TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 6240 - 6244

Breast Cancer Prognosis with Apache Spark Random Forest Pipeline

Timmana Hari Krishna
Dept. of Computer Science and Engineering,
Bharath Institute of Higher Education and Research, Chennai, India

Dr C. Rajabhushanam
Dept. of Computer Science and Engineering,
Bharath Institute of Higher Education and Research, Chennai, India

Article Info
Volume 82
Page Number: 6240 - 6244
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 30 January 2020*

Abstract:
Brest cancer is one of the most common cancers diagnosed in women in western countries. Breast cancer research and awareness supports the improvements in cancer diagnosis and treatment. Early detection of Breast cancer improves the survival rates and decreases the number of deaths related to this disease. Recently Computer concepts are spread across all domains including medical and healthcare. Data science and machine learning techniques are used in cancer prediction and analysis to get rapid accurate results. The cancer prediction involves the identification malignant cells from breast cells. Researchers and Pathologists used the several machine learning algorithms like K-Nearest Neighbors, logistic support vector machine, artificial neural networks and decision tree in cancer prediction. They did not conclude the feasible method for cancer prediction. In this paper we propose a scalable, fault tolerant pipeline model that analyses big cancer data in and predicts the cancerous cells in real time. This model is developed on Apache Spark using Machine Learning Pipeline. In this paper, we implemented our pipeline using Random Forest algorithm to compare with baseline model in terms of accuracy and performance.

Keywords:*Apache Spark, Machine Learning pipeline, Cancer Prediction, Random Forests.*

I. INTRODUCTION

Breast cancer is a type of cancer that grows from breast tissue. It is one of the major cancersdiagnosed in western countries. Breast cancer research and awareness supports the improvements in cancer diagnosis and treatment. Early detection of Breast cancer improves the survival rates and decreases the number of deaths related to this disease. In the view of cancer, the cells are of two types: (a) Benign cells that are non-cancerous and (b) Malignantcells that are cancerous. The cancer prediction involves the identification malignant cells from breast cells. In this paper we described the different type of cancers and the symptoms with prevention methods.Pathologists use several techniques to predict breast cancer. Researchers used the different statistical, machine learning methods in breast cancer prediction. But they did not able to conclude the best technique. We propose a model developed with spark machine learning pipeline that processesvoluminous data fast and accurate. We used Original cancer data set of UCI Machine learning repository as input for analysis. Apache Spark is a scalable, fault tolerant in memory computing engine that handles big data. It provides

January - February 2020
ISSN: 0193 - 4120 Page No. 6240 - 6244

rich library to implement machine learning algorithms in an effective manner. In Machine learning, Pipeline is used to chain the prediction process which executes continuously. We model Random Forest algorithm with pipe line and without pipeline. Then we compare both models in terms of accuracy and performance.

II. BREAT CANCERFUNDAMENTALS

In this section, we describe the different types of the breast cancer and their common symptoms along with prevention methods. Cancer cells are differentiated as malignant cells from benigncells i.e. normal cells. The samples of breast cancer cell and normal cell are as shown in Fig 1.

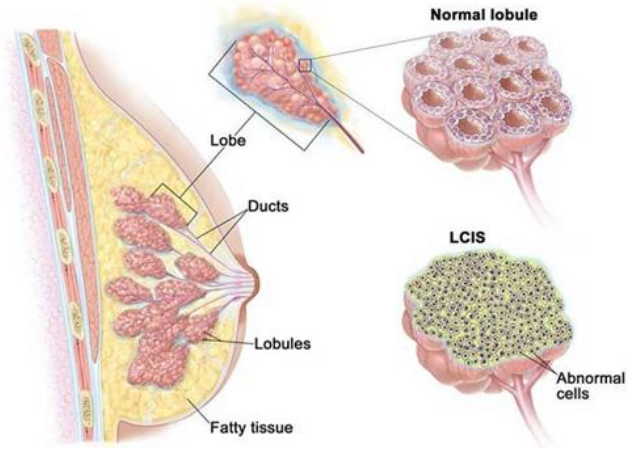

Fig 1. Normal and Cancerous Breast cells (source: http://cancer.gov)

A. Different types of Brest Cancers

Different cancer types include:(a) Angiosarcoma is grown as padding of the blood vessels and lymph vessels. The lymph vessels are part of immune system those collect bacteria, viruses and waste products and dispose of them from the body. It forms on the skin of head, neck, breast along with deeper tissues like liver and hearts. The treatment options contain surgery, radiation therapy and chemotherapy. (b) Ductal carcinoma in situ (DCIS) is initial and non-offensive form of breast cancer that forms inside the milk conduit. Radiation and breast-conserving surgery are used to treat this type of cancer (c)Inflammatory breast cancer is a rare type of breast cancer that looks like normal breast infection. This makes breast swollen, tender and red in color. It spreads very fast by blocking the lymphatic vessels that covers the Brest. (d) Invasive lobular carcinoma is a type of breast cancer that begins in the milk-producing glands (lobules) of the breast, not from lump. It spreads throughout the body along with lymph nodes by broken out of lobules. (e) Male breast cancer forms in the men breast tissue. Generally, people think breast cancer occurs only in women but appears in men also. Due to unawareness this cancer is diagnosed in advanced stages. It occurs at all ages, mostly in old. Early stage diagnosis of this cancer has good chance of cure. (f) Paget's disease occurs in the women of age 50 that starts on the nipple and spreads to the dark circle of skin (areola) around the nipple. (g) Recurrent breast cancer is breast cancer that occurs again after treatment due to survived cancerous cells of initial treatment. These undetected cancer cells increase and lead to recurrent breast cancer

