International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com)
© 2020 IJSRSET | Volume 7 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
DOI : https://doi.org/10.32628/IJSRSET1207232

Breast Cancer Prediction Using Machine Learning Algorithm with Big Data Concept

R. Nirmalan[1], M. Javith Hussain Khan[2], V. Sounder [3], A. Manikkaraja[4]

[1]Assistant Professor, Department of Computer Science and Engineering, Bannari Amman Institute of Technology Sathyamangalam, Erode, Tamil Nadu, India

[2,3,4] UG Students, Department of Computer Science and Engineering, Bannari Amman Institute of Technology Sathyamangalam, Erode, Erode, Tamil Nadu, India

ABSTRACT

The evolution in modern computer technology produce an huge amount of data by the way of using updated technology world with the lot and lot of inventions. The algorithms which we used in machine-learning traditionally might not support the concept of big data. Here we have discussed and implemented the solution for the problem, while predicting breast cancer using big data. DNA methylation (DM) as well gene expression (GE) are the two types of data used for the prediction of breast cancer. The main objective is to classify individual data set in the separate manner. To achieve this main objective, we have used a platform Apache Spark. Here,we have applied three types of algorithms used for classification, they are decision tree, random forest algorithm, support vector machine algorithm which will be mentioned as SVM .These three types of algorithm used for producing models used for breast cancer prediction. Analyze have done for finding which algorithm will produce the better result with good accuracy and less error rate. Additionally, the platforms like Weka and Spark are compared, to find which will have the better performance while dealing with the huge data. The obtained outcome have proved that the Support Vector Machine classifier which is scalable might given the better performance than all other classifiers and it have achieved the lowest error range with the highest accuracy using GE data set

Keywords : Classification, Machine Learning, SVM, DNA

I. INTRODUCTION

In this era, organizations which are in various sectors are obtaining and crossing the huge limits of data than the last decade The information we are capturing are of with various types, large cases are including the huge data from the field of biomedical, sensors, spatiotemporal stream networks, social networks, etc. The data we are getting in a day to day life is due to spending our life blindly with the modern inventions all the time without concerning our traditional life style .

In our modern world to manage this huge amount of data cannot be possible without the modern techniques used these days. One of the major field which occupy this entire world is Machine learning. Machine Learning is the concept of making even the machine to think by themselves like a human. In the other word , it is like creating the sense to the human to make it work like a human without the help of anyone. This can be achieved by training the machine by providing the training data set to make an analysis and can perform a task by its own with the data. Here the training of machine can be obtained through three types of learning they are unsupervised, and

supervised, and reinforcement learning.For each type, of learning several algorithms and techniques willexist.

II. LITERATURE SURVEY

In Literature survey there been a many projects and implementation about prediction of disease or anything using artificial intelligence, neural networks and machine learning and in all the projects there will be a common thing which get existing and there wont be unique implementation and in all the prediction projects the common problem identified is the prediction would not support for the Big data, which is the most efficient one to take place in this modern century. We have found a many studies which deals with the prediction problem and recurrence using the vast concept, that is data mining techniques such as decision trees. Delen et al. Have used the concepts like decision trees, logistic regression and artificial neural networks to obtain the data models using the dataset. Lundin et al. used logistic regression models to predict breast cancer survival at different age person. They have analysed and study the most of the patient details with the factors like age, tubule formation, tumor necrosis etc . In this they have used the data mining concept for finding patterns which is the common factor and link between all the breast cancer patient.

A.Punitha et al. 2007 discussed the concept of breast cancer using genetic algorithm, adaptive resonance theory. They have trained totally 700 samples with 17 data missing, and 683 with the symptoms causing breast cancer. In that 64% are in beginning stage and 36% are in malignant.

III. EXISTING SYSTEM

1. If a patients came for check up, the previously existed methodology will only gather patients data individually. This is the only option that exists in our medical field till 2000.During that period the doctor should go through patients data manually. Sometimes there may be the probability of human errors. The rate of finding the cancer for more accurately may take several medical test over several period of time.

2. The new method which existing after 2000 will collect each and ever medical data of the patients over the database. Storing medical information over database will help the doctor to visualizing the data in digital form. By the help of modern medical records the probability of finding the diseases in the patients will be much more easy. The accuracy of the result by using the modern methodology like Machine learning will have a higher diseases prediction over the traditional one.

