Predicting Breast Cancer using Apache Spark Machine Learning Logistic Regression

S.Sujithra[1] Dr.L.M.Nithya[2] Dr.J.Shanthini[3]
[1]P.G. Student [2]Head of Dept. [3]Associate Professor
[1,2,3]Department of Information Technology
[1,2,3]SNS College of Technology, Coimbatore, India

Abstract— *In real world Breast Cancer Diagnosis and Prognosis are two medical applications pose a great challenge to the researchers. There are many scientific technologies that has rich information in taking medical decisions but that might not be accurate and properly used to its potential. The use of machine learning and data mining techniques has revolutionized the whole process of breast cancer Diagnosis and Prognosis. The objective of these predictions is to assign patients to either as "benign" group that is noncancerous or a "malignant" group that is cancerous. In the existing system it is been summarized with different types of data mining algorithms in order to obtain good mortality rate Strong and sophisticated algorithms like Bagging Logistic Regression ,Support Vector Machine, k-Nearest Neighbors algorithm, Decision tree and Artificial Neural Networks have been used and concluded that there was not a single best algorithm depending on the features of the large dataset, it was also a challenge for single-node tools with limited memory and computing power. In proposed system the main goal is to identify the sample observation as malignant or not. And it visualize the data by their exact attributes through logistic regression analysis which improves performance through intelligent optimizations and also to achieve single node analysis through spark framework. Spark not only raises processing speed and real-time performance but also achieves high fault tolerance and high scalability based on in-memory computing.*

Keywords— Big Data, Hadoop Framework, Cancer Prediction, Map Reduce

I. INTRODUCTION

Cancer is a potentially fatal disease caused mainly by environmental factors that mutate genes encoding critical cell-regulatory proteins. The resultant aberrant cell behavior leads to expansive masses of abnormal cells that destroy surrounding normal tissue and can spread to vital organs resulting in disseminated disease, commonly a harbinger of imminent patient death. More significantly, globalization of unhealthy lifestyles, particularly cigarette smoking and the adoption of many features of the modern Western diet will increase cancer incidence.

Data mining technique involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods in early detection of cancer. In classification learning, the learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying unseen examples. In association learning, any association among features is sought, not just ones that predict a particular class value. In clustering, groups of examples that belong together are sought. In numeric prediction, the outcome to be predicted is not a discrete class but a numeric quantity. In this study, to classify the data and to mine frequent patterns in data set Decision Tree algorithm is used. Data Mining techniques are implemented together to create a novel method to diagnose the existence of cancer for a particular patient. When beginning to work on a data mining problem, it is first necessary to bring all the data together into a set of instances. Integrating data from different sources usually presents many challenges. The data must be assembled, integrated, and cleaned up. Then only it can be used for processing through machine learning techniques. This developed system can be used by physicians and patients alike to easily know a person's cancer status and severity without screening them for testing cancer.

Cancer is a multifaceted disease which involves many stages of development over a long period of time. The root cause of breast cancer is difficult to pin down, however a closer look at the science involved in it can help us to understand why exposure to some harmful radiations and chemicals can be a potential source to this risk.

This paper is organized as follows: in section II, we discuss about related works and the characteristics of Big Data, section III focuses on the Hadoop for Big Data and proposed system implementations, and discuses about the Data Set considered for the Experiment, section VI shows the experimental results and discussion as part of big data analytics and section VII concludes the work.

II. RELATED WORKS

Early detection of cancer is essential for a rapid response and better chances of cure. Unfortunately, early detection of cancer is often difficult because the symptoms of the disease at the beginning are absent. In the existing system it is been summarized with different types of data mining algorithms in order to obtain good mortality rate Strong and sophisticated algorithms like bagging logistic support vector machine, logistic regression, KNN, decision tree and artificial neural networks have been used and concluded that there was not a single best algorithm depending on the features of the dataset Provided that the dataset is large enough, it was also a challenge for single-node tools with limited memory and computing power.

