Informatics in Medicine Unlocked 5 (2016) 39–51

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Enhanced gene ranking approaches using modified trace ratio algorithm for gene expression data

Shruti Mishra*, Debahuti Mishra*

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

ARTICLE INFO

Keywords:
Gene regulatory network
Gene selection
Information gain
Trace ratio
Canonical correlation analysis
Classification

ABSTRACT

Microarray technology enables the understanding and investigation of gene expression levels by analyzing high dimensional datasets that contain few samples. Over time, microarray expression data have been collected for studying the underlying biological mechanisms of disease. One such application for understanding the mechanism is by constructing a gene regulatory network (GRN). One of the foremost key criteria for GRN discovery is gene selection. Choosing a generous set of genes for the structure of the network is highly desirable. For this role, two suitable methods were proposed for selection of appropriate genes. The first approach comprises a gene selection method called *Information gain,* where the dataset is reformed and fused with another distinct algorithm called *Trace Ratio* (TR). Our second method is the implementation of our projected modified TR algorithm, where the scoring base for finding weight matrices has been re-designed. Both the methods' efficiency was shown with different classifiers that include variants of the Artificial Neural Network classifier, such as Resilient Propagation, Quick Propagation, Back Propagation, Manhattan Propagation and Radial Basis Function Neural Network and also the Support Vector Machine (SVM) classifier. In the study, it was confirmed that both of the proposed methods worked well and offered high accuracy with a lesser number of iterations as compared to the original Trace Ratio algorithm.

1. Introduction

Genes, as good as their products (proteins) are the essential construct blocks of animation that do not function autonomously. Rather for a cell to function appropriately, they act together with each other and form an intricate network [1]. One such application to understand the behavior of the genes and their expression levels is to construct a gene network that signifies the relationship between sets of genes which harmonize to achieve different tasks. For the understanding of the core biological process and its molecular system, Gene Regulatory Network (GRN) [2] plays a crucial part. However, modeling of these networks is a significant challenge that needs to addressed.

Apart from this, understanding the construction and functionalities of GRN is a basic problem in biology. With the accessibility of gene expression data and whole genome sequences, several computational approaches have been developed to discover, their regulatory network by enabling the recognition of their regulatory state component [3]. In the current era, formation of precise GRN models [4] is reaching a major percentage of importance in biomedical research. The gene expression of the microarray data monitors the behavior of thousands of genes simultaneously that provides a maximum chance to look into large scale regulatory networks. Lastly, an absolute GRN model allows us to incorporate experimental facts about the elements and interactions of the factors which leads to knowing the final state or the dynamical behavior of the network.

Gene selection [5,6] acts as a major criterion. Gene selection from microarray data (which is a high dimensional dataset) is statistically difficult problem. Usually, the number of samples is quite less as compared to thousands of genes whose expression levels are measured. Hence, it is important to restrain down to few disease related genes from thousands of microarray genes by the operation of selection or ranking. There are many gene selection or feature selection methods [7,8] that deal with the problem of curse of dimensionality in microarray data. Apart from this, it also helps to reduce the time and memory complexities which always create issues. Generally, gene selection or feature selection methods are split into two categories: classifier independent and classifier dependent. Filter methods [9] are believed to be a classifier dependent as the choice is based on some heuristic criterion and score, whereas wrapper and embedded methods are thought to be a part of the classifier dependent method. Wrapper method [10] assesses a subset of variables according to their efficacy to a given predictor whereas in embedded methods, a variable selection is

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

http://dx.doi.org/10.1016/j.imu.2016.09.005
Received 10 May 2016; Received in revised form 25 September 2016; Accepted 26 September 2016
Available online 30 September 2016

performed as a part of the learning practice and are usually precise to a given learning machine. Other than gene selection, gene ranking is also an important factor of consideration for which different methods are available in the literature for study of class data. Some of them are Fold Change (FC), moderated t-statistics, Significance Analysis of Microarrays (SAMs) etc. There is another method called as RP method that is the only rank based non-paramteric method. This method independently handles up-regulated and down-regulated genes under one class and therefore produces two separate ranked gene lists.

Separate from these existing techniques, there are various computational techniques and methods for gene selection. Model et al. [11] established how phenotypic classes can be predicted by amalgamating feature selection methods and discriminant analysis for methylation pattern based discrimination between acute lymphobatic leukemia and acute myeloid leukemia. They used SVM to the methylation data for using every CpG position as a separate dimension. Li et al. [12] studied the problem of edifying the multi-class classifier for tissue classification based on gene expression datasets. They stated that for datasets with a small number of classes the results are good and for datasets with a large number of classes the accuracy is moderately less. Mundra and Rajapakse [13] used the famed *t*-statistics for gene ranking in the analysis of microarray data. Here, they have divided the *t*-statistics into two parts: relevant and irrelevant data points. A backward elimination based iterative approach was projected to rank genes using only the relevant sample points and *t*-statistics. It was found that the proposed method performed considerably better than the standard *t*-statistic approach. Kira et al. [14] partitioned the information points into clusters using k-d-tree and chose random data point from each cluster, and then performed feature selection by means of Relief which looks for frontier points to estimate feature weights. Pechenizkiy et al. [15] used the principal component analysis for dimensionality reduction after partitioning large datasets with *k-d*-tree. Cavill et al. projected a GA/*k*-NN based move for concurrent feature and sample selection from metabolic profiling data [16].

