International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Classification of Metabric Clinical Dataset using Naive Bayes Classifier

E. Jenifer Sweetlin, D. Narain Ponraj

Abstract*: The rapid growth of the internet and its applications makes data grow to huge volumes. The Relational Database Management Systems are inefficient to handle huge volumes of data and so nowadays, Big Data technology is being used by many organizations such as Facebook, Twitter etc. Big Data technology is very useful for organizations to take proper decisions to attain their goals and in mounting themselves organization to full fledge. The use of this technology is broadly widened across all fields of Science, Medicine, Technology, and Business, so it is mandatory to acquire knowledge about Big Data concepts. Thus, acquiring knowledge on the technological revolution from traditional Database Management System to Big Data is significant. In this paper, we have discussed about big data and its evolution, characteristics, data sources, formats, Stages of Big Data process. A huge volume of clinical dataset has been considered and it is analyzed using Naive Bayes Classifier.

***Keywords* : Metabric dataset, big data, naive bayes**

I. INTRODUCTION

Big Data is a popular term where the data is big in size, emerges from variety of sources at a high rate. Various sources of Big Data includes sensors, digital systems, mobile devices, social media platforms, satellites, etc., It is a technology which supports many characteristics related to heterogeneous data coming from many different sources, involves Big Data in many disciplines such as statistics, data mining, machine learning, networking, algorithms and security. The storage as well as data processing tactic differs from conventional methods. It requires advanced software tools, applications, frameworks and libraries to process and manage the data. Using Big Data the storage capability is increased from gigabytes (1000MB) to petabytes (1000TB). It plays an important role in many applications and some of them are healthcare, transport, social media, stock ex-change, geo spatial data, marketing and so on.

The data is being generated and consumed by the people, machine and organization in massive amount. The generated data need to be processed and analyzed efficiently by means of analytical tools and computational techniques.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
E. Jenifer Sweetlin*, Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India. Email: jsweetlin@gmail.com
D. Narain Ponraj, Department of ECE, Karunya Institute of Technology and Sciences, Coimbatore, India. Email: narainpons@gmail.com

Many tools and techniques are available these days. But knowledge about the tools and techniques for their usage and the environments on which they have been designed to work in is required. Only then, the right tool can be selected for a specific application.

II. ABOUT BIGDATA ANALYTICS

Big Data Analytics is the process of analyzing huge amounts of varied datasets (Big Data) to obtaining insights, identification of hidden patterns, unknown correlations, extracting the meaningful and appropriate information to take effective business decisions makes the business strategy and operation successful.

A. *Stages of Analytics*

The different stages of analytics where it helps to find insights from the data to take effective business decisions to improve the business plan based on the pre-diction is shown in Fig.1.

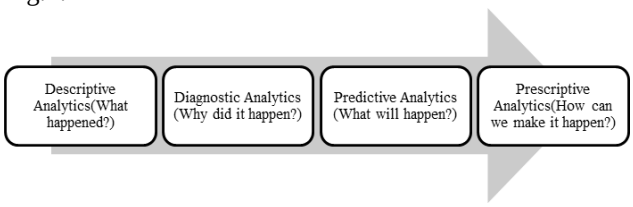


Fig.1 Stages of Analytics

- Descriptive Analytics:

In descriptive analytics the information present in the data is obtained and summarized. It is primarily involved in finding all the statistics that describes the data.

- Diagnostic/Discovery Analytics:

This discovery stage involves in finding the reasons for the statistics deter-mined in the descriptive stage.

- Predictive Analytics:

This stage involves predicting the possible future events based on the information obtained from the descriptive and or discovery analytics stages. Here in this stage the possible risks can be identified and it mainly helps to attain insights for future.

- Prescriptive Analytics:

Based on the prediction data obtained from the predictive analytics stage it involves planning actions or making decisions to improve the business [1].

B. *Uses of Big Data Analytics*

The following are some of the uses of analytics being applied to Big Data and it plays a dynamic role in the industries.

- Personalized Marketing:

Many shopping companies uses Big Data Analytics for personalized marketing to make their customers happy.

These companies are using recommendation engines to improve their product sales. Examples such as Amazon, Flipkart, EBay, etc.,[2].

- Mobile Advertising:

The Big Data analytic engine of a shopping company knows the personalized needs of its customers from shopping history. When offers come up on the products of their interest in a particular place where the customer is around, they are informed over their mobile phones. The Big Data source associated with customer's geographical position is used here [3].

- Retail Industry:

Big Data analytics improves the functioning of any associated organization or business. Data Analytics can improve the following aspects in the areas of procurement, product development, manufacturing, distribution, marketing and price management [4].

- Financial Services:

The data analytics improves the financial services in the following aspects by means of the following aspects. Credit scoring uses to find people with highest credit worthiness. Fraud detection uses to predict fraudulent transactions and customers in the financial services and it helps to formulate strategies and to prevent damage from it [5].

