Journal of Parallel and Distributed Computing 138 (2020) 78–98

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

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

A Parallel Multilevel Feature Selection algorithm for improved cancer classification ☆

Lokeswari Venkataramana [a,*], Shomona Gracia Jacob [b], Rajavel Ramadoss [c]

[a] Sri Sivasubramaniya Nadar College of Engineering, Department of CSE, Chennai, 603110, India
[b] Ibri, Muscat, Oman
[c] Sri Sivasubramaniya Nadar College of Engineering, Department of ECE, Chennai, India

ARTICLE INFO

Article history:
Received 30 April 2019
Received in revised form 20 September 2019
Accepted 21 December 2019
Available online 28 December 2019

Keywords:
Parallel Multilevel Feature Selection
Parallel Random Forest
Horizontal & Vertical partition
Oncogenes and proteins
Classification accuracy

ABSTRACT

Biological data is prone to grow exponentially, which consumes more resources, time and manpower. Parallelization of algorithms could reduce overall execution time. There are two main challenges in parallelizing computational methods. (1) Biological data is multi-dimensional in nature. (2). Parallel algorithms reduce execution time, but with the penalty of reduced prediction accuracy. This research paper targets these two issues and proposes the following approaches. (1) Vertical partitioning of data along feature space and horizontal partitioning along samples in order to ease the task of data parallelism. (2) Parallel Multilevel Feature Selection (M-FS) algorithm to select optimal and important features for improved classification of cancer sub-types. The selected features are evaluated using parallel Random Forest on Spark, compared with previously reported results and also with the results of sequential execution of same algorithms. The proposed parallel M-FS algorithm was compared with existing parallel feature selection algorithms in terms of accuracy and execution time. The results reveal that parallel multilevel feature selection algorithm improved cancer classification resulting into prediction accuracy ranging from ∼85% to ∼99% with very high speed up in terms of seconds. On the other hand, existing sequential algorithms yielded prediction accuracy of ∼65% to ∼99% with execution time of more than 24 hours.

1. Introduction

CANCER is a life threatening disease that affects life at all ages. Early prediction of oncogenes and onco-proteins will help oncologists and pharmaceutical scientists to design drugs that would target these cancer causing genes and protein mutants for their therapeutic function. Oncogenes could be obtained from Microarray Gene Expression (MGE) data and oncoproteins from Tumor Protein 53 (TP53) mutant data and Amino acid changes in TP53. Micro-array data has lot of noisy or irrelevant genes and missing data [1,50]. Microarray gene data is high-dimensional (thousands of genes) and low-sample data set [19,42]. This affects accuracy of predicting a disease. It is very much necessary to weed out the irrelevant genes and select only important genes for improving class prediction [16]. This necessitates feature selection before classification and choosing appropriate techniques that may lead to good classification performance for a high-dimensional dataset [9].

Clinical and biological data grows day by day and analyzing this data would incur huge resources, time and manpower. Parallel Algorithms play an important role in exploiting multiple machines in a cluster to process huge data in parallel and produce results in less time. It was ascertained from the review of previous work that most algorithms that have been proposed in the past were unable to execute on larger data sets. Even if they did scale to handle large data, it was at the cost of prediction accuracy and prolonged execution time. As data is split and gets stored in different storage nodes in a cluster and each node performs task independently, this could affect the accuracy of results produced by individual nodes compared to processing done at single machine. The rationale behind reduced prediction accuracy was due to movement of data between nodes in a cluster, interdependency between results produced by nodes in a cluster, synchronization and communication overhead among nodes in a cluster. This is a serious issue since biological data is prone to grow in multiples, serial execution of algorithm could not scale for this growing data and even though the existing parallel algorithms could scale for massive amount of data, they do not offer prediction accuracy.

The well-known parallel programming frameworks are Apache Hadoop & Spark [56]. Hadoop is a parallel programming framework that is used to process very large data sets.

☆ The co-authors have also contributed to the paper.

* Corresponding author.
E-mail addresses: lokeswaricts@gmail.com, LOKESWARIYV@SSN.EDU.IN (L. Venkataramana), graciarun@gmail.com (S.G. Jacob), rajavelr@ssn.edu.in (R. Ramadoss).

https://doi.org/10.1016/j.jpdc.2019.12.015
0743-7315/