MapReduce Based Feature Selection and Classification of Microarray Dataset

by

Nitish Kumar Rath

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela, Odisha, 769 008, India

May 2015

MapReduce Based Feature Selection and Classification of Microarray Dataset

Thesis submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Nitish Kumar Rath

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela, Odisha, 769 008, India

May 2015

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India.

Certificate

This is to certify that the work carried out in the project titled "***MapReduce Based Feature Selection and Classification of Microarray Dataset***" by **Nitish Kumar Rath**, is an original work carried out by him under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in the Department of Computer Science & Engineering, National Institute of Technology, Rourkela. Neither this thesis not any part of it has been submitted for any degree or academic award anywhere else.

Place: NIT Rourkela
Date: 11^{th} May 2015

Prof. Santanu Ku. Rath
Professor, CSE Department
NIT Rourkela, Odisha

Acknowledgment

No undertaking is completed completely by an individual, many individuals have helped all through for the finishing of this task and each of their input has been valuable. My deepest gratitude goes to my supervisor, **Dr. Santanu Kumar Rath**, Professor, Department of Computer Science & Engineering, for his guidance, support and encouragement throughout the period of this work.

I would like to thank **Mr. Mukesh Kumar**, for his valuable and important suggestions. His readiness for consultation at all times, his educative comments, his concern and assistance even with practical things have been valuable.

I must acknowledge the academic resources that we have acquired from NIT Rourkela. I want to thank the staff from the office who have been sufficiently benevolent to help in this project.

I am thankful to all my friends. I sincerely thank everyone who has provided me with inspirational words, a welcome ear, new ideas, constructive criticism, and their invaluable time.

Last but not the least, I would like to dedicate this project to my family, for their love, patience and understanding.

Nitish Kumar Rath

*Roll No.-**111cs0155***

Author's Declaration

I hereby declare that all the work contained in this thesis is my own work unless otherwise acknowledged. Also, all of my work has not been previously submitted for any academic degree. All sources of quoted information have been acknowledged by means of appropriate references.

Nitish Kumar Rath

National Institute of Technology, Rourkela

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela, Odisha, 769 008, India

May 2015

Abstract

Gene expression profiling has emerged as an efficient technique for classification, diagnosis and treatment of various diseases. The data retrieved from microarray contains the gene expression values of the genes present in a tissue. The size of such data varies from some kilobytes to thousand of Gigabytes. Therefore, the analysis of microarray dataset in a very short period of time is essential.

The major setback of microarray dataset is the presence of a large number of irrelevant information, which hinders the amount of useful information present in the dataset and results in a large number of computations. Therefore, selection of relevant genes is an important step in microarray data analysis. After retrieving the required number of features, classification of the dataset is done.

In this project, various methods based on MapReduce are proposed to select the relevant number of feature. After feature selection, Nave Bayes Classifier and N-Nearest Neighbor is used to classify the datasets. These algorithms are implemented on Hadoop framework. A comparative analysis is done on these methodologies using microarray data of different sizes.

Keywords: Gene Expression, Classification, Feature Selection, MapReduce, Hadoop

Nitish Kumar Rath

*Roll No.-**111cs0155***

Contents

List of Figures

Chapter 1

Introduction

1.1 Microarray Data

A microarray is a complex lab-on-chip that consists of a large number of microscopic DNA spots that act as probes for measuring the expression levels for various genes in the human genome. Figure 1.1 shows a microarray chip and Figure 1.2 gives the structure of the chip at a microscopic level.

Figure 1.1: Microarray chip

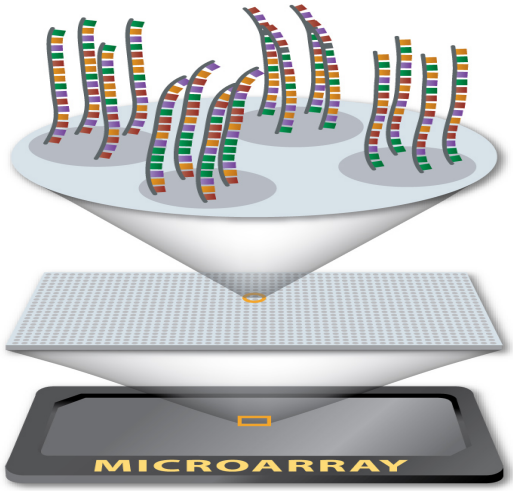

Figure 1.2: DNA spot on the Microarray chip

When a sample of the cultured tissue is applied to the chip the spots on the chip lighten up. The intensity of the light from a spot gives the expression levels for that corresponding gene. The microarray chip is scanned by a microarray scanner which converts the light intensity to its corresponding numerical value. These values are used in the analysis of the microarray dataset. The expression values of the genes are different in different individuals depending on whether host

