Summary of Research Article:
"On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context"

Problem statement

The proliferation in medical science modernized medical conditions examination of patients. Nowadays, a massive amount of data can be gathered and processed like gene expressions (GE) and DNA methylation (DM).

Breast Cancer prediction is a matter of concern as 2.3 million cases were reported in 2020 with six lakhs+ deaths. According to the experts, early diagnosis could prevent the adversities of breast cancer.

The authors discussed how to scale machine learning classification tasks for large datasets using the Apache spark in this work. Results of three different classification algorithms are presented in comparison with WEKA.

Tools

- Spark Architecture consists of Master node, worker nodes, and cluster manager.
- Libraries from Spark MLlib
- RDD for data storage
    - ❖ Stores 3 copies of data to ensure fault tolerance.
    - ❖ Automatic partitioning of data and distribution across clusters.
- Waikato Environment for Knowledge Analysis - WEKA (Command line)

Solution Methodology

Step 1. Preprocessing

- Preparing dataset for spark context

- Data formatting for WEKA

Step 3. Initialize spark context
Adding all configurations and reformatting dataset to LabeledPoint type in RDD.

Step 4. Classification models
The GE and DM datasets are processed using the following algorithms on Spark and WEKA.

- Linear SVM for binary classification
- Decision Trees
- Random Forest

Step 6. Prediction and validation
Dataset divided into 60% training and 40% testing; classification models evaluated on the following metrics.

- Accuracy
- Error Rate
- Area under ROC curve

Step 7. Result Analysis
In particular, results are presented for both WEKA and spark for each of the 9 nine models corresponding to GE, DM, and [GE DM] combined.
Apache Spark's SVM outperforms every other model with 99.68% accuracy.

Source of Data
The Cancer Genome Atlas—Data Portal
Gene expression and DNA Methylation Datasets
Details:
254 Samples with the following distribution for both the datasets
Cancer present - 215

Normal - 39

16,077 genes are present per sample.

Implementation Idea

We plan to apply the same steps as mentioned in the solutions methodology section.

The only challenge that we are facing is that the original dataset cannot be obtained. We also dropped messages to the authors to provide the dataset; we are yet to receive a reply from them.

The original dataset was huge in terms of gene expression [16,077 genes per sample].

We searched for some similar publicly available datasets to continue with the work.

Mendeley Data: Gene Expression Profiles of Breast Cancer

Total no. of samples: 590

Cancer Diagnosed - 529

Normal - 61

17,814 genes per sample.

This dataset has almost double the number of samples compared to the original dataset and a comparable size of gene expression.

In the end, we will illustrate the results from WEKA and Spark both.

Student Name	*Student ID (DAIICT)*	*Student ID (IIT Jammu)*
Ambuj Mishra	202116003	2021PCS1017
Arpita Nema	202116004	2021PCS1018