Similarly, Cawley et al. [17] proposed a straight forward Bayesian approach which gets rid of the regularization parameter fully, by integrating it out systematically using an uninformative Jeffrey's prior. The anticipated algorithm (BLogReg) uses two or three orders of magnitude faster than the original algorithm, as there is no longer a necessity for a model selection step. Two new dimensionality reduction techniques were proposed by Fitzgerald et al. [18]. These methods use the minimum and maximum information models. These are information theoretic extensions of Spike-Triggered Covariance (STC) with the intention that can be practiced with non-Gaussian stimulus distributions to locate relevant linear subspaces of random dimensionality. Piao et al. [19] projected an Ensemble Correlation-Based Gene Selection algorithm based on symmetrical indecision and Support Vector Machine. In the method, symmetrical indecision was used to analyze the importance of the genes and the diverse preparatory points of the pertinent subset were used to produce the gene subsets where Support Vector Machine was used as an assessment criterion of the wrap.

Nie *et al.* [20] proposed an optimized subset-level score and algorithm to proficiently discover the global optimal feature subset such that the subset-level score is maximized. This algorithm is called as Trace Ratio (TR) which uses the Fisher and Laplacian score as the evaluation criterion. It's essentially a graph based feature selection algorithm. Zhao et al. [21] introduced the trace ratio linear discriminant analysis (TR-LDA) algorithm for dememtia diagnosis. They also proposed the ITR algorithm (iITR) to resolve the TR-LDA problem. This process integrates with the sophisticated missing value imputation method and is used for the probe of the nonlinear datasets in many real-world medical diagnosis problems. Wang et al. [22] proposed a amalgamated objective to flawlessly hold trace ratio formulation and *k*-means clustering process in a manner that the trace ratio criterion is extended to unsupervised model. They also proposed an unsupervised feature selection method by integrating unsupervised trace ratio formulation and ordered sparsity-inducing norm regularization. This method was able to strap up the discriminant power of trace ratio criterion, and thus it tends to select discriminating features. The major disadvantage of using this trace ratio algorithm [23] is that though theoretically the algorithm converge and global optimum of the solution is achieved, but by extensive study it is found that sometimes the algorithm does not converge as the basic stopping criteria is not met. Hence, we do forcefully terminate the algorithm by providing some stopping criteria to it.

In our study, we have proposed two methods in which the trace ratio algorithm has been explored properly. In our first method, we have not altered any criteria of TR algorithm. Rather, we improvised and structured the dataset on the basis of information gain values. In our second method, we have modified the existing and original TR algorithm by changing the scoring criteria which is one of the fundamental steps in TR algorithm. Instead of using the Fisher's score, the canonical correlation analysis score is used to calculate the weight matrices within-class and between class. Canonical correlation score being a statistical technique aims at providing a better rank list when merged with the TR algorithm as compared to the existing Fisher's score. It is also relevant as it is expected to provide a far better classification accuracy rate when compared with the original TR algorithm. Both the proposed method is examined and evaluated on the basis five datasets i.e. Colon [24], Leukemia [25], Medulloblastoma [26], Lymphoma [27] and Prostate Cancer [28]. The nature of the dataset is quite large in terms of the number of genes, but have a small sample size. It was found that the information gain with the original TR algorithm and the modified TR algorithm provided promising results as compared to the unmodified TR algorithm.

The rest of the paper is divided as follows: the first section depicts the materials and methods that have been used for this work such as datasets used, the methods and the algorithm like information gain, TR algorithm, Canonical correlation analysis, Performance Metrics etc. The next section deals with the experimental evaluation where pre-processing of the data, parametric discussion and schema diagram of the proposed model are discussed. Following this section, the result of the proposed technique along with the original technique have been critically analyzed and summarized. Lastly, the conclusion of the work is briefed with some future direction.

2. Materials and methods

2.1. Datasets used

Expression profiling of colon cancer or colorectal adenomas and normal mucosas from 32 patients were downloaded from Gene Expression Omnibus [24] (SOFT Matrices files were download and for the same log transformation was used as the data were mostly skewed to the right). This set consists of 32 adenomas and 32 normal mucosas sample (64 samples) having 43,237 genes. To illustrate the molecular developments underlying the alteration of normal colonic epithelium, the transcriptomes of 32 prospectively collected adenomas were measured along with those of normal mucosa from the same entities. Similarly, the Leukemia dataset was collected from [25] where the dataset consist of 10,056 genes with 48 samples of both ALL and AML (24 ALL- Acute Lymphocytic Leukemia and 24 AML- Acute Myeloid Leukemia each). Apart from these two, few more datasets were taken into consideration like the Medulloblastoma dataset [26] having 5893 genes with 34 samples of 25 C and 9 D samples (Medulloblastoma have four molecular sub types out of which two less well defined sub types are group C and group D), Lymphoma dataset [27] having 7070 genes having 77 samples of 58 DLBCL (Diffuse Large B-cell Lymphoma) and 19 FL (Follicular Lymphoma) samples (Affymetrix HuGeneFL array), and the prostate cancer dataset [28] having 12,533 genes with 102 samples of 50 normal and 52 tumor