Journal of Biomedical Informatics xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier

Mukesh Kumar *, Nitish Kumar Rath, Santanu Kumar Rath

Department of Computer Science and Engineering, NIT Rourkela, Orissa 769008, India

ARTICLE INFO

Article history:
Received 18 September 2015
Revised 28 February 2016
Accepted 2 March 2016
Available online xxxx

Keywords:
Big data
Classification
Hadoop
K-nearest neighbor
MapReduce
Microarray

ABSTRACT

Microarray-based gene expression profiling has emerged as an efficient technique for classification, prognosis, diagnosis, and treatment of cancer. Frequent changes in the behavior of this disease generates an enormous volume of data. Microarray data satisfies both the veracity and velocity properties of big data, as it keeps changing with time. Therefore, the analysis of microarray datasets in a small amount of time is essential. They often contain a large amount of expression, but only a fraction of it comprises genes that are significantly expressed. The precise identification of genes of interest that are responsible for causing cancer are imperative in microarray data analysis. Most existing schemes employ a two-phase process such as feature selection/extraction followed by classification. In this paper, various statistical methods (tests) based on MapReduce are proposed for selecting relevant features. After feature selection, a MapReduce-based K-nearest neighbor (mrKNN) classifier is also employed to classify microarray data. These algorithms are successfully implemented in a Hadoop framework. A comparative analysis is done on these MapReduce-based models using microarray datasets of various dimensions. From the obtained results, it is observed that these models consume much less execution time than conventional models in processing big data.

1. Introduction

Microarray-based gene expression profiling has emerged as an efficient technique for cancer prognosis, diagnosis, and treatment purposes [1]. In recent years, the DNA microarray technique has had a great impact in determining *informative genes* that cause cancer [2,3]. The major drawback that exists in microarray data analysis is the curse of dimensionality problem, which hinders usefulness of information from a dataset and leads to computational instability [4]. Therefore, the selection/extraction of relevant features (genes) remains an imperative task in analyzing cancer microarray datasets, which is a critical step towards effective classification.

In literature, many feature (gene) selection/extraction techniques have been proposed by various researchers and practitioners [5]. Meanwhile, recent developments in microarray chip technology have helped in assaying thousands/millions of genes simultaneously, generating a huge amount of data. However, it is difficult to process the data on a conventional system (data are stored on a standalone machine) with standard computational power. The MapReduce programming model and its implementation in the Hadoop framework provides substantial foundation for processing large datasets, in particular for high-dimensional genomic data such as microarray data, in a distributed manner [6]. Apache Hadoop, developed by Doug Cutting in 2008, is open-source software that provides an effective way of storing and processing big data in a distributed fashion on large-scale clusters of commodity hardware. It follows a master/slave architecture for both distributed storage and distributed computation, thus accomplishing two tasks, i.e., massive data storage and faster processing [7].

There are various machine learning methods explored in the area of bioinformatics (typically, microarray data) [8–12]. These methods consume substantial amounts of time to analyze and explore large datasets on a conventional system with standard computational abilities. To counter this problem, the concept of distributed computing has been adopted, wherein the data is distributed over various nodes in a cluster, and various parallel processing paradigms are used.

In recent years, big data applications have increasingly become the focus of attention because of the enormous increase in data generation and storage that has taken place. Extracting information from the data becomes a challenge because current data mining techniques are not adapted to the new space and time requirements. To overcome these challenges, many paradigms like

* Corresponding author.
E-mail addresses: mkyadav262@gmail.com (M. Kumar), nitish.rath@gmail.com (N.K. Rath), skrath@nitrkl.ac.in (S.K. Rath).

http://dx.doi.org/10.1016/j.jbi.2016.03.002
1532-0464/

MapReduce and MPI have been considered for developing scalable algorithms. Researchers have accomplished a significant array of relevant, intelligent techniques in data science development; these approaches deal with imprecision, uncertainty, learning, and evolution in posing and solving computational problems [13].

