Available online at www.sciencedirect.com
ScienceDirect
journal homepage: www.elsevier.com/locate/bbe

Original Research Article

# A hybrid gene selection method for microarray recognition

Alok Kumar Shukla *, Pradeep Singh, Manu Vardhan

*Department of Computer Science & Engineering, NIT, Raipur, Chhattisgarh 492010, India*

ARTICLE INFO

*Article history:*
Received 27 April 2018
Received in revised form
29 July 2018
Accepted 14 August 2018
Available online xxx

*Keywords:*
Accuracy
Ensemble
Adaptive genetic algorithm
Gene selection
Support vector machine

ABSTRACT

DNA microarray data is expected to be a great help in the development of efficient diagnosis and tumor classification. However, due to the small number of instances compared to a large number of genes, many of the computational learning methods encounter difficulties to select the low subgroups. In order to select significant genes from the high dimensional data for tumor classification, nowadays, several researchers are exploring microarray data using various gene selection methods. However, there is no agreement between existing gene selection techniques that produce the relevant gene subsets by which it improves the classification accuracy. This motivates us to invent a new hybrid gene selection method which helps to eliminate the misleading genes and classify a disease correctly in less computational time. The proposed method composes of two-stage, in the first stage, EGS method using multi-layer approach and f-score approach is applied to filter the noisy and redundant genes from the dataset. In the second stage, adaptive genetic algorithm (AGA) work as a wrapper to identify significant genes subsets from the reduced datasets produced by EGS that can contribute to detect cancer or tumor. AGA algorithm uses the support vector machine (SVM) and Naïve Bayes (NB) classifier as a fitness function to select the highly discriminating genes and to maximize the classification accuracy. The experimental results show that the proposed framework provides additional support to a significant reduction of cardinality and outperforms the state-of-art gene selection methods regarding accuracy and an optimal number of genes.



## 1. Introduction

Biological data, like microarray, contain many irrelevant and redundant genes. Nowadays, DNA microarray has gained considerable attention due to its ability to measure the expression levels of hundreds or thousands of genes in a single experiment [1]. Finally, it produces gene expression data that contain valuable statistics on genomics, and prognosis for researchers [2]. Therefore, there is a need to identify the essential genes that contribute to predict the state of cancers [3]. At the abstract level, the main problem of this experiment

* *Corresponding author at*: Department of Computer Science & Engineering, NIT, Raipur, Chhattisgarh 492010, India.
E-mail address: akshukla.phd2015.cs@nitrr.ac.in (A.K. Shukla).

https://doi.org/10.1016/j.bbe.2018.08.004



Please cite this article in press as: Shukla AK, et al. A hybrid gene selection method for microarray recognition. Biocybern Biomed Eng (2018), https://doi.org/10.1016/j.bbe.2018.08.004

is thousands number of genes compared to the small number of instances, misleading genes, and noisy data [4]. To resolve this limitation, researchers have been used prominent gene selection method to select a subset of relevant genes that maximize the capability of the classifier to classify instances more accurately. The instances can be unlike from many points of view like genotype, phenotype or other relevant biological or clinical record, respectively. In the field of bioinformatics, feature selection is also known as gene selection [5].

The gene selection (GS) method tries to select the critical genes among all the features available in the data, which are useful for the application of learning algorithms. Furthermore, it is an essential application of data reduction to avoid challenges, such as over-processing, the high cost of the calculation, and the low interpretability of the final model [6]. Furthermore, the main problem in high-dimensional data sets is the "*curse of dimensionality*." To solve this problem, researchers have introduced a large number of gene selection methods, many of them derived from the need to analyze microarray data to select the best discriminating gene, called as "*biomarker*" [7]. In previous studies, gene selection methods are widely used before data classification in bioinformatics domains [8]. The gene selection approaches are efficient and straightforward, and has several merits:

1. It can retain or enhance the classification performance.
2. It can diminish the data dimensionality.
3. It can discard meaningless and irrelevant genes.
4. It can reduce the computational complexity while performing experiments.