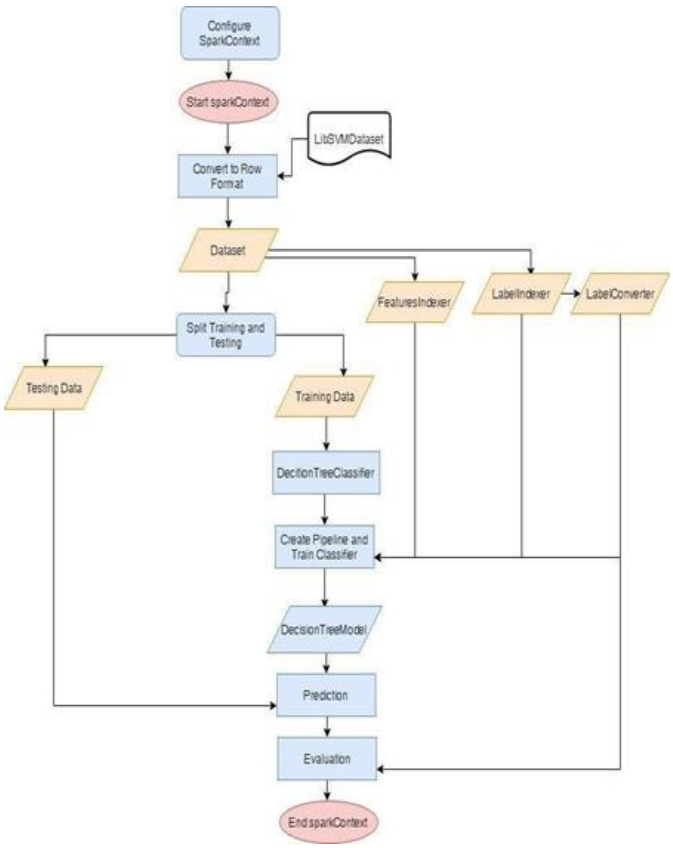
IV. PROPOSED SYSTEM

We here built and constructed the various models with the help of three algorithms which are used for classification : Random Forest algorithm, Decision tree algorithm and support vector machine algorithm. Library like ML lib in Spark used for implementations of the algorithms which will used for dealing with the big data.

Support Vector Machine algorithm. Used for classification on field like bio informatics, because of its features like performance, efficiency, and ability for dealing with the major concern like GE data, high-dimensional feature space its features effect is considered very high.

At the initial stage, Spark Context will get initiate and the major configuration get determined, like number of nodes to used for processing the huge data. Additionally they used Resilient Distributed Data set to store the data set. It provides a interface which are alike, and can deal with varieties of data emerging from various resources. The data which get stored in Resilient Distributed Data set can be partitioned and it can be distributed on cluster of data which allows the parallel processing of every portion

simultaneously. The work mentioned on the Spark context get distributed on clusters for performing the process on individual portion of data. The RDD may provides a mechanism called fault- tolerance in way of distributing individual three copies of partition of data on various cluster.



V. BRIEF METHODOLOGY

The samples of data used with methods of machine-learning will be described by the various features which will be of various and different types of values. The data's nature will be used to decide the different type of Machine-learning algorithm techniques used for obtaining information which are valuable.

The major Work is scaling up the algorithms used in machine learning for classification by applying individual data set jointly and separately. The various algorithms used for classification like support vector machine, random forest, decision tree for creating models which will be used for breast cancer prediction.

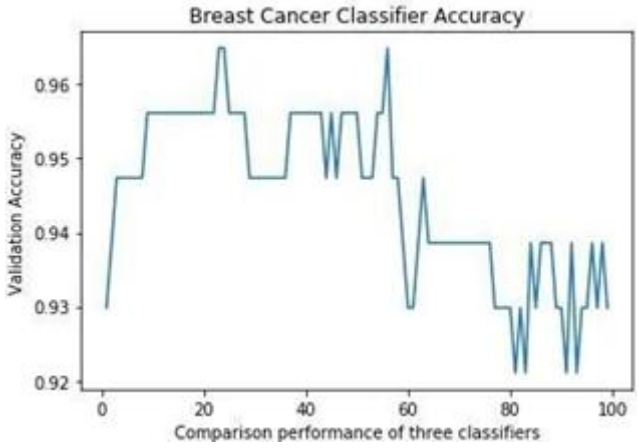


Fig 1.1

In Fig 1.1 it shows the comparison performance of three classifiers

Models	Accuracy	Precision
J48	0.896	0.909
REPTree	0.938	0.922
Random Tree	0.965	0.971

Fig 1.2

Fig 1.2 shows the accuracy of all the three algorithm which we have been used.

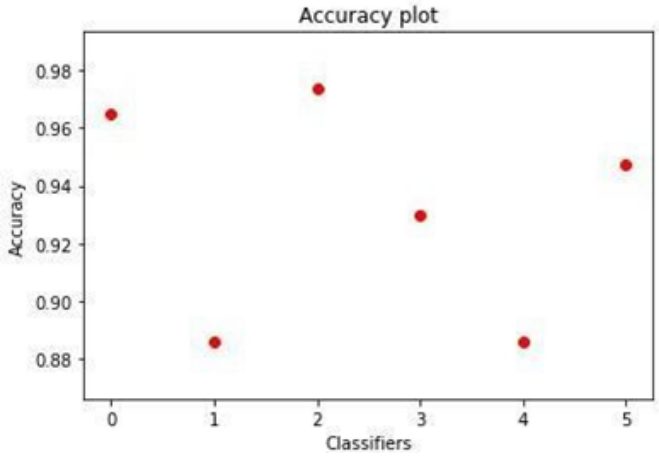


Fig 1.3

Fig 1.3 shows the accuracy plot of the classifiers and its accuracy.