B. Brest Cancer Symptoms

Breast cancer symptoms includes:(a) Brest lump differs from neighboring tissues(b) Breast size, shape and form will be changed (c) Skin over the breast will be lumped. (d) Nipple looks overturned. (e) Skin pigmented area encased by Brest skin is peeled, scrambled, covered and crumbled. (f) The breast skin turns into orange color

C. Brest Cancer Prevention methods

Breast cancer riskprevention methods includes:
(a) Frequent breast cancer screening exams helps to identify the cancer in early stage (b)Breast awareness may help the patient in cancer detection (c) Reduce or avoid the consumption of alcoholic products (d) Body exercise for 30 minutes a day reduces the risk of cancer (e) Hormonotherapy is one of the risks of cancer occurrence (f) Maintaining proper weight by daily exercise. (g) Proper and healthy diet that contains fruits and vegetables, whole grains, legumes, and nuts

January - February 2020
ISSN: 0193 - 4120 Page No. 6240 - 6244

III. Apache Spark Pipeline and Model

Big data is a vast data that is difficult to handle by conventional systems and is in unstructured, structured and partially structured formats. Machine Learning is concept to implement algorithm to train the system which rivalled human learning. Pipeline is machine learning standard that allows chaining the linear sequence of data transforms to evaluate the modeling process. It ensures that all pipeline steps are designed to evaluate the available data including training data set and cross validation process.

Spark machine learning library provides standard application programming interface to implement machine learning algorithms simpler. It provides feasibility to develop pipelines and workflow by combining the multiple algorithms. Each ML pipeline contains following components.

Data Frame: This is the basic component of the spark SQL that handles columned data. It is used to store feature vectors, labels, predictions and data. Spark ML API can work with data frames.

Transformer: This is a ML algorithm that transform data frame features into predictions.

Estimator: A ML learning algorithm that trains the data frames to generate the model

Pipeline: A process that creates a workflow by chaining multiple estimators and transformers.

Parameters: A parameter set that can be shared by transformers and estimators.

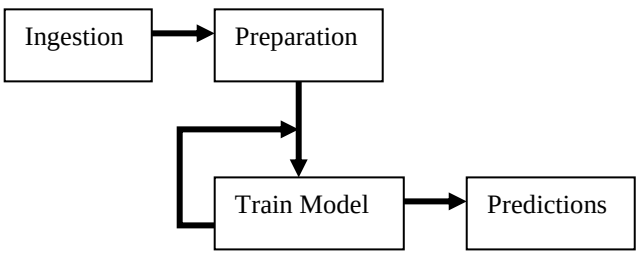

Fig 2 A Typical Pipeline Model

The sample Pipeline process is shown in Fig 2. It consists of four phases: (a) Data Ingestion: In this phase, input data loaded into spark system and converted into data frames. (b) Data Preparation: Input data may not be proper and may contain missing values. In this phase, prepare the proper input data for the model by eliminating or predicting the missing values. Output data is used to build and train the model. (c) Train and build Model: In this phase, features are added to data frames and transform them to predictions with data frames by training the model. (d) Predictions: In this phase we evaluate the predictions from the model using train data and test data.

IV. Input DataSet and Features

In this section we project the description about the input Cancer Data Set Attributes and Features. Original Wisconsin Breast Cancer Database of UCI repository is used as input data set for Experiments. This dataset was created by Dr. WIlliam H. Wolberg, University of Wisconsin Hospitals, Madison, Wisconsin, USA and donated by OlviMangasarian and David W. Aha. This multivariate dataset consists of 699 samples with 11 attributes. The attributes are namely Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses and class values that contain 2 for benign, 4 for malignant.

A. Dataset Attributes

This dataset contains following attributes. (a)Clump thickness: Cancer cells are accumulated in multiple layers where normal cells are grouped in monophonic layers. (b) Uniformity of cell size/shape: Cancer cell differs in size and shape. This attribute is use to identify that whether the sample is cancerous or not. (c) Marginal adhesion: Normal cells tend to penetrate together whereas cancer cells cannot penetrate. (d) Epithelial cell size: Epithelial cells that are significantly enlarged if it is impacted by cancer (e) Bare nuclei: This value is available only in normal cells. (f) Bland Chromatin: It deals with the surface of sample cell. If the cell is cancerous the surface will be rough where it is softer for normal cell. (g) Normal nucleoli: The small visible structures of nucleus is called Nucleoli. This

January - February 2020
ISSN: 0193 - 4120 Page No. 6240 - 6244

Nucleoli is too small and invisible in normal cell whereas very noticeable in cancerous cells.