Qian et al. [14] proposed a parallel and hierarchical attribute reduction method that can be applied to big data to analyze the intended data more efficiently. This model is able to mine decision rules under different levels of granularity. The proposed algorithms are implemented in the Hadoop framework using MapReduce, which can distribute the data and tasks in parallel and efficiently deal with big data. Li et al. [15] developed a dominance-based rough sets approach, which is an extension of classical rough sets theory, for selecting relevant features efficiently. It processes information within the preference-ordered attribute domain and then uses it for multi-criteria decision analysis. Ayadi et al. [16] used the concept of biclusters, using DNA gene expression for microarray data. Initially, they considered a new tree structure, called the modified bicluster enumeration tree (MBET), in which biclusters are represented by the profile shapes of genes. In the next phase, they proposed an algorithm called BiMine+, which uses a pruning rule to avoid both trivial biclusters and the combinatorial explosion of the search tree. The performance of BiMine+ was assessed on both synthetic and real DNA microarray datasets. Triguero et al. [17] described the MROSEFW-RF algorithm based on a MapReduce parallelization strategy. This algorithm ensembles several highly scalable re-processing and mining methods. It performs the balancing of class distribution, detects cost-relevant features, and builds an appropriate random forest model. Islam et al. [18] proposed a MapReduce-based parallel gene selection method, that utilizes sampling techniques to reduce irrelevant genes by using the ratio of between-group to within-group sums of squares (BW). The BW ratio indicates the variances among gene expression values. After gene selection, it applies the MRkNN technique to execute multiple kNN in parallel using the MapReduce programming model. Finally, the effectiveness of the method was verified through extensive experiments using several real and synthetic datasets. Wang et al. [19] proposed a new method for calculating correlation and introduced an efficient algorithm based on MapReduce to optimize storage and correlation calculations. This algorithm is used as a basis for optimizing high-throughput molecular data (microarray data) correlation calculations. He et al. [20] described a parallel implementation of several classification algorithms (e.g., K-nearest neighbor, naive Bayesian models, and decision trees), which were executed concurrently on various clusters using the iris dataset.

Parametric and non-parametric statistical tests are elegant procedures to analyze the behavior of data [21]. The statistical tests are used as a feature selection method by assuming a hypothesis, i.e., the null hypothesis and an alternate hypothesis. Based on the correctness of the hypothesis, features are either selected or rejected. The K-nearest-neighbor classifier provides a simple non-parametric procedure for the assignment of a class label to the input pattern, which is based on the class labels represented by the K-nearest training samples [22]. The major contributions of this paper are:

- To deal with high-dimensional data, various feature selection methods based on statistical tests are proposed. Statistical tests like ANOVA, Kruskal–Wallis, and Friedman tests are implemented in the Hadoop framework using the MapReduce paradigm, which processes the data in a distributed manner.
- After precise identification of the features, MapReduce-based K-NN (mrKNN) is proposed to classify the datasets. The proposed mrKNN classifier is completely different from the mrKNN algorithm proposed by Islam et al. [18]. They executed multiple kNN models concurrently on each data node using the MapReduce paradigm. In contrast, in the proposed classifier, the Euclidean distance between the training set and testing set has been calculated concurrently on each data node using mappers, and then the partial results from the nodes are collected in the reducer phase. The final decision of the testing samples is made based on the results obtained from the reducer phase.

These algorithms are implemented in the Hadoop framework to process and obtain results for various microarray datasets. The performance of the algorithms are tested on a Hadoop cluster with four slave (data) nodes and a conventional system.

The rest of the paper is organized as follows: Section 2 presents the proposed work for selecting features and classifying the microarray data using statistical tests and K-NN based on the MapReduce programming paradigm. Section 3 highlights the basic concepts of Hadoop and its components. Section 4 presents the implementation details for the proposed approach. Section 5 highlights the results obtained and the interpretations drawn from them, and also presents the comparative analysis for gene classification of microarray data. Section 6 concludes the paper and highlights the scope for future work.

2. Proposed work

The presence of a large number of insignificant and irrelevant features degrades the quality of the analyses of diseases like cancer. To counter this, it is essential to analyze the dataset from the proper perspective. This section presents an approach for classifying microarray data, which consists of two phases:

i. The input data are preprocessed using methods such as missing data imputation, normalization, and feature selection using statistical tests based on the MapReduce programming model.
ii. After selection of relevant features, MapReduce-based K-NN (mrKNN) is applied to classify a microarray dataset into cancerous/non-cancerous samples.

Fig. 1 shows the graphical representation of the proposed approach. A brief description is discussed as follows:

a. *Data collection*
The dataset for classification, which acts as requisite input for the models, is obtained from the National Center of Biotechnology Information (NCBI GEO, http://www.ncbi.nlm.nih.gov/gds/).
b. *Missing data imputation and normalization of datasets*
Missing data for a feature (gene) in a microarray dataset are imputed by using the *mean* value of the respective feature. Input feature values are normalized over the range $[0, 1]$ using the min–max normalization technique [23].
c. *Division of dataset*
The dataset is divided into two categories: a training set and testing set (Section 5).
d. *Feature selection*
MapReduce-based statistical tests such as ANOVA, Kruskal–Wallis, and Friedman tests are applied to select features with high relevance values, and thus the curse of dimensionality issue is addressed.
e. *Design of a classifier*
MapReduce-based K-NN (mrKNN) is built to classify the microarray dataset. The training of mrKNN is done using stratified 3-fold cross validation (CV), and the parameter K is obtained.