Data Preprocessing

Preprocessing of data contains the way which will turn the data sets to a format in which the rows and columns related to the patients and genes. For working in spark environment, we may change the data sets to Lib SVM which is input for Spark, For Weka data sets loaded in format of comma separated values files. Additionally, configuration made for the Class Assigner of Weka get map for Class column. As a result the instances get classified in a way of "Patient"or "Normal".

Support Vector Machine:

Support Vector Machine is the important study which will be used for a supervised learning in for various purpose like recognize and analyzing the data patterns, for the purpose of regression analysis and classification. The SVM is one of the standard mechanism uses entire set of data for input and for prediction, SVM used to build a model which can allocate examples for one category or for other category to make it as a non-probalistic binary lineary classifier

Decision tree:

Decision tree is the efficient method and well-equipped methods used for classification. It gives benefit by generating rules which are explainable with conditions that makes to understand the correlation between gene and their share in occurance of breast cancer.Spark decision tree is a algorithm which undertakes a binary partitioning of space in a recursive way. In The decision tree every partition will occured by selecting a better split from the splits which are possible. This can be done to increase the information obtain at a tree node. It will handle the each and every data set in a way of row by row. It will be used for splitting the instances of data set.

VI. CONCLUSION AND FUTURE SCOPE

This various machine-learning techniques used in prediction of breast cancer. The best challenge is to obtain the accurate classifier, which will be used for all type of and huge amount of data. Here we have used three major techniques which is used for classification to ensure the best prediction of breast cancer. Here which our major aim is achieved using the Apache Spark which will be useful in case of big data. The additional comparison of using spark framework and traditional Weka framework is made by using it in the process.

We compared the the method of predicting the cancer with the high accuracy. The result obtained ensures that the SVM having a highest accuracy with the best performance at lower error rate. Further different studies should be conducted for improving classification techniques and its performance using different dataset and selection techniques. Another way is measuring deep learning performance using its architecture for predictions.

VII. REFERENCES

[1]. K. P. Murphy, Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning. Cambridge, Mass.: MIT Press, 2012.

[2]. M. Guller, Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large Scale Data Analysis. Berkeley: Apress, 2015.

[3]. International Agency for Research on Cancer (IARC) and World Health Organization (WHO). GLOBOCAN 2018: Age standardized (World) incidence and mortality rates, breast. Online]. Available: https://gco.iarc.fr/today/data/factsheets/cancers/ 20-Brea st-fact- sheet.pdf

[4]. "DNA Deoxyribonucleic Acid," 2016. Online]. Available: http://www.myvmc.com/anatomy/dna-deoxyribonuclei c- acid/

[5]. Y. Lu, and J. Han, "Cancer classification using gene expression data," Information Systems, vol. 28, no. 4, pp. 243–268, 2003.
[6]. M. M. Babu, "Introduction to microarray data analysis," Computational genomics: Theory and application, vol. 17, no. 6, pp. 225–49, 2004.
[7]. T. Mikeska, and J. M. Craig, "DNA methylation biomarkers: cancer and beyond," Genes, vol. 5, no. 3, pp. 821–864, 2014.
[8]. S. B. Baylin, "DNA methylation and gene silencing in cancer," Nature Reviews Clinical Oncology, vol. 2, no. S1, p. S4, 2005.
[9]. A. Einstein, B. Podolsky, and N. Rosen, "Can quantum-mechanical description of physical reality be considered complete?" Physical Review, vol. 47, no. 10, p. 777, 1935.
[10]. Spark 2.1.0," 2018.
[11]. "Apache Spark™ - Unified Analytics Engine for Big Data," Spark.apache.org, 2018. Accessed on: Nov. 10, 2018 Online]. Available: http://spark.apache.org/
[12]. "Spark Programming Guide – Spark 2.0.1 Documentation," Spark.apache.org, 2018. Accessed on: Oct.15, 2018 Online]. Available: https://spark.apache.org/docs/2.0.1/programming- guide.html

Cite this article as :

R. Nirmalan, M. Javith Hussain Khan, V. Sounder, A. Manikkaraja , "Breast Cancer Prediction Using Machine Learning Algorithm with Big Data Concept", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 2, pp. 123-127, March-April 2020. Available at doi : https://doi.org/10.32628/IJSRSET1207232
Journal URL : http://ijsrset.com/IJSRSET1207232