B. Cancer Observation Schema

The sample breast cancer schema is as shown in Table I.

C. Random Forests Algorithm

The pseudo code of Random Forest prediction is as follows:

- Step 1: Read the input data set for "n" features.
- Step 2: Choose subset of features and name it as "i" from "n" features randomly
- Step 3: Compute the node "n" among "i" features base on finest fit
- Step 4: Base on best split, divide the node into child nodes.
- Step 5: Repeat 2-4 steps until got "j" nodes i.e. trees
- Step 6: Repeat 2-5 step for "m" times to get "m" number of trees to create a forest.
- Step 7: Test features and generated discussion trees are used for prediction. This is stored as target
- Step 8: Calculate the generalized value called vote for each predicted target
- Step 9. The high voted target is considered as final prediction.

TABLE I. BREAST CANCER SCHEMA

Attribute Name	Id	Type
Sample code number	seqno	Integer
Class: (2-benign, 4-malignant)	class	Integer
Clump Thickness	thickness	Double
Uniformity of Cell Size	size	Double
Uniformity of Cell Shape	shape	Double
Marginal Adhesion	madh	Double
Single Epithelial Cell Size	epsize	Double
Bare Nuclei	bnuc	Double
Bland Chromatin	bchrom	Double
Normal Nucleoli	nNuc	Double
Mitoses	mit	Double

V. EXPERIMENTS AND DISCUSSION

In this section we compared the performance parameters of the algorithm with pipeline and without pipe line. The experiment results are captured in tables II.This experiment shows that Pipeline gives good accuracy and less error compared to without pipeline.

TABLE II. COMPARISION BETWEEN BASELINE AND PIPELINE

Parameter	Baseline	Pipeline
Accuracy	0.9951273532668882	0.9955149501661129
Mean Squared Error	0.030150753768844216	0.025125628140703522
Mean Absolute Error	0.03015075376884422	0.02512562814070352
Root MSE Squared	0.17363972405196978	0.15851065623706037
Root Squared	0.8677740863787375	0.8898117386489479
Explained Variance	0.22802454483472642	0.22951440620186364

VI. CONCLUSION

In this paper, we have presented the fundamentals of Breast cancer, types of Breast cancers and their symptoms along with the prevention methods. We proposed Apache spark Machine Learning pipeline model that processes the cancer data set and predicts the cancer fast and accurate. We implemented the Random Forests algorithm modelusing both base line and pipeline. We have compared the both model in terms of accuracy,error and performance.

REFERENCES

1. T.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

January - February 2020
ISSN: 0193 - 4120 Page No. 6240 - 6244

2017,.
2. Dana Bazazeh and RaedShubair,"Comparative study of machine learning algorithms for breast cancer detection and diagnosis",5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), 2016
3. AlirezaOsareh and BitaShadgar, "Machine learning techniques to diagnose breast cancer",5th International Symposium on Health Informatics and Bioinformatics (HIBIT), 2010
4. Ahmed F. Seddik and Doaa M. Shawky,"Logistic regression model for breast cancer automatic diagnosis", SAI Intelligent Systems Conference (IntelliSys), 2015
5. Ganesh N. Sharma, Rahul Dave,JyotsanaSanadya,Piush Sharma and K. K Sharma,"VARIOUS TYPES AND MANAGEMENT OF BREAST CANCER: AN OVERVIEW", Journal of Advanced Pharmaceutical Technology & Research, pp.109–126,Apr-Jun 2010
6. Chandresh Arya and Ritu Tiwari, "Expert system for breast cancer diagnosis: A survey", International Conference on Computer Communication and Informatics (ICCCI), 2016
7. PubMed Forums,"Breast cancer: Overview",Informed Health Online,July 27, 2017.
8. Spark Machine learning documentation site at https://spark.apache.org/docs/2.2.0/ml-pipeline.html
9. PubMed Forums,"Female Breast Cancer (Female Breast Carcinoma): Symptoms"
10. Aiello EJ1, Buist DS, White E, Seger D and Taplin SH, "Rate of breast cancer diagnoses among postmenopausal women with self-reported breast symptoms",The Journal of the American Board of Family Medicine,Dec 2004..
11. Big data wiki Available at https://en.wikipedia.org/wiki/Big_data
12. Spark Documentation site available at https://spark.apache.org/
13. UCI Breast Cancer Data Set at https://archive.ics.uci.edu/ml/datasets/breast+cancer
14. Mona Botros and Kenneth A Sikaris,"The De Ritis Ratio: The Test of Time", The Clinical Biochemist Reviews, pp.117-130, Nov 2013