COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 131 (2016) 191–206

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A MapReduce approach to diminish imbalance parameters for big deoxyribonucleic acid dataset

Sarwar Kamal [a], *Shamim Hasnat Ripon* [a], *Nilanjan Dey* [b], *Amira S. Ashour* [c,*], *V. Santhi* [d]

[a] *Computer Science and Engineering, East West University, Dhaka, Bangladesh*
[b] *Techno India Institute of Technology, Kolkata, India*
[c] *Department of Electronics and Electrical Communications Engineering, Faculty of Engineering, Tanta University, Tanta, Egypt*
[d] *School of Computing Science and Engineering, VIT University, Vellore, Tamil Nadu, India*

ARTICLE INFO

Article history:
Received 9 December 2015
Received in revised form
18 March 2016
Accepted 6 April 2016

Keywords:
MapReduce
K-nearest neighbor
Big data
DNA (deoxyribonucleic acid)
Computational biology
Imbalance data

ABSTRACT

Background: In the age of information superhighway, big data play a significant role in information processing, extractions, retrieving and management. In computational biology, the continuous challenge is to manage the biological data. Data mining techniques are sometimes imperfect for new space and time requirements. Thus, it is critical to process massive amounts of data to retrieve knowledge. The existing software and automated tools to handle big data sets are not sufficient. As a result, an expandable mining technique that enfolds the large storage and processing capability of distributed or parallel processing platforms is essential.
Method: In this analysis, a contemporary distributed clustering methodology for imbalance data reduction using k-nearest neighbor (K-NN) classification approach has been introduced. The pivotal objective of this work is to illustrate real training data sets with reduced amount of elements or instances. These reduced amounts of data sets will ensure faster data classification and standard storage management with less sensitivity. However, general data reduction methods cannot manage very big data sets. To minimize these difficulties, a MapReduce-oriented framework is designed using various clusters of automated contents, comprising multiple algorithmic approaches.
Results: To test the proposed approach, a real DNA (deoxyribonucleic acid) dataset that consists of 90 million pairs has been used. The proposed model reduces the imbalance data sets from large-scale data sets without loss of its accuracy.
Conclusions: The obtained results depict that MapReduce based K-NN classifier provided accurate results for big data of DNA.

1. Introduction

Biological data are complex and accessible in various formats. The key problem of the current era is to discover knowledge from these complex/large biological databases. DNA is the central building block of any living organism. The information accumulated in the DNA is used to construct RNA (ribonucleic acid), which is considered more transient. There are sets of exons and introns that exist in the biological dataset.

* *Corresponding author.* Department of Electronics and Electrical Communications Engineering, Faculty of Engineering, Tanta University, Tanta, Egypt. Tel.: +966591859749.
E-mail address: amirasashour@yahoo.com (A.S. Ashour).

http://dx.doi.org/10.1016/j.cmpb.2016.04.005
0169-2607/

These exons play a significant role in molecular interactions and processes. Also, the exons are modified into several parts as mature messenger RNA (mRNA) by transcriptions method and this is accomplished in ribosomes [1]. Similarly, introns are also considered an important factor of eukaryotic genomes segments. These introns are not directly modified into proteins. They are available in human cells, yeast and bacteria.

Typically, human genomes are constituted with introns and exons sequences. Researchers observed that 26,536 annotated genes contain 233,785 exons and 207,785 introns, where each gene contains 8.8 exons and 7.8 introns. The length of exons and introns are also varied, approximately 80% of the exons have length less than 200 bp, 0.01% of the total introns have less than 20 bp and 10% introns have more than 11,000 bp [2]. In genome biology, large volume of genomic data sets is considered a vital framework to determine ration of exons and introns [3]. Swift developments of sequenced eukaryotes permit exons–introns data into various classifications such as the phylogenetic tree analysis [4].

Introns classifications have supported to reconstruct the taxonomic relationship among gene lineages as well as subgroups gene families [5,6]. Thus, the classification of exons is essential for physical protein–protein interaction (PPI) in living organs [7,8]. Meanwhile, imbalance data indicate the variations of exons and introns present in the DNA and RNA (ribonucleic acid). Typically, imbalanced data sets are considered a special case for the classification problems, where the class distribution is not uniform among the classes, i.e. the classification contains unequal distribution of data sets. Consequently, the data classes are unequally represented if the training dataset is imbalanced.

