Computers in Biology and Medicine 81 (2017) 11–23

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Wrapper-based gene selection with Markov blanket

Aiguo Wang[a], Ning An[a,*], Jing Yang[a], Guilin Chen[b], Lian Li[a], Gil Alterovitz[c,d,e]

[a] School of Computer and Information, Hefei University of Technology, Hefei, China
[b] School of Computer and Information Engineering, Chuzhou University, Chuzhou, China
[c] Center for Biomedical Informatics, Harvard Medical School, Boston, USA
[d] Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, USA
[e] Children's Hospital Informatics Program at the Harvard/MIT Division of Health Sciences and Technology, Boston, USA

ARTICLE INFO

Keywords:
Gene selection
Microarray data
Symmetric uncertainty
Markov blanket
Wrapper methods

ABSTRACT

Gene selection seeks to find a small subset of discriminant genes from the gene expression profiles. Current gene selection methods such as wrapper-based models mainly address the issue of obtaining high-quality gene subsets. However, they are considerably time consuming, due to the existence of irrelevant and redundant genes. In this study, we present an improved wrapper-based gene selection method by introducing the Markov blanket technique to reduce the required wrapper evaluation time. In addition, our method can identify targeting genes while eliminating redundant ones in an efficient way. We use ten publicly available microarray datasets to evaluate the proposed method. The results show that our method can handle gene selection effectively. Our experimental results also show that wrapper-based method combined with the Markov blanket outperforms other competing methods in terms of classification accuracy and time/space complexity.

1. Introduction

The rapid development and maturing of microarray technology enables researchers to measure the expression profiles of thousands of genes in a single experiment simultaneously [1], and the analysis of microarray data is a good alternative to the diagnosis of cancers and the discovery of disease biomarkers at the molecular level [2,3]. Accordingly, various statistical analysis methods and machine learning models have been utilized to analyze gene expression profiles, whereas the intrinsic nature of microarray data that are characterized by small sample sizes and high dimensionality largely hinders their meaningful applications in practice [4,5]. For example, in the diagnosis of cancer with microarray data, since the number of genes typically exceeds the number of available samples, classifiers that are directly constructed on such data may suffer from poor generalization capacity and weak robustness [6]. In addition, there are relevant studies suggesting that only a few discriminant genes are associated with a certain cancer but predictive for cancer diagnosis [7], and that the original gene space consists of a wealth of noisy and redundant genes, which deteriorates the performance of a classification model. Naïve Bayes, for example, is sensitive to redundant features, and nearest neighbor-based learners are susceptible to irrelevant features in handling classification problems [8]. One feasible way to mitigate this problem is to select a small subset of discriminant genes from original gene space using an effective gene selection method [9,10].

Feature selection, also known as gene selection in the context of microarray data, plays an important role in the analysis of gene expression profiles, ranging from cancer diagnosis and gene clustering to tumor subtype classification and disease gene discovery [5]. Feature selection is a process of finding a small subset of informative features that are relevant to a specific task by discarding irrelevant and redundant features [11]. Besides reducing the high dimensionality, feature selection offers a multitude of benefits, including reducing time costs in classifier training, enhancing the generalization capacity of the constructed classifier, and helping biologists understand the underlying biological mechanisms and biologically validate the drug targets efficiently [12,13]. According to the framework proposed by Dash and Liu [14], feature selection methods typically consists of two components: a feature subset generator module and an evaluator module. The former exploits a given search strategy to generate candidate feature subsets, while the latter evaluates the quality of a feature or a subset of features and feeds the evaluation information to the feature subset generator to guide the next-round search of candidate feature subset. In feature selection, establishing powerful evaluation criteria for measuring the goodness of a feature subset largely determines the quality of finally selected features. Depending on

* Correspondence to: Hefei University of Technology, Hefei Tunxi Road 193, Hefei 230009, China.
E-mail addresses: wangaiguo2546@163.com (A. Wang), ning.g.an@acm.org (N. An), jsjyj0801@163.com (J. Yang), glchen@chzu.edu.cn (G. Chen), llian@hfut.edu.cn (L. Li), gil_alterovitz@hms.harvard.edu (G. Alterovitz).

http://dx.doi.org/10.1016/j.compbiomed.2016.12.002
Received 14 September 2016; Received in revised form 17 November 2016; Accepted 2 December 2016

whether a classifier is used as the evaluation function, we can group existing feature selection methods into three categories: filter methods, wrapper methods, and embedded methods [15]. Filter methods are independent of a classification model and measure the quality of a feature subset using only the intrinsic properties of training samples, so they are flexible in combination with various classifiers and have lower computational complexity. Further, commonly used filter metrics include distance-, dependency-, consistency-, and information theory-based metrics [14,16,17]. Compared with other three metrics, feature selection methods with information theory have drawn much more attention because of their effectiveness and efficiency, and the capacity in reflecting the non-linear relationships among variables and capturing high order statistics of data. Correspondingly, researchers have proposed and developed a number of feature selectors on the basis of mutual information, such as symmetric uncertainty (SU), fast correlation based filter (FCBF), mutual information feature selection (MIFS), conditional mutual information maximum (CMIM), minimum redundancy maximum relevance (mRMR), and joint mutual information (JMI) [17]. In contrast to filter methods, wrapper methods use a specific classifier to evaluate the quality of a feature, and often use the classification accuracy or error rate as an evaluation criterion [18,19]. Because wrapper methods search for a feature subset that is best suited to a classifier, they generally obtain better classification performance but at the cost of high time complexity [19]. Embedded methods are special cases of wrapper methods, and feature subsets are obtained when they are used to construct the classifier. This makes them usually more tractable than wrapper methods [20], and there are many embedded methods available and many of them support multiple class problems, such as random forest feature selection, multi-task lasso [21].