Association Rule Mining Based Predicting Breast Cancer Recurrence on SEER Breast Cancer Data [1] to anticipate regardless of whether breast cancer will recur for the breast cancer patient in view of SEER (Surveillance, Epidemiology, and End Results) dataset of Program of the National Cancer Institute (NCI). Evolutionary Conformal Prediction for Breast Cancer Diagnosis [2] we introduce rule-based Genetic Algorithms (GAs) as a method for building a CP, and we apply the resulting algorithm to the problem of breast cancer diagnosis. Support vector

machines combined with feature selection for breast cancer diagnosis [3] A medical decision making system based on SVM combined with feature selection has been applied on the task of diagnosing breast cancer. A Novel Gene Selection Algorithm for cancer identification based on Random Forest and Particle Swarm Optimization [4] method utilizes K –means clustering technique to grouped similar genes from microarray cancer datasets in to the same clusters. Then applying Random Forest ranking and select top scored genes from each cluster to obtain filtered genes. Identification of Gene Signatures for Classifying of Breast Cancer Subtypes Using Protein Interaction Database and Support Vector Machines [5] the proteins corresponding the selected genes in the first phase were identified in the PPI network. Then, all the selected genes corresponding with the proteins having direct relationship with those identified in the fust step were added to the set of genes selected in the fust phase. This will pool the genes into a bigger set of data. Identifying Informative Genes for Prediction of Breast Cancer Subtypes [6] a tree-based approach that conducts gene selection and builds the classifier simultaneously. We conducted computational experiments to select the minimal number of genes that would reliably predict a given subtype. Predicting Breast Cancer Survivability Using Data Mining Techniques [7] to predict the survivability rate of breast cancer patients. In our study, we have used the SEER data and have introduced a pre-classification approach that take into account three variables: Survival Time Recode (STR), Vital Status Recode (VSR), and Cause of Death (COD).

III. PROPOSED WORK

The main aim of this research project is predicting the cancer and for future prevention of the disease in risk level of a patient. Big data is processing huge volume of data in real time situation.

The objective is to identify the accurate cancer stage and to visualize the data by their exact attributes through logistic regression analysis which improves performance through intelligent optimizations. The cancer prediction system will help medical professionals to reach into conclusions based on the clinical data of patients and also perform with high levels of accuracy. This system supersedes existing prediction systems with its ability to handle big data in cloud and produce results of high accuracy levels.

A. Cancer Observation Schema

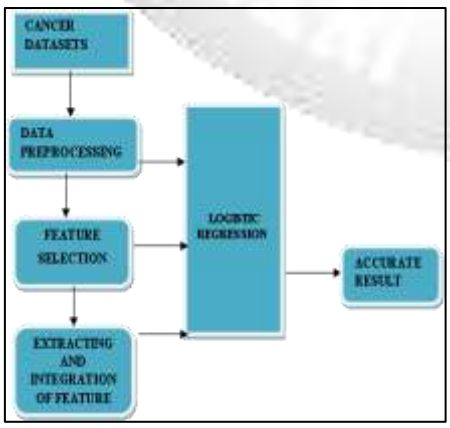

Fig. 1: Architecture Diagram

Data has been collected is from the Wisconsin Diagnostic Breast Cancer (WDBC) Data Set which categorizes breast tumor cases as either benign or malignant. based on the collected data it is further tuned by random forest algorithm into 9 features to predict the accurate result.

B. Data Pre-Processing

Initially Wisconsin Diagnostic Breast Cancer (WDBC) dataset which is available from the UCI machine learning repository is been analyzed The dataset has 32 attributes and 569 instances. It has two class labels (Malignant, Benign) for binary classification. All values of the features are real and continuous, which indicates that logistic regression is more appropriate to be used instead of decision tree or Apriori algorithm which prefers discrete values. Initially, the raw data was normalized by z-score standardization in view of the heterogeneity of medical data. The raw data was normalized by z-score standardization in view of the heterogeneity of medical data. The raw data in CSV format was converted into loaded to RDD.