Nowadays, machine learning techniques help to categorize imbalance datasets. Moreover, testing data need to be differentiated from training data. There are numerous data preprocessing techniques that exist to process imbalance dataset, such as the SMOTE (Synthetic Minority Over-sampling Technique) [9], re-sampling techniques [10], data distributions [11] and Iterative Performance Filter (IPE-SMOTE) [12]. Nonetheless, the illustration and knowledge mining from large biological datasets become very difficult for most of the existing data mining tools and techniques [13,14]. Data mining processes should be accommodated toward materialized engineering [15] to exceed their limitations.

In the age of digitalization, most of the large datasets are imbalance, and there are overlapping sets. In this regard, unequal labels are used in datasets during classification [16]. The DNA sequences are highly imbalance due to its huge size and variations in length. A mathematical approach named Imbalance Ratio (IR) is imposed to investigate the dimensions of imbalance and noises in collected datasets [17,18]. Generally, some techniques such as re-arrangements, cost-effective learning and algorithmic approaches are used [10]. Since DNA sequences are multi-label datasets (MLDs), there are less coherency among datasets [19]. Recently, some algorithmic processes have initiated to solve the factors [20,21]. Frenay and Verleysen [22] conducted a survey on label noise with the introduction of facts regarding misclassification. Kuncheva and Rodríguez [23] used weight base probabilistic framework to capture the voting system. An imbalance dataset disturbed both performance and robustness during the development of classifiers. Equalized Loss of Accuracy (ELA) addresses the problems for noisy datasets on imbalance impact. This approach achieves a balance between robustness as well as performances. However, this balance hampers on the overall performances as well as accuracy [24]. Recently, the use of Fuzzy Unordered Rule Induction Algorithm (FURIA) assisted the diagnosis of the dyslexia symptoms using vague (imbalance) datasets. However, fuzziness is not always perfect while qualitative datasets are captured [25]. Discretization is a new form of data representation in data management. This approach provided better orientation of datasets in machine learning, where the primary goal of discretization is to simplify the representation of datasets along with discrete value. Some popular techniques such as Bayesian network or rule learning approach support this analysis for better outcomes. However, discretization sometimes missed some values and losses the information [26]. There are sets of instances in multi-instance data classifications. During this analysis, some objects are properly labeled and some are not. The un-labeled datasets create problems for proper data management. Fuzzy rough set based operations assist in labeling the un-labeled datasets. However, this process limits two levels of classifications such as bag-level and instance support level [27]. Other classification factors should also be considered. Multi-mode operations were absent in this process. Group datasets are classified by ordering centric pruning metrics for large imbalance datasets. This classification improves overall impact of the system. Nevertheless, the grouping creates complexity for the classification process [28].

To address the previously mentioned drawback, the MapReduce scheme [29,30] with distributed files system [31] has been proposed. MapReduce technique generally supports parallel data processing. Data shrinkage methods [32] turn up as data pre-processing algorithms that aim to abridge and brighten the initially collected data as raw data. These algorithms enable faster and efficient ways to reduce noisy, complex, unequal and redundant datasets. There are some popular data reduction techniques in different environments [33,34]. However, these algorithms suffer when the data size increases in regular time intervals.

To address the unequal classification and clustering problem arising from Bioinformatics, K-Nearest Neighbor (K-NN) is evaluated on collected DNA datasets [35,36]. Big data volumes are categorized into sub-groups to identify the new sample points of exons and introns. The Bioinformatics database of human DNA sequence could be analyzed to extract its characteristics, such as the number of exons–introns, protein–protein interactions, DNA breaks, DNA consensus and RNA orientations. These characteristics are measured numerically using algorithmic approaches. However, some processes are quite expensive and costly as DNA sequence contains gene coding (exons) and non-coding (introns) segments in a repeated manner. For example, the collected long DNA sequences of base pairs in symbolic representations as $\{D_1, D_2, D_3, \ldots\}$ reflects sequences in computer memory. The prevalent problem is the dataset processing into certain areas of limited resources. The DNA datasets are symbolically represented as set D. The suffix values of D indicate the segments as well as its positions. So, suffix number should be properly addressed by $\{D_1, D_2, D_3, \ldots\}$.