ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 7, July 2016

Big data Analytics of Medicare Disabled Patients with Hepatitis Disease

[1] N. Navaneetha, [2] R. Sasi Rekha, [3] L. Prashanti
[1][2][3] Asst. Prof, Dept. of CSE, CMR Engineering College, Hyderabad, India

***Abstract*— In the growing era of technology, concentration is on the analysis of large amount of structured and unstructured data. The processing applications are inadequate to deal with these data are termed as BigData since in large amounts. In this work, an initial stage for analysing medical informatics using Rstudio by R programming is attempted by two algorithms. The biomedical data is used because they are concerned with the real time usage and is an open access journal aiming to facilitate the presentation, validation, use, and re-use of datasets, and can be modifiable with focus on publishing biomedical datasets that can serve as a source for simulation and computational modelling of diseases and biological processes. Random forest technique and support vector machine (SVM) techniques are used to derive features from the database and are able to differentiate various disease supports. The aim of this paper is to provide a comparison between the various techniques that are involved in the field of sorting the data and analysing them in large numbers. For this the process of data mining is used. Data mining is the process of extracting valuable information from a large set of databases. The latter technique produces more appropriate results that has less deviation from the reference taken from the hepatitis profile. By this method one can get the lead vision of the results that are produced by medical science. Therefore the SVM technique can be implemented practically in the medical field.**

Keywords—BigData; Hepatitis; Random forest; SVM; Error analysis .

I. INTRODUCTION

In the growing era of Technology, focus is on the Analysis of large amount of Structured and Unstructured data. The processing applications are inadequate to deal with these data are termed as Big Data , since in large amounts. In this work, an initial stage for analyzing medical informatics by advanced algorithms. The biomedical data is used because they are concerned with the real time usage and we can facilitate the reuse of dataset also data be modified. The focus of biomedical datasets can be a source for simulation and computational modeling of diseases and biological processes. Big data gives the optimal results on analysis of huge amount of database that are considered. In Hepatitis, there are various enzymes that are involved in the secretion of the bile serum .Hence they are sorted using Big data. The Hepatitis is caused due to the inflammation of the liver. Cholylglycine (CG) and Sulfolithocholylglycine (SLCG) in fasting and postprandial serum were determined in patients with liver diseases by radioimmunoassay also in diseases like cancer. These are the two vital parameters that are used to classify to form the decision trees. Methodology of database analysis is the database is generated synthetically with the preferred values of the Hepatitis. After the prediction of factors from the plot ,the techniques are applied. In the hepatitis disease analysis the important part is the plot prediction .It helps to analyze the factors which are important for the result analysis. Here the comparison of the various factors which is taken for the result analysis in the datasets is provided. The main factors are age, CG, and SLCG are main factors which are used for the clinical analysis of liver disease hepatitis. It gives the comparative results of the normal and abnormal rate of the patient. The results are the main factor in the synthetic data is to be compared with the parameters in the datasets. Finally in this analysis the bioinformatics are used for the prediction and computation .Since the current medical field are concerned with the natural data from the patients, the advanced algorithms can be used for classification for any medical data. The factors being obtained from the generalized plot is tabulated and the results are obtained by the advanced algorithms. From these results the deviation can be predicted from the techniques. Since in the medical field a minute deviation can produce a larger impact and consequences are large . By this analysis the techniques proves the complexity and less memory requirements, also optimal design for multiclass, also finding solution for Discrete data.

II. RELATED WORK

The hepatitis C virus (HCV) infection in the past among those born 1945-1965 and HCV's extended latency period of three to five decades, it is expected that progressively more infected individuals currently 50-70 years of age will start exhibiting symptoms of liver damage such as cirrhosis and hepatocellular carcinoma (liver cancer), dramatically increasing the burden on the health care system. Besides liver disease, hepatitis C virus infection (HCV) is also known to cause systemic metabolic and vascular disorders including stroke, cardiac failure, metabolic syndrome, and renal disease,2-5 but antiviral treatment may help in reducing the risk for those extra hepatic outcomes. The Healthy People initiative and the Patient-Centered Outcomes. Previous studies have examined the prevalence of HCV in people with developmental limitations but are of small sample size, outdated, and did not evaluate HCV treatment or management patterns.

