Biochemical and Biophysical Research Communications 446 (2014) 850–856

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

Robust and stable feature selection by integrating ranking methods and wrapper technique in genetic data classification

Maryam Yassi [a,*], Mohammad Hossein Moattar [b]

[a] Young Researchers and Elite Club, Mashhad Branch, Islamic Azad University,, Mashhad, Iran
[b] Department of Software Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran

ARTICLE INFO

Article history:
Received 19 February 2014
Available online 19 March 2014

Keywords:
Microarray classification
Dimension reduction
Filter method
Wrapper method
Support vector machine
Imbalance classes

ABSTRACT

High dimensional data increase the dimension of space and consequently the computational complexity and result in lower generalization. From these types of classification problems microarray data classification can be mentioned. Microarrays contain genetic and biological data which can be used to diagnose diseases including various types of cancers and tumors. Having intractable dimensions, dimension reduction process is necessary on these data. The main goal of this paper is to provide a method for dimension reduction and classification of genetic data sets. The proposed approach includes different stages. In the first stage, several feature ranking methods are fused for enhancing the robustness and stability of feature selection process. Wrapper method is combined with the proposed hybrid ranking method to embed the interaction between genes. Afterwards, the classification process is applied using support vector machine. Before feeding the data to the SVM classifier the problem of imbalance classes of data in the training phase should be overcame. The experimental results of the proposed approach on five microarray databases show that the robustness metric of the feature selection process is in the interval of [0.70, 0.88]. Also the classification accuracy is in the range of [91%, 96%].

1. Introduction

Applying microarray technology which makes studying the expression of thousands of genes possible simultaneously, has led to the production of massive amounts of gene expression data recently. Statistical analysis of these data includes feature selection, normalization, and classification.

The expression levels of genes have important information about biological networks, cellular states, and the understanding of gene function. An objective of gene expression data analysis is to determine how the expression of each individual gene affects the expression of other genes or genetic networks. Another goal is to determine how these genes are expressed in healthy and diseased cells. Using data mining techniques and artificial intelligence techniques for analyzing data obtained from this technology can be useful for diagnosis and treatment.

Collections of genetic data have high dimensions and small size and are usually imbalanced. Data collection and analysis and discovering unknown relations among these data are complex tasks. High dimensions increase the complexity of the classification process and the prediction of disease type and since there are redundancies and duplications of genes, the classification accuracy may degrade. Experiments which are carried out to extract the gene expression matrix are costly and due to the limited number of experiments, we are faced with the small size of the data set. The small size of the data set leads to another challenge. Among those people that are tested most are not likely to have the disease and those which are suspicious to have cancer or tumor are less likely. In this case, we are faced with minority in genetic databases who are sick.

Different strategies have been proposed over the last years for feature selection, including filter, wrapper [28], embedded [29], and more recently ensemble techniques [30].

The feature selection is an essential problem for pattern classification. The proposed feature selection approach is based on ranking methods [2] to select each characteristic individually, and without considering the relationship between features while other feature selection approaches are based on the selection of the best subset of features considering the interaction between the features [3].

Feature ranking based selection methods evaluate the significance of features according to some measurements, such as distance [31,32], and information theory [33]. Among the distance based measures, Relief, which is firstly proposed by Kira and

[*] Corresponding author.
E-mail addresses: Yassimaryam@gmail.com (M. Yassi), Moattar@mshdiau.ac.ir (M.H. Moattar).

http://dx.doi.org/10.1016/j.bbrc.2014.02.146
0006-291X/© 2014 Elsevier Inc. All rights reserved.

Rendell [31] is one of the most successful ones and adopt Euclidean distance to assign a relevance weight to each feature. The key idea of Relief is to iteratively estimate feature weights according to their ability to discriminate between nearby instances. However the optimality of Relief is not guaranteed because Relief randomly picks out an instance from training dataset.

Subset selection algorithms search the set of possible features for the optimal subset in which features are relevant in the given model. One critical problem for feature subset selection methods is that exhaustive search and evaluation of all the possible feature subsets, which usually ends in a considerably high computational complexity [34]. Thus, many heuristic subset search strategies have been introduced [35], such as sequential forward/backward selection, random selection [36], and branch and bound search [37]. A good feature subset is one that contains features highly correlated with the class, yet uncorrelated with each other [38].