ID_REF	VALUE_DS
AFFX-BioB-5_at	1126.74
AFFX-BioB-M_at	1167.66
AFFX-BioB-3_at	840.345
AFFX-BioC-5_at	2572.97
AFFX-BioC-3_at	3291.99
AFFX-BioDn-5_at	6171.86
AFFX-BioDn-3_at	11723.4
AFFX-CreX-5_at	27573.8
AFFX-CreX-3_at	28086.4
AFFX-DapX-5_at	350.522
AFFX-DapX-M_at	777.328
AFFX-DapX-3_at	1236
AFFX-LysX-5_at	58.9167
AFFX-LysX-M_at	113.665
AFFX-LysX-3_at	188.708
AFFX-PheX-5_at	130.262
AFFX-PheX-M_at	108.407
AFFX-PheX-3_at	256.371
AFFX-ThrX-5_at	82.8532
AFFX-ThrX-M_at	136.71

Total number of rows: **54675**

Figure 1.3: Sample of Microarray

is healthy or infected. Figure 1.3 shows a snap shot of a microarray dataset.

1.2 Problem Definition

Analysis of microarray data is the key for accurate diagnosis and treatment of various diseases. Constructing a classification model would help in early prediction of a disease.

But analysing it from a proper perspective is difficult due to its large dimension. It leads to computational instability and makes the model of the classifier more complex. Hence selection of relevant genes is an imperative in the process of microarray analysis.

After selection of relevant genes a reduced dataset is formed with improved

information quality and computationally less complex. This dataset is used for constructing the classification model.

1.3 Distributed computing

A distributed computing is a system in which different systems in a cluster or grid convey and coordinate their activities by passing messages. Different distributed techniques are in use for many years, but they require an efficient network for communication between them. Thus network bandwidth becomes a bottleneck in such situations. Hadoop comes into rescue in this situation. Hadoop is an open source framework for parallel storage and processing of large datasets like microarray datasets. It follows a master-slave architecture. It contains three major components:-

1.3.1 Hadoop Distributed File System(HDFS):

It is a distributed file system, in which the datasets are stored as blocks in the individual member nodes of the cluster. It consists of a Namenode and Datanode. The datanodes store the data in the form of blocks while the namenode keeps a track of blocks on different datanodes, i.e, it stores the metadata of the cluster.

1.3.2 YARN:

It is the resource manager for the framework. It keeps a track of the resources of the cluster and their allocation It consists of a central ResourceManager on master node and Nodemanager on every slave node for its working.

1.3.3 MapReduce Programming Model:

It is a programming paradigm for distributed processing of the datasets. It consists of two major phases. The first phase is the map phase, where the map function is applied to (mapped to) the blocks in the datanodes and the obtained output is writtento an intermediary file in the HDFS. This is then sent to the reducer

where the results from the mappers are combined and simplified (reduced) to get the final output.

1.4 Thesis Organization

The rest of the thesis is organized as follows.

- Chapter 1 gives the introduction of the theory and brief introduction about the project.
- Chapter 2 states the Review of related work & Motivation.
- Chapter 3 briefs the proposed feature selection techniques.
- Chapter 4 briefs the proposed classification techniques.
- Chapter 5 shows the results & implementation.
- Chapter 6 gives the conclusion.

Chapter 2

Literature Review

As mentioned earlier analysis of microarray dataset helps in early diagnosis and treatment of diseases. But due to their huge dimension, it is wiser to explore the field of distributed computing for efficient storage and parallel processing. This chapter discusses the related work in this field and the motivation for carrying the project.

2.1 Review of related work:

A number of gene selection techniques have been proposed by various practitioners in the past. A. K. M. Tauhidul Islam et al. [1] have proposed a parallel gene selection technique based on MapReduce , that utilizes sampling techniques to reduce irrelevant genes by using Between-groups to Within-groups sum of square (BW) ratio. The BW ratio indicates the variances among gene expression values. After gene selection, it applies KNN technique in parallel using MapReduce programming model. Finally, the accuracy of the method is verified through extensive experiments using several datasets.