Representing almost three quarters of all Medicare patients diagnosed with HCV disabled patients infected with HCV and eligible for Medicare due to the Social Security Disability Benefit (<65 years of age) are an understudied population that faces numerous barriers to the availability, delivery, and quality of care and treatment. Little is known about the management and treatment of the infection among Medicare disabled HCV patients, who due to their high prevalence of psychiatric, substance abuse, and physical comorbidity are likely to be poor candidates for pre-treatment evaluation, treatment receipt, and response to therapy.22 Additionally, an evidence gap exists in our understanding of whether HCV treatment is effective at reducing the incidence of extra hepatic manifestations of HCV in a population facing an elevated risk of metabolic/vascular disorders.

III.ASSUMPTIONS

Explore quality of care (QC) patterns and define differential levels of QC receipt in Medicare HCV disabled patients.

1. Adapt the measurement of previously validated quality of care (QC) metrics to Medicare administrative claims data and construct QC groupings predictive of peg-interferon treatment receipt.

2. In descriptive analyses, contextualize the relationship between patient- and county-level characteristics and the receipt of differential levels of quality.

Determine the factors associated with differential QC receipt and peg interferon Treatment initiation.

1. Examine patient- and county-level determinants of quality of care receipt.

2. Examine patient- and county-level determinants of peg-interferon initiation.
   a) in the whole population
   b) among high, good, fair, and low quality of care recipients.

Assess whether the receipt of peg-interferon therapy for at least 24 weeks is Associated with reduced risk for metabolic and vascular disorders, compared to Untreated patients.

1. In the overall analytic cohort of patients and regardless of underlying comorbidity or race, examine if peg-interferon therapy is associated with slower progression to developing mild, severe, and mild or severe metabolic/vascular disorders.

2. In a healthier and more treatment responsive cohort of White patients free of metabolic/vascular comorbidity at baseline and no diagnosis of diabetes, examine if peg-interferon therapy is associated with slower progression to developing mild, severe, and mild or severe metabolic/vascular disorders.

IV.LITERATURE SURVEY

The hepatitis is caused due to the inflammation of theliver.Cholylglycine (CG) and sulfolithocholylglycine (SLCG) in fasting and postprandial serum were determined in patients with liver diseases by radioimmunoassay also in diseases like cancer, etc. These are the two vital parameters that are used to classify to form the decision trees. In liver disease, serum bile acids [11] were elevated in both the acute and the chronicdisorders. The greatest increase was found in acute viralhepatitis [4] but moderate or slight increase was also found in chronic active hepatitis, liver cirrhosis, and hepatoma and in others. Insignificant elevation of bile acids was found postprandially in patients with liver diseases as well as normal controls and postprandial bile acids were not more sensitive than fasting ones,hence taken in both cases. It has also been suggested that measuring serum bile acids two hours after a meal was the method of choice for detecting hepatobiliary disease [5] rather than conventional liver function testing. Thus the enzyme levels are taken in both the stages since there is a variation in the secretion of the bile serum [12].The values are obtained from the profile of acid range in liver disease from the literature [3], where they had described the estimated levels of bile acids various stages. These values are taken within a set of range and to validate them in BigData by using various techniques which are generated.

V. METHODOLOGY

1. Random forest Algorithm.
2. Support vector Machine.

Introduction to Random Forests

Random forests are an ensemble method used for classification. The methodology includes construction of decision trees of the given training data and matching the test data with these. Random forests are used to rank the importance of variables in a classification problem. To measure the importance of a variable in a data set Dn={(Xi,Yi)}i=1n we fit a random forest to the data. During the fitting process the error for each data point is calculated and averaged over the forest. To measure the importance of the i-th feature after training, the values of the i-th feature are permuted among the training data and the error is again computed on this data set. The importance score for the i-th feature is computed by averaging the difference in error before and the score is done by the standard deviation of these differences. Features which produce large values for this score are more important than features which produce small values. Random forests provide information about the importance of a variable and also the proximity of the data points with one another.