Generally, gene selection techniques are categorized into three important phases: filter [9], wrapper [10], and hybrid [11]. In the first phase, the filter approach is the first technique to select subsets of genes before applying the inductive algorithm. On the other hand, the wrapper method [12] uses inductive learning as a fitness function and looks for the optimal subset of genes in the space of all characteristics. The irrelevant and noisy genes are produced by the high computational effort and also decrease the classification performance. Meanwhile, these genes increase the dimensionality of datasets thereby classifier gives the bias results. Besides filter and wrapper approach, the hybrid method takes the advantages of individual approaches such as filter and wrapper. Over the few decades, several hybrid approaches have been developed primarily, an integration of filter and wrapper method to select the useful genes for accurate diagnosis [13,14]. It takes advantage of both approaches by integrating the complementary strengths [15]. In general, the gene selection method uses a fitness function with a search strategy to attain the subsets of characteristics. The fitness function tries to measure the ability to discriminate a gene and to classify the different labels. In the current literature, there are five measures available, i.e., distance, information or uncertainty, dependence, consistency and correlation [16].

A large number of gene selection methods are available on high dimensional datasets for a tumor or cancer classification, over the past decades [17]. There are, in general, two methods for gene selection is used such as filter and wrapper. The filter methods have been incredibly increased since it can be used to define various relationships between pairs of features to select informative genes from gene classification datasets [18]. For instance, a gene decided in an earlier step cannot be reconsidered in later phases; however, a feature in the gene subset referred as relevant may lose its importance when the gene subset is updated with new feature sets. To address this limitation, researchers have used the powerful nature-inspired optimization technique namely GA [19] which was introduced by John Holland in early 1975, and another algorithm is PSO [20] used as wrapper approach. The improved version of the genetic algorithm is used in this study called adaptive genetic algorithm (AGA). The AGA can improve the growth of the solution quality by adjusting the values of the regulation parameters and controlling the premature convergence, and stagnation.

During the previous years, several machine learning (ML) approaches have been used an ensemble learning model for better prediction. It is a way toward building multiple logical models which convey us to solve the gene selection (or classification) problem with the help of machine learning approaches. Ensemble learning has affected classification problems, on the other hand, it can be used as gene selection for enhancing additional machine learning capacity [21]. The primary cause of inaccuracy in the machine learning is due to noise, bias, and variance. Ensemble strategy helps to minimize these factors. This method is designed to improve the stability and the accuracy of machine learning algorithms. The combinations of multiple gene selection methods may decrease variance, especially in the case of unstable techniques and produce a more reliable representation than an individual gene selection method.

The ensemble can be formed in many ways, but we are using ensemble framework in the form of multi-layer approach which is based on ranking of gene selection method such as "minimum redundancy-maximum relevance" (mRMR) [22], Relief-F [23], chi-square (CS) [24], joint mutual information (JMI) [25], and information gain (IG) [26]. A large number of techniques have widely applied to gene expression datasets for classification as shown in Table 1. Still, there is no agreement on which technology is better than others; particular gene selection method can accomplish better correlated with class than others in respect of specified dataset, while a further approach can outperform the others when dealing with different datasets. This uncertainty about which methods produce an optimal gene subset; is overcome in this research by introducing a new gene selection method that able to exploit the potential of existing gene selection methods.

Availability of statistics in bioinformatics fields is a critical problem toward selecting the relevant genes and overlook the extraneous genes that contribute to a carcinomatous stage. In this paper, we have developed a hybrid model, combination of ensemble gene selection (EGS) and AGA algorithm on biological datasets for identifying the informative features and reduces the computational cost of the learning algorithm. The proposed method uses EGS as a filtering approach to select highest ranked genes that will be passed to AGA algorithm. Furthermore, AGA is combined with SVM and NB to search for the most top-rated genes obtained from EGS genes to find the most revealing genes that will satisfy cancer classification accurately. This process continues until a sufficiently reached



Please cite this article in press as: Shukla AK, et al. A hybrid gene selection method for microarray recognition. Biocybern Biomed Eng (2018), https://doi.org/10.1016/j.bbe.2018.08.004