Robust feature selection algorithms are related to the sensitivity of the algorithm facing with various real world conditions, such as disruptions in the training data. If a feature selection algorithm is not robust, increasing and decreasing the training samples will lead to different results. To enhance the strength of feature selection and not to get stuck in local optima, authors of [4,5] describe the integration of various criteria to build a more robust and appropriate dataset. The point which is an important factor in the feature selection algorithm is the stability of feature selection.

Considering imbalanced databases is a comprehensive look at the real world. Authors of [6] offer a feature ranking method based on density estimation to deal with the problem of imbalanced classes. The authors of [7], using genetic algorithm combined with fuzzy theory to select the most distinctive features. The Huang-index method using fuzzy c-means is employed to enhance cluster validity and achieve consistent clusters of the features. Also a new entropy-based feature evaluation method is formulated for the authentication of relevant features. Then, multivariate statistical analyses are utilized to solve the redundancy between relevant features [39]. Criteria to describe the relationship between features includes: Entropy, Mutual information and Information-F.

The authors of [8] determine the relevance of each feature with other features using similarity relationships among the data sets. The method used is based on an unsupervised process. First, the discrimination of each feature is calculated. In the next, step of the features that are rated higher in terms of discrimination are selected in an iterative process. Authors of [40] offer a dynamic weighting-based feature selection algorithm, which not only selects the most relevant features and eliminates redundant features, but also tries to retain useful intrinsic groups of interdependent features. The primary characteristic of the method is that the feature is weighted according to its interaction with the selected features and the weight of features will be dynamically updated after each candidate feature has been selected. The authors of [9,10] try to overcome the weaknesses of theoretical methods using cooperative game. This means that if the discrimination of a single feature is weak, it will improve when placed alongside other features. Therefore, this method represents the strength of features to make any discrimination imbalanced nature of the dataset is another problem in this field.

Due to the great importance that today genetic data has, it is are required to present a method which has an appropriate performance and also can overcome mentioned challenges. The method proposed in this paper tries to solve the problems of complexity and overcome the above mentioned challenges. One of the challenges related to high dimensional data sets of genetic is the dimension reduction based on feature selection to specify distinct and superior genes.

The first stage of the feature selection process consists of two parts. In the first part, the process of feature selection integrates ranking methods. The integration of ranking methods causes better robustness and feature selection stability. In the second part of the feature selection process, a wrapper technique is applied [1] which has the ability to express interactions between genes. In the first part, the unique characteristics of selected features are analyzed; while in the second part the interaction between features are analyzed.

The second phase of the proposed approach is based on SVM classifier. Using SVM classifier is due to the high generalization ability. As previously mentioned, genetic data sets have small sizes and low generalization ability. Thus, SVM is appropriate to be used to overcome this problem. As previously mentioned, the genetic datasets are imbalanced and SVM is highly sensitive against imbalanced datasets. Separating hyper plane would not be located properly between data of two classes and will be close to the majority class. In this paper, a method is proposed to solve this problem by removing the data points from majority class which are far from the decision boundary.

Organization of the materials presented in this paper is as follows: In the Section 2 the proposed method is introduced. In Section 3, the simulation results obtained from the proposed method are discussed. Conclusions and future works are described in Section 4.

2. Materials and methods

Due to the complexity of high-dimensional problems, we need to provide an intelligent way to select appropriate features. The best feature set is that with the highest performance and the lowest classification error.

2.1. Ranking method

The proposed method of feature selection is grounded on filtering techniques based on the ranking of features. Top ranked features which are selected based on statistical approaches and represent enough information and functionality to enhance the performance of classification is chosen. Ranking methods used in this study is tabulated in Table 1.

In the above mentioned, x_i is attribute values of ith sample, $\bar{x}_1$ is its average value and σ_{x_i} the characterizes the variance of samples. C is the class label and parameters such as n_1 and n_2 demonstrates the number of samples belonging to any particular features in the corresponding class label. Also S_W and S_B are the within-class and between-class scatter matrix, respectively.

2.2. Integrating ranking methods

The voting system we are going to use in this study merges the output of each of the ranking methods. The output of each method is a ranking list in the descending order. In this voting system, voters are ranking functions and volunteers are the entire set of features. Finally, the integration of the outputs of the ranking methods is a list of features sorted by the votes earned from each ranking function.

Our proposed voting system determines which features are in the top of the ranking functions. Distinctive and more superior features are selected based on total votes received for that feature. This study develops an integration method which provides the stability of feature selection process.

2.3. Integrating wrapper method with ranking methods

Integrating ranking methods using voting system is already investigated. But feature relations and response characteristics has not been studied. To combine the integrated ranking method