- Transportation:

Big Data analytics greatly improves transportation services. The sensors in the handheld devices, on the roads and on vehicles helps to detect real time traffic in-formation in the road. The data obtained from the sensors are helpful for the traffic authorities to enhance their work in all possible ways to change the routes in the current time or in near future [6].

III. BIGDATA PROCESS

The characteristics of Big Data gives the basic understanding of what makes Big Data a challenge for organizations. However Big Data is not only about the data, the way it is handled is of prime concern. The process of Big Data is presented in detail as follows:

A. *Acquire*

The data coming from multiple data sources like enterprise data, social media, Internet of Things (IoT), sensors, mobile devices, geo spatial data with different formats and the data from the Cloud are need to be obtained for storing, processing and analyzing[7].

B. *Storage*

Data attained from multiple sources with different formats is to be processed and analyzed to obtain insights. Before storing, processing and analyzing the data, it's vital to do data cleaning. Data cleaning includes steps to filter, cleanse and prepare the data for further analysis. From a storage perspective, the copy of the data is first stored in its acquired format, then the data is cleaned and further the prepared data is need to be stored.

The highly scalable and cost-effective technologies and solutions are developed for storing the data. One such storage in Big Data technology is Hadoop Distributed File System (HDFS) which is used for the data storage across multiple nodes in the cluster [8]. So it is important to know some of the aspects and the underlying mechanism behind storage technologies as far as Big Data is concerned.

C. *Processing*

Data acquired should be processed before analyzing it. So during data processing the large dataset should be partition into small datasets to achieve high performance and to get speedy results. In Big data, files are stored in distributed file systems or distributed database so the larger datasets are already partitioned and stored. In traditional relational databases the processing is done in a centralized manner but in Big Data, processing is done in parallel in a distributed location where the data is stored. MapReduce is used to process the data. The data processing may be done either in online or in offline mode.

D. *Accessing*

The data accessing such as querying, searching and indexing the data from the database can be performed through many Big Data tools. One such Hadoop Ecosystem is MongoDB which is a NoSQL database, open source and good to manage the unstructured and semi structured data [9].

E. *Analyzing*

The processed data needs to be analyzed by applying some analytical techniques such as machine learning algorithms, data mining algorithms, statistical methods, etc.,

F. *Visualization*

After analysis is done, the data should be visualized by means of charts, histograms, etc., to presenting the results. R is a programming language which provides statistical and graphical techniques for the data representation that can be shown effectively [10].

IV. RESULTS

The clinical breast cancer dataset can be given as input to Naive Bayes classifiers. Various other classifiers like random forest, decision tree, SVM classifiers can also be used to classify the clinical dataset. However, for our analysis we have considered the Naive Bayes classifier. The dataset we have used contains the information of 2504 breast cancer patients with 35 features. The information of some of the patients has been missed so we have omitted the missed features from the dataset, the number of patients has been reduced from 2504 to 1385. By reducing many features, we had taken only 7 features and 1385 patients which are relevant to breast cancer treatments. We evaluated the performance of two-class classification problem.

We have considered the breast cancer patients survival based on the age at diagnosis, tumor stage and the treatments the patients undergone. In this paper we have grouped the tumor stages of the patients before applying classification algorithms.

The dataset was classified using Naive Bayes Classifier. The false positive rate and true positive rate obtained using the classifier is analysed.

Table 1 represents the graph between FPR vs TPR for the Naive Bayes classifier.

Table 1: FPR vs TPR

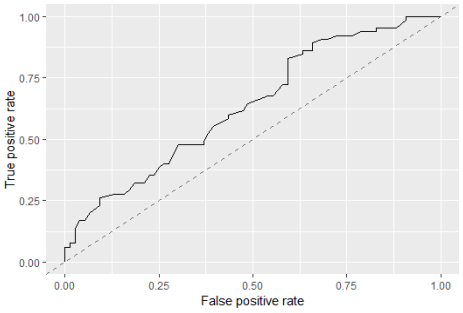
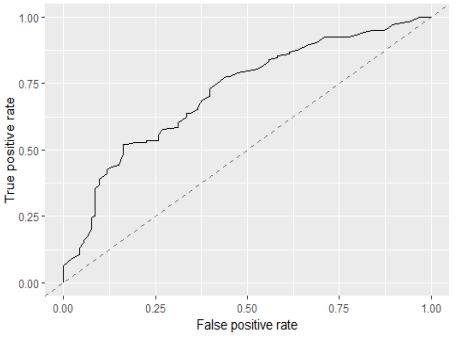
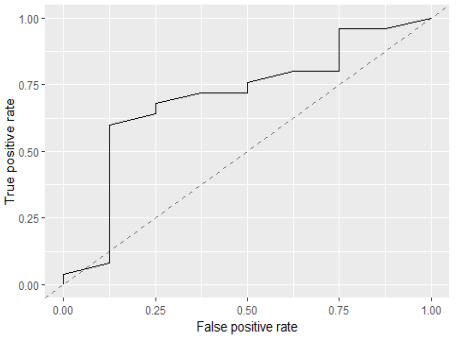
Tumor Stage	Naive Bayes
1	
2	
3	

Table 2 represents the confusion matrix for different tumor stages. The naive bayes classifier performs well at greater tumor stages. It has obtained an accuracy of 58% for first stage tumor and 70% accuracy for stage three.

