Available online at www.sciencedirect.com

ScienceDirect

Karbala International Journal of Modern Science 1 (2015) 86—96
http://www.journals.elsevier.com/karbala-international-journal-of-modern-science/

SVM-BT-RFE: An improved gene selection framework using Bayesian T-test embedded in support vector machine (recursive feature elimination) algorithm

Shruti Mishra*, Debahuti Mishra*

Siksha 'O' Anusandhan University, Bhubaneswar 751030, Odisha, India

Received 21 September 2015; revised 19 October 2015; accepted 23 October 2015
Available online 28 November 2015

Abstract

Gene Regulatory Network (GRN) has always gained considerable attention from bioinformaticians and system biologists in understanding the biological process. But the foremost difficulty relics to appropriately select a stuff for its expression. An elementary requirement stage in the framework is mining relevant and informative genes to achieve distinguishable biological facts. In an endeavor to discover these genes in several datasets, we have suggested a strategic gene selection algorithm called Support Vector Machine Bayesian T-Test Recursive Feature Elimination algorithm (SVM-BT-RFE), which is an extended variation of support vector machine recursive feature elimination (SVM-RFE) algorithm and support vector machine t-test recursive feature elimination (SVM-T-RFE). Our algorithm accomplishes the goal of attaining maximum classification accuracy with smaller subsets of gene sets of high dimensional data. Each dataset is said to contain approximately 5000—40,000 genes out of which a subset of genes can be selected that delivers the highest level of classification accuracy. The proposed SVM-BT-RFE algorithm was also compared to the existing SVM-T-RFE and SVM-RFE where it was found that the proposed algorithm outshined than the latter. The proposed SVM-BT-RFE technique have provided an improvement of approximately 25% as compared to the existing SVM-T-RFE and more than 40% of improvement as compared to the existing SVM-RFE. The comparison was performed with regard to the classification accuracy based on the number of genes selected and classification error rate of 5 runs of the algorithm.

Keywords: Gene regulatory network; Gene selection; SVM-RFE; SVM-T-RFE

1. Introduction

The complex patterns of gene expression are engendered in response to specific cellular activities at different levels. One of the most challenging objectives of systems biology is to provide qualitative and quantitative models for reviewing the intricate patterns of gene interaction [1,2]. Amongst all the models, gene

* Corresponding authors.
E-mail addresses: shruti_m2129@yahoo.co.in (S. Mishra), mishradebahuti@gmail.com (D. Mishra).

Peer review under responsibility of University of Kerbala.

http://dx.doi.org/10.1016/j.kijoms.2015.10.002

regulatory networks (GRNs) [3] are most crucial. Construction of GRN is the process of identification of genes that interact in a gene–gene interaction network. This helps researchers define the diverse biologic functions and undercurrents of molecular activities taking place in a human body. However, identifying GRNs cannot be accurate because of high density of the network, deficiency of information about biological organism and uproar in the expression measurement [4,5].

Conventionally, gene selection [6–9] is considered as a vital technique with microarray data. The DNA microarray technology has provided us several prospects to detect gene expression levels for many thousands of genes simultaneously. The problem is to select a small subsection of genes from a huge pattern of expressions. The challenging task is to choose relevant genes that are highly correlated to classify because of small sample size of the expression data. Gene selection methods are usually categorized into four different approaches: filter [10], wrapper [11], embedded [12], and hybrid [13]. Filter methods [14] evaluate based on the individualities of the data and relation of each gene with the class label. It usually considers statistical properties of the data without any learning model. The wrapper methods [15] are quite popular in machine learning tasks and applications. This method evaluates the fitness of subset of selected genes iteratively by a specific learning classifier model in the genetic selection process. In the embedded method [16], using a preliminary gene set, a learning classifier model is skilled to establish a criterion to measure the rank values of genes. The hybrid approach [17] takes the benefits of the filter and the wrapper approaches. In the hybrid approach at a first subset of genes is chosen based on the filter approach and then the wrapper approach is active to select the final gene set.

Various computational approaches have been proposed to select genes based on the information they provide, simplicity and computational efficiency. Diaz-Uriarte and de Andres [18] discussed a method for gene selection and data classification based on a random forest where the method yields small sets of genes that provides high classification accuracy. Shreem et al. [19] proposed an approach that embeds the Markov Blanket with the harmony search algorithm for gene selection. Cai et al. [20] too proposed a feature weighting algorithm for gene selection called LHR. LHR estimates the feature weights through local approximation based on Relief F. Han et al. [21] proposed and suggested the gene-to-class sensitivity subjugated by a single hidden layered feedforward neural network in a hybrid gene selection. They used *k*-means clustering and binary particle swarm optimization for filtering irrelevant genes.

Similarly, Guyon et al. [22] proposed a feature elimination technique using support vector machines [23–27] known as *support vector machine recursive feature elimination* (SVM-RFE). In this algorithm, the genes are removed recursively based on the SVM classifier weights and later classifies the samples with SVM. Studies based on SVM-RFE approach have drawn a huge curiosity among the researchers for selection of relevant genes. But the major problem with this algorithm is that, it consumes a huge amount of training time, the problem of over-fitting persist and at each iteration it eliminates only one gene at a time. Li et al. [28] proposed SVM-T-RFE, a gene selection algorithm that extended the SVM-RFE algorithm by incorporating the Welch's t-test. This method combined the statistical Welch's t-test to predict higher accuracy and more significant genes. But this method too has a problem. Unlike the previous algorithm, this algorithm is entirely dependent on the threshold value which is somewhat in the range of 0–1 with a variation of 0.01 at each step. Hence, the algorithm need to iterate till the set {0, 0.01, 0.02 … 1} is covered. This leads to a time consuming process as the algorithm has to execute for the above threshold set. In our study, we have proposed a technique known as *Support Vector Machine Bayesian T-Test Recursive Feature Elimination* algorithm (SVM-BT-RFE) based on the SVM-RFE and a statistical test (Bayesian t-test).

For our experiment, we have considered five datasets i.e. colon dataset [29], Leukemia dataset [30], medulloblastoma dataset [31], Lymphoma dataset [32] and prostate cancer dataset [33]. The nature of the dataset is quite large enough in terms of the number of genes, but have a small sample size. In this work, the SVM-RFE algorithm is merged with Bayesian T-test for selecting genes from the high dimensional datasets. The ranking criteria that is considered as the vital parameter in this algorithm is redefined with the help of the Bayesian T-test. Our approach helps us to select a smaller subset of genes as compared to the generalized SVM-RFE. Amongst SVM-BT-RFE, SVM-T-RFE and SVM-RFE, the proposed approach takes slight an extra time for its two parameters value calculation but computationally the number of genes selected is much appropriate and less as compared to the remaining of the two algorithms.

This article is structured as follows: firstly, the section depicts the materials and methods that have