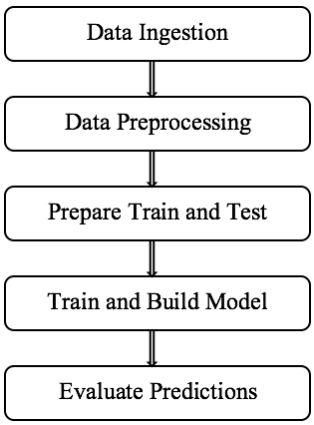
*Coefficient of determination:*the coefficient of determination (r^2) reviews the explanatory power of the regression model and is computed from the sums-of-squares terms.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Sum of squares total (sst) = $\sum(y - \bar{y})^2$

Sum of squares regression (ssr) = $\sum(y' - \bar{y}')^2$

Sum of squares error (sse) = $\sum(y - \bar{y}')^2$




```
sgpt< 21.5
|   gammagt< 20.5
|   |   gammagt< 14.5
|   |   |   alkphos< 70.5
|   |   |   |   sgot< 14.5
|   |   |   |   |   sgpt<9.5: 2 (2/0)
|   |   |   |   |   sgpt>= 9.5: 1 (1/0)
|   |   |   |   sgot>= 14.5: 2 (30/0)
```

Fig 2. Apache spark mechine learning flow

VI. EXPERIMENTS AND RESULTS

We used ubuntu distribution installed on desktop that is running on intel core i5 processor with8gb ram and 1tb hard disk. We installed latest versions of apache hadoop and apache spark for this experiment. We configured the system such that the spark job runs on yarn

We implemented the gradient boosted and random forest algorithm using spark machine learning library.this code is developed on scala language that executes 100 time faster than java. As part of experimental study we found the rmms value of each algorithm and compare them with each other.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 4, Issue 2, February 2019

For experiment, we used the bupa liver disorders dataset from bupa medical research ltd as input data set. This is a multivariate date set contains 7 variables and 345 samples with no missing values. We defined the sixth column as: class 0 was defined as "drinks < 3"and class 1 was defined as "drinks ≥ 3". Seventh column can be split into training set and testing set for classification. The sample tree generated using random forests algorithm is described as below.

The comparison between mse, mae, r2 and rmse calculated using random forest and gradient boosting algorithms are shown in table 1.

COMPARISION OF RANDOM FOREST AND GRADIENT BOOSTERS

Parameter	Random forest	Gradient boosting
Area under roc	0.992912513842746	0.967109634551495
Area under pr	0.986631993583347	0.962227734219895
Mse	0.0499264996059791	0.0290560542546096
Mae	0.06180230114773302	0.07346573904883415
R2	0.872574883218516	0.781047695360312
Rmse	0.170458365164663	0.223442385428501

```
sgpt< 21.5
|   gammagt< 20.5
|   |   gammagt< 14.5
|   |   |   alkphos< 70.5
|   |   |   |   sgot< 14.5
|   |   |   |   |   sgpt<9.5: 2 (2/0)
|   |   |   |   |   sgpt>= 9.5: 1 (1/0)
|   |   |   |   sgot>= 14.5: 2 (30/0)
```

Fig 3. Sample random forest tree

CONCLUSION

In this paper we have analyzed and compared the spark random forests algorithm with spark gradient booster tree algorithm. We discussed about liver disorder overview, types and implications along with schema attributes in section ii. We briefed about big data, apache spark concepts in section iii., we presented the basics of machine learning, spark machine learning code work flow, random forests algorithm and gradient boosting algorithms in section iv. We present the evaluation parameters of machine learning algorithms in section v. We discussed the experiments and resultsin section vi. Finally concluded the paper based on the experiment results.

REFERENCES

[1] hari krishna and dr c. Rajabhushanam,"mininet implementation of sdn towards network softwarization", international journal of innovative research in management, engineering and technology, vol. 2,issue 5, pp.1-4,may 2017,.

[2] Sina bahramirad, aida mustapha and maryam eshraghi, "classification of liver disease diagnosis: a comparative study" ,ieee second international conference oninformatics and applications (icia),2013 .

[3] Seema sharma,jitendra agrawal,shikha agarwal, and sanjeev sharma,"machine learning techniques for data mining: a survey", ieee international conference on computational intelligence and computing research (iccic), 2013.

[4] Archana ganapathi, harumi kuno, umeshwar dayal,janet l. Wiener, armando fox, michael jordan and david patterson,"predicting multiple metrics for queries: better decisions enabled by machine learning", ieee 25th international conference on data engineering, 2009. Icde '09

[5] Marko krstić and milan bjelica, " performance metrics for personalized program guides",ieee 13th symposium on neural networks and applications (neurel), 2016.

[6] Gürol canbek,seref sagiroglu,tugba taskaya temizel, and nazife baykal, "binary classification performance measures/metrics: a comprehensive visualized roadmap to gain new insights", international conference on computer science and engineering (ubmk), 2017.
[7] Yoshihiro tanaka, keitaro oka, takatsugu ono and koji inoue, "accuracy analysis of machine learning-based performance modeling for microprocessors", fourth international japan-egypt conference on electronics, communications and computers (jec-ecc), 2016.
[8] Detlef schuppan and nezam h. Afdhal, "liver cirrhosis", nih public access, 2008,pp. 838–851.
[9] Mark g swain,"fatigue in liver disease: pathophysiology and clinical management", canadian journal of gastroenterology and hepatology,pp.181–188,mar 2006.
[10] Radan bruha, karel dvorak, and jaromir petrtyl, "alcoholic liver disease", world journal of hepatology pp81-90, mar 2012.
[11] Big data wiki available at https://en.wikipedia.org/wiki/big_data
[12] Spark documentation site available at https://spark.apache.org/
Mona botros and kenneth a sikaris,"the de ritis ratio: the test of time", the clinical biochemist reviews, pp.117-130, nov 2013.