Advantages

1. It provides accurate predictions for many types of applications
2. It can measure the importance of each feature with respect to the training data set.
3. Pairwise proximity between samples can be measured by the training data set.

Disadvantages

1.For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels.
2. If the data contain groups of correlated features of similar relevance for the output, then smaller groups are favored over larger groups.

Support vector machine

Support vector machine(SVM)[5] is the one of the classification algorithms used for the pattern analysis. From the different types Support Vectors are simply the co-ordinates of individual observation. For instance, (45,150) is a support vector which corresponds to a female. Support Vector Machine is a frontier which best segregates the Male from the Females. In this case, the two classes are well separated from each other, hence it is easier to find a SVM. This is another popular algorithm that is used in many real life applications like text categorization, image classification, sentiment analysis and handwritten digit recognition. Support vector machine algorithm can be used both for classification as well as for regression. Spark has the implementation for linear SVM which is a binary classifier. If the datapoints are plotted on a chart the SVM algorithm creates a hyperplane between the datapoints. The algorithm finds the closest points with different labels within the dataset and it plots the hyperplane between those points. The location of the hyperplane is such that it is at maximum distance from these closest points, this way the hyperplane would nicely bifurcate the data.

REFERENCES

1. Gali Halevi & Henk F.Moed,“Evolution of BigData as Research and Scientific Topic: overview of literature”,in Biometrices, September 2012.

2. Kuo lane Chen, Huei Lee, “The Impact of BigData on the Healthcare Information system Transcation of the International Confrence on health information technology advanvement,2013.

3. T R Prajwala, “ A comparative study on Decision Tree and Random forest using R tool”, IJRCCE,vol.4, issue 1, Janauary 2015.

4. Olusegun Adekanle, A.Dennis, Samuel Anu Olowookere Oluwasegun Ijarotimi, and Kayode Thaddeus Ijadunola,“Knowledge of Hepatitis B Virus Infection Immunization with Hepatitis B Vaccine, Risk Perception, and Challenges to Control Hepatitis among Hospital Workers in a Nigerian Tertiary Hospital”,Volume 2015 (2015), Article ID 439867, 6 pages.

5. Sara Romani, Seyed Masoud Hosseini, Seyed Reza Mohebbi Shabnam Kazemian, Shaghayegh Derakhshani, Mahsa Khanyaghma, Pedram Azimzadeh, Afsaneh Sharifian, and Mohammad Reza Zali,

7,“Interleukin-16 Gene Polymorphisms Are Considerable Host Genetic Factors for Patients’ Susceptibility to Chronic Hepatitis B Infection”,Volume 2014 (2014), Article ID 790753.

6. Jason Brwle, “SVM for machine leaning”,machine learning algorithm,april 2016.

7. Segal M. Medicare Part D Utilization and Spending on Hepatitis C Virus (HCV) Medications. Centers for Medicare & Medicaid. Office of Enterprise Data and 2015.

8. Chirikov VV, Shaya FT, Howell CD. Contextual analysis of determinants of late diagnosis of hepatitis C virus infection in medicare patients. Hepatology. Mar 7 2015.

9. Meckley L, Wang Z, Miyasato G, Scaife J, Sanchez H. Medicare Hepatitis C Patients: Are patients under 65 different? Poster Presentation.ISPOR 19th International Meeting, Montreal, Canada, June 2014. 2014;

10. Hong Bo Li, Wei Wang, Hong Wei Ding, Jin Donh, ”Trees Weighting Random Forest Method for Classifying High- Dimensional Noisy Data”, IEEE 7th International Conference, Publication Year: 2010,pages:160-163.

11. Min Ja Kim,Dong Jin Suh, “Profile of Serum Bile Acids in Liver Disease”, korrean J intern med,PMC.

12. M .Mostafa , E. Behairy , M. Azza , A. Sameh , and E.Ehab , “Serum Inter-Alpha-Trypsin Inhibitor Heavy Chain 4 (ITIH4) in Children with Chronic Hepatitis C: Relation to Liver Fibrosis and Viremia”, Volume 2014 .