Shicai Wang et al. [2] have proposed a method for calculating correlation and introduced an algorithm based on MapReduce to optimize storage and correlation calculation. This algorithm is used as a basis for optimizing high throughput molecular data (microarray data) correlation calculation.

Statistical tests are sophisticated procedures to analyze the behavior of data [3]. The statistical tests are used as a gene selection method by assuming the hypothe-

ses, i.e., Null hypothesis and alternate hypothesis. Based on the correctness of the hypothesis, the features are either selected or rejected. The K-Nearest Neighbor classifier provides for a non-parametric procedure for the assignment of a class label to the input based on the class labels represented by the K-nearest training samples [4].

2.2 Motivation

Analysis of microarray datasets would help physicians for early diagnosis and treatment of various diseases. But due to the curse of dimensionality it is difficult to analyse them. Presence of huge amount of genes not only leads to loss of useful information but also computational instability. Hence avenues for faster processing and efficient storage must be explored.

2.3 Objective

Motivated by the requirement of faster processing and efficient storage of datasets, the following objectives were undertaken:

- To explore the concept of Distributed Processing
- To implement algorithms on Distributed Systems

Chapter 3

Gene selection

3.1 Gene Selection

Gene selection a is the method by which a given gene is selected or discarded so as to get a revised dataset with a reduced number of genes and increased information. Reduced number of genes leads to reduced computations and increase in the information quality of the dataset.

3.2 Advantages of Gene Selection

The Advantages of gene selection are:

- Less number of computations.
- Improve in information quality of dataset.
- Less complex classification models.

3.3 Proposed Algorithms

In this work, hypothesis testing algorithms are used for gene selection. In this type of testing, two hypotheses are assumed

- **Null Hypothesis:** It assumes that the gene sets are a random samples of the same population, thus there is no significant difference between them
- **Alternate Hypothesis:** This assumes that there exists some significant difference between the gene sets. Thus they represent the data effectively.

Hypothesis testing algorithms test the basic properties of the dataset like mean, variance, ranks etc. and a statistic score is obtained from them. Based on this score it is decided whether to accept or discard a gene.

In this work, ANOVA (Analysis of Variance) test is performed for gene selection.

Chapter 4

Classification Of Microarray Data

4.1 Introduction

Classification is a technique for predicting the class or type of a new input sample based on the knowledge obtained from the training samples. In this work, two types of classifiers are explored. The K-Nearest Neighbor classifier and Naive Bayes classifier.

4.1.1 K-Nearest Neighbor

K-Nearest Neighbor (KNN) classifier has been used to create a classification model for the datasets. KNN is an example of instance based classifier. In KNN the distance between the training sets and a testing sample is calculated. The smallest k distances are then enumerated and the testing sample is classified into the modal class of the k training samples.

Chapter 5

Results And Interpretation

In this chapter the result obtained for the proposed algorithms are discussed. For the purpose of execution of the algorithm LEukemia dataset is used.

5.1 Gene Selection

The proposed algorithms for gene selection were implemented using MapReduce programming model and executed on Hadoop Framework. The time taken for the execution on a conventional machine as well as a Hadoop cluster were recorded as shown in figure 5.1.

As is evident from the graphs, the time taken for execution on a Hadoop cluster is much less than that taken by a normal machine. After gene selection, 43604 genes were selected in the case of the ANOVA test and 43376 genes in the case of kruskal-wallis test.

5.2 Classification

After selection of relevant features, the data set is reduced and divided into a training and testing set for the purpose of classification. The third sample in each dataset is taken as a testing sample and the rest as training sample.The algorithms were implemented using MapReduce programming model and executed on Hadoop framework. An optimal model was found by changing the number of input genes to the classification model. The accuracy plots aregiven in figure 5.2, As it can be seen the accuracy first increases until a certain number of features and from