C. Feature Selection

To find the relevant features. Information gain is the difference between the original information content and the amount of information needed. The features are ranked by the information gains, and then the top ranked features are chosen as the potential attributes used in the classifier. Frequent item sets capture all the dominant relationships between items in a dataset. The features is been extracted by using random forest method and the data is been labeled with 2 classes:

- Label → malignant: 0 or 1
- Features → {"thickness", "size", "shape", "madh", "epsize", "bnuc", "bchrom", "nNuc", "mit"}

D. Random Forest Algorithm Steps

1) Step l: For a given training dataset, extract a new sample set by N times repeated random sampling using bootstrap method. For example, from the data (xl5 yi),...,(xn,yn) to build a sample (xi , yi)....(xN , yN) .
2) Step 2: Build a decision tree or regression tree based on sample set resulted from step1;
3) Step 3: Repeat step 1-2, result in many trees, composing a forest.
4) Step 4: Let every tree in the forest to vote for xL
5) Step 5: Calculate the sum of votes for every class, the class with highest number of votes is the classification label for xL
6) Step 6: The percentage of incorrect classification is the classing error ratio of random forest.

IV. RESULT AND DISCUSSION

This work is carried out in core i3 processor with 2GB RAM in Linux platform through spark platform with two stages 1) data pre-processing and 2)by logistic prediction Cancer observations is done throughout in spark a parallel computing frame work spark ML provides a uniform set of high-level APIs built on top of Data Frames. In the first phase the raw data is been converted into the unstructured data of rdd strings and data frames are extracted using the feature extracted method with the random forest techniques in the second phase the extracted features is then used as

input to the logistic regression model where the data frames is been tested and trained using elastic net regulation technique and finally the calculation of metrics is been obtained with the accurate results that predicts the exact stage of cancer .

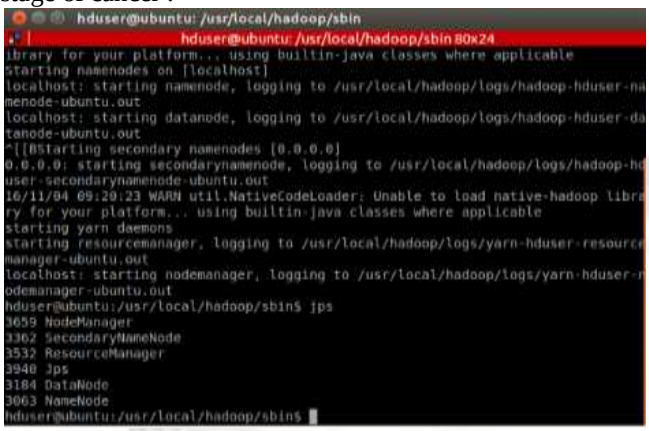

Fig. 2: Generating Namenode, Datanode

Fig. 3: Data frames

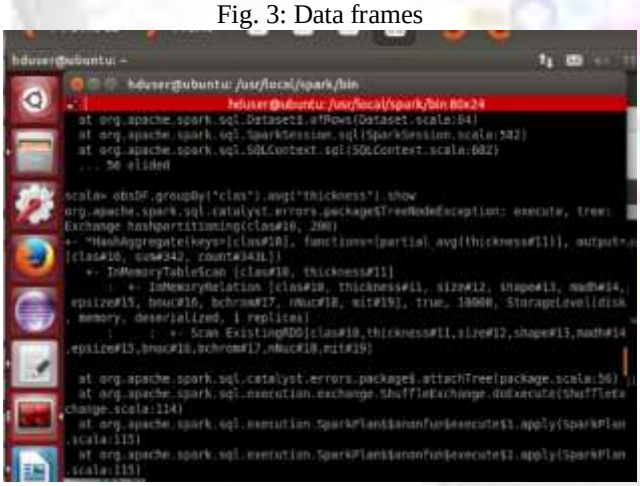

Fig. 4: Feature Extraction

In this chapter the three module of proposed system is implemented. Algorithm and modules are discussed in a narrative way, the existing and proposed are clearly explained.