Though wrapper methods generally achieve better classification accuracy than filter methods, a major disadvantage is that they are considerably time-consuming. For a dataset with N features, wrapper methods approximately evaluate the quality of $O(N^2)$ feature subsets when using the sequential selection scheme [8], and even incremental wrapper methods handle a linear or sub-quadratic number of candidate feature subsets [22,23]. Such a large number of wrapper evaluations would require a large amount of CPU time when they work on high-dimensional microarray data. To this end, we present a novel model that combines wrapper-based feature selection with the Markov blanket technique. Markov blanket is a cross-entropy based technique that considers the relevance between features, and is capable of explicitly identifying and removing redundant genes. Given the Markov blanket, the eliminated features are conditionally independent of the target class [24], then they have no relevance to the target class, thus can be removed safely. This enables us to identify redundant features in a filter way rather than in a wrapper way and further reduce the number of wrapper evaluations, which leads to better time performance. In addition, it obtains better classification accuracy compared with other methods without introducing Markov blanket, as shown in our preliminary experimental results [25]. The main contributions of this study are as follows. (1) We propose to combine wrapper-based gene selection with the Markov blanket technique to accelerate the feature selection process without degrading the classification performance. Two types of specific feature selectors are implemented based on our approach in this paper. (2) We conducted extensive experiments to verify the effectiveness and efficiency of the proposed methods on ten benchmark microarray datasets with three popular classifiers. The results show our approach outperforms other competing methods. (3) We analyze the theoretical space and time complexity of the proposed approach, and find it is superior in practice. (4) By conducting the feature subset consistency analysis, we find that the resulting set of cancer-predictive genes is not unique. It indicates that there probably exist different subsets of genes in achieving similar or equal predictive classification performance in cancer diagnosis, which facilitates the comprehensive study of disease specific genes.

The rest of this paper is organized as follows. Section 2 briefly illustrates the wrapper-based feature selection methods, symmetric uncertainty, as well as the relevance criteria for feature inclusion. In Section 3, we first introduce several definitions and the Markov blanket, and then detail the proposed feature selection methods. Experimental setting and results are illustrated in Section 4, and Section 5 analyzes the theoretical space and time complexity. Finally, we conclude it with a brief summary.

2. Wrapper-based feature selection

2.1. Wrapper-based feature selection with sequential forward selection

Because wrapper methods use a classifier to measure the quality of a feature subset, they generally obtain low classification error rates due to the specific interaction between the classifier and training set. Obviously, enumerating all combinations of features and evaluating their qualities in turn guarantee obtaining the globally optimal one, but at the cost of high computational complexity that grows exponentially with the number of features [18]. In practice, such high time complexity is often unacceptable, particularly for the gene expression profiles with high dimensionality. To accelerate this process, researchers have proposed various search strategies to generate candidates. In feature selection, commonly used search schemes include, but not limited to, sequential forward selection (SFS), sequential backward selection (SBS), sequential floating search, bidirectional search, random search, and heuristic search [18]. Among these search strategies, SFS achieves a better tradeoff between the quality of the obtained feature subset and the computational complexity. Specifically, initializing the selected feature subset to be empty, SFS selects the first feature that is most relevant to the target class, and then searches for the next candidate feature that most reduces the classification error rate. Continue with the procedure until there is no candidate feature left or no further improvement in classification performance. If k features are finally selected from the total N features, wrapper methods with SFS approximately evaluate $O(kN)$ candidate feature subsets. Algorithm 1 presents corresponding pseudo-code. The *evaluate*() subroutine (Line 7) is the evaluation process for measuring the quality of a candidate gene. The criteria used to select a candidate feature and the notations used in Algorithm 1 are given in subSection 2.2.

2.2. Relevance criteria

In wrapper-based feature selection, the criterion to add a candidate feature f into the selected feature set S is to conduct an inner five-fold cross-validation on training set *Data* projected over $\{S, f\}$ and class label C of *Data*. We use the symbol "↓" to represent the projection over a dataset. For example, $Data^{\downarrow S}$ indicates that we obtain a new dataset that consists of $|S|$ column vectors (selected according to S) from *Data*, i.e., the new dataset is a slice of *Data*. Rather than use the average accuracy of the five-fold cross validation and do a t-test over the cross-validation results [22,29], we adopt the following criteria: (1) a five-fold cross-validation is used on *Data*; (2) the new feature f is selected only if the average accuracy of the five-fold cross-validation over $Data^{\downarrow\{S\cup f\cup C\}}$ is higher than that of the five-fold cross-validation on $Data^{\downarrow\{S\cup C\}}$, and at least MinFoldersBetter (mf) out of the five accuracies over $Data^{\downarrow\{S\cup f\cup C\}}$ is not lower than the average accuracy over $Data^{\downarrow S\cup}$. Such a strategy avoids the criticism for the use of a statistical test on a dataset of small size. Notably, mf is a user-specified threshold. For the better control of low-confidence and over-fitting issues, recommended empirical values for mf are 2 or 3 [8]. The quality of a candidate feature is measured by *evaluate*(*classifier*, $Data^{\downarrow\{S_{new}\cup C\}}$), which returns two items: the average accuracy acc_{new} of the five-fold cross-validation and the number num representing how many times the five accuracies obtained from the five-fold cross-validation over $Data^{\downarrow\{S\cup f\cup C\}}$ are better than average accuracy over $Data^{\downarrow\{S\cup C\}}$.