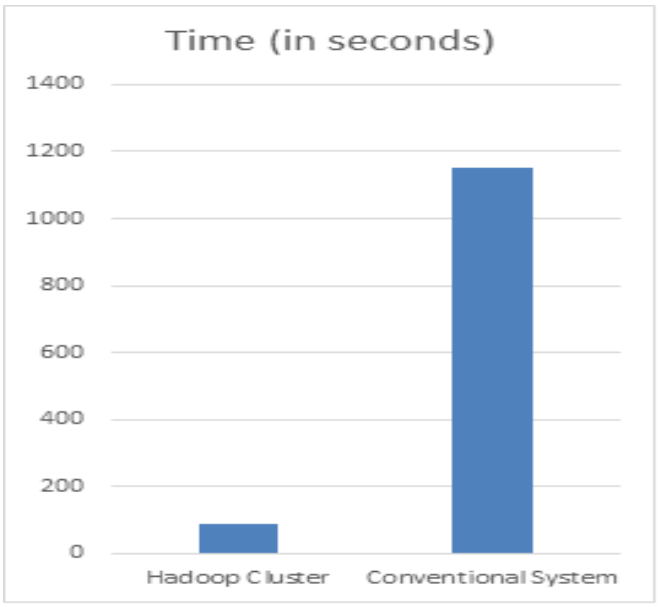

Figure 5.1: Time Taken for gene selection test

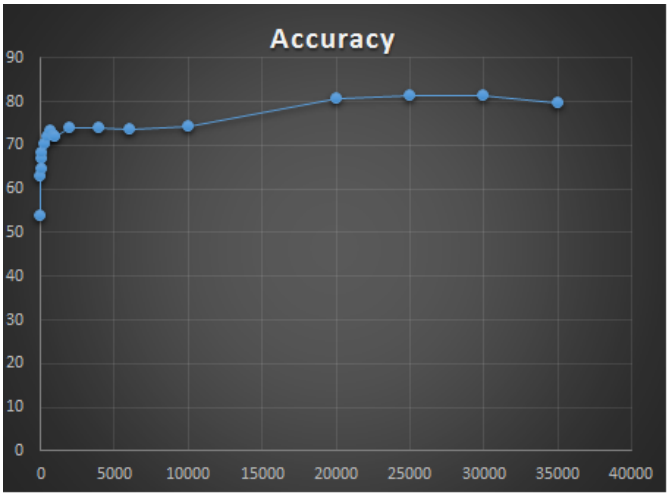

Figure 5.2: Accuracy plot for Classifier

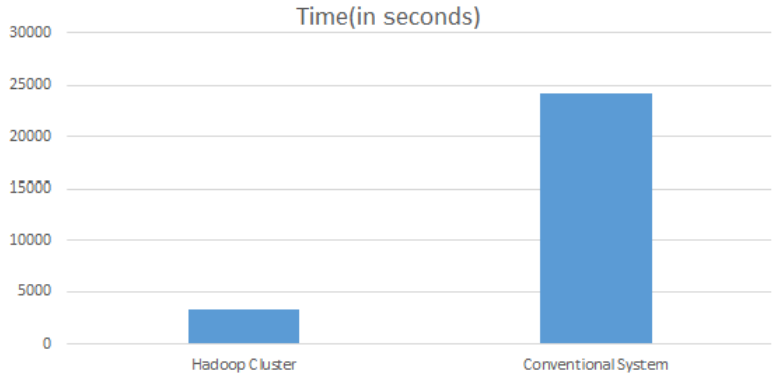

Figure 5.3: Time plot for Classifier

that point onwards it either decreases or remains constant. The time plot for the classification algorithms are shown in figure 5.3. As can be seen from the time plots, the time taken by a Hadoop cluster is much less than that taken by a normal system.

Chapter 6

Conclusion

In this work an attempt has been made to explore various algorithms for analysis of microarray datasets. Algorithms were developed for execution in a parallel manner. The optimal accuracy for various models were found by varying the number of genes in the classification model. The major contributons of the project are :

- Use of distributed computing techniques
- Development of algorithms to be executed in parallel manner

Bibliography

[1] A. T. Islam, B.-S. Jeong, A. G. Bari, C.-G. Lim, and S.-H. Jeon, "Mapreduce based parallel gene selection method," *Applied Intelligence*, pp. 1–10, July 2014.

[2] S. Wang, I. Pandis, D. Johnson, I. Emam, F. Guitton, A. Oehmichen, and Y. Guo, "Optimising parallel r correlation matrix calculations on gene expression data using mapreduce," *BMC bioinformatics*, vol. 15, p. 351, November 2014.

[3] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2003.

[4] M. Kumar and S. Kumar Rath, "Classification of microarray data using kernel fuzzy inference system," *International Scholarly Research Notices*, vol. 2014, p. 18 pages, August 2014.

Dissemination of Work

Accepted

1. Mukesh Kumar, Nitish Kumar Rath, Santanu Kumar Rath. Feature selection And Classification Of Microarray Data using Mapreduce based ANOVA K-Nearest Neighbour. *11th International conference on Data Mining And Warehousing,* Bangalore, India, 2013.