V. CONCLUSION

Medical data in various organizational forms is voluminous and heterogeneous, it is meaningful and significant to utilize efficient data mining techniques to explore the development rules and the correlation of diverse diseases and discover the actual effect of treatments. However, it is a challenge for single-node data analysis tools with limited memory and computing power, therefore, distributed and parallel computing is in great demand.

In this phase a comprehensive medical data mining method consisting of mainly two steps: 1) data preprocessing; 2) logistic regression with Spark is done .Initially, the raw data in CSV format was converted RDD ,the memory based data objects of spark. The data has been visualized by their exact attributes and all the works is been implemented on the single node Environment .Spark MLlib has been implemented logistic regression, their experimental results proved that logistic regression in Spark MLlib was 100× faster than MapReduce

In the proposed system the goal is to train and test the logistic regression model with Elastic Net Regularization and to evaluate calculation metrics based on the predicted values by the ROC (Recall curve points) curve were the accurate results can be obtained.

REFERENCE

[1] Mandeep Rana, Pooja Chandorkar Alishiba Dsouza,"Breast Cancer Diagnosis And Recurrence Prediction Using Machine Learning Techniques" IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | p ISSN: 2321-7308,2015.

[2] Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader Benyettou –" Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules", International Journal of Computer Applications January 2013

[3] T.S. Subashini, V. Ramalingam, S. Palanivel "Breast mass classification based on cytological patterns using RBFNN and SVM" Expert Systems with Applications 36 (2009) 5284–5290,2009

[4] B Nithya," An Analysis on Applications of Machine Learning Tools, Techniques and Practices in Health Care System," Volume 6, Issue 6, June 2016

[5] Dengju Yao, Jing Yanglb, Xiaojuan Zhan," Predicting Breast Cancer Survivability Using Random Forest and Multivariate Adaptive Regression Splines", International Conference on Electronic & Mechanical Engineering and Information Technology,2011

[6] L. Chin, J. N. Andersen, and P. A. Futreal, "Cancer genomics: from discovery science to personalized medicine," Nature medicine, 2011.

[7] A, Jafari S. An expert system for detection of breast cancer using data preprocessing and Bayesian network Int J Adv Sci Technol, 2011.

[8] Jian pan" Bagging-Based Logistic Regression with Spark: A Medical Data Mining Method "2nd International Conference on Advances in Mechanical Engineering and Industrial Informatics AMEII 2016. 38, pp. e178-e178, 2016.

[9] Łukasz Neumann; Robert M. Nowak; Rafał Okuniewski; Witold Oleszkiewicz; Paweł Cichosz; Dariusz Jagodziński; Mateusz Matysiewicz." Preprocessing for classification of thermograms in breast cancer detection", Proc. SPIE 10031, Photonics

Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2016, 100313A (September 28, 2016); doi:10.1117/12.224930

[10] Amruta V. Shelke,"A Bayesian Network Based Classification of Breast Lesion in Digital Mammogram" International Journal of Scientific Engineering and Research (IJSER ISSN (Online): 2347-3878,Volume 3 Issue 2, February 2015

[11] Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence," 2008

[12] Jaree Thongkam, Guandong Xu and Yanchun Zhang, "AdaBoost Algorithm with Random Forests for Predicting Breast Cance Survivability," International Joint Conference on Neural Networks (IJCNN 2008), pp. 3062-3069, 2008

[13] Liu D C, Nocedal J. On the limited memory BFGS method for large scale Optimization[J].Mathematical programming, 2007

[14] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome," BMC bioinformatics, vol. 12, p. 323, 2006.

[15] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," 2010.

[16] Qiu H, Gu R, Yuan C, et al. "YAFIM: A Parallel Frequent Itemset Mining Algorithm with Spark.Parallel & Distributed Processing Symposium Workshops", 2014