Table 2: Confusion Matrix

Tumor Stage	Confusion Matrix
1	predicted true DECEASED LIVING DECEASED 34 31 tpr: 0.52 fnr: 0.48 LIVING 28 48 fpr: 0.37 tnr: 0.63 ppv: 0.55 for: 0.39 lrp: 1.42 acc: 0.58 fdr: 0.45 npv: 0.61 lrm: 0.76 dor: 1.88
2	predicted true DECEASED LIVING DECEASED 84 60 tpr: 0.58 fnr: 0.42 LIVING 29 64 fpr: 0.31 tnr: 0.69 ppv: 0.74 for: 0.48 lrp: 1.87 acc: 0.62 fdr: 0.26 npv: 0.52 lrm: 0.61 dor: 3.09
3	predicted true DECEASED LIVING DECEASED 17 8 tpr: 0.68 fnr: 0.32 LIVING 2 6 fpr: 0.25 tnr: 0.75 ppv: 0.89 for: 0.57 lrp: 2.72 acc: 0.7 fdr: 0.11 npv: 0.43 lrm: 0.43 dor: 6.38

V. CONCLUSION

From the analysis we have done, we have observed that the Naive Bayes classifier performs better even at higher tumor stages. Even for the missing values, Naive Bayes Classifier performs better due to its ability in handling so many variables. The features showed significant characteristics difference between the various classes thus increasing the accuracy. We conclude that the Naive Bayes classifier performs better with multiple feature variable in all the test cases. This work can be further extended by considering a data set with many values and to adapt deep learning framework for classification.

REFERENCES

1. Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li., "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial", IEEE Access, 2014, 2,pp.652–687, doi:10.1109/access.2014.2332453.
2. Ducange, P., Pecori, R., and Mezzina, P.,"A glimpse on big data analytics in the framework of marketing strategies", Soft Computing, 2017, 22(1), pp.325–342, doi:10.1007/s00500-017-2536-4.
3. Ahmed, E., Yaqoob, I., Hashem, I. A. T., Shuja, J., Imran, M., Guizani, N., and Bakhsh, S. T, "Recent Advances and Challenges in Mobile Big Data", IEEE Communications Magazine, 2018, 56(2), pp.102–108,
4. Aloysius, J. A., Hoehle, H., Goodarzi, S.,and Venkatesh, V., "Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes", Annals of Operations Research, 2016, doi:10.1007/s10479-016-2276-3.
5. Duan, L., and Xiong, Y., "Big data analytics and business analytics. Journal of Management Analytics", 2015, 2(1),pp.1–21, doi:10.1080/23270012.2015.1020891.
6. Shtern, M., Mian, R., Litoiu, M., Zareian, S., Abdelgawad, H., and Tizghadam, A., "Towards a Multicluster Analytical Engine for Transportation Data", International Conference on Cloud and Autonomic Computing, 2014, doi:10.1109/iccac.2014.37.
7. Gueidi, A., Gharsellaoui, H., and Ahmed, S. B., "A NoSQL based Approach for Real Time Managing of Embedded Data Bases", World Symposium on Computer Applications and Research (WSCAR),2016, doi:10.1109/wscar.2016.17.
8. Okman, L., GalOz, N., Gonen, Y., Gudes, E., and Abramov, J.," Security Issues in NoSQL Databases", IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications, 2011, doi:10.1109/trustcom.2011.70.
9. Samadi, Y., Zbakh, M., and Tadonki, C, "Comparative study between Hadoop and Spark based on Hibench benchmarks", 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech), 2016.
10. Tang, Shanjiang, Bingsheng He, Ce Yu, Yusen Li, and Kun Li, "A Survey on Spark Ecosystem for Big Data Processing", 2018.

AUTHORS PROFILE

Ms. E. Jenifer Sweetlin is currently pursuing her doctoral studies in CITE at Manonmaniam Sundaranar University, Tirunelveli. She has completed her M.Tech in Information Technology from Satyabhama University in the year 2010. Her area of interest includes Big data analytics and Data Science.

Dr. D. Narain Ponraj is Assistant Professor in the department of Electronics and Communication Engineering in Karunya Institute of Technology and Sciences, India. He received his doctoral degree from the same institute in 2017. His research area is medical imaging. He has authored 1 book chapter and several papers in reputed journals. He is a member of the International association of computer sciences and information technology (MIACSIT) and a member of the International association of engineering (MIAENG).