Pattern Recognition 42 (2009) 1761 – 1767

Contents lists available at ScienceDirect

Pattern Recognition

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

Gene boosting for cancer classification based on gene expression profiles

Jin-Hyuk Hong, Sung-Bae Cho*

Department of Computer Science, Yonsei University, 134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Republic of Korea

ARTICLE INFO

Article history:
Received 11 February 2008
Received in revised form 12 November 2008
Accepted 7 January 2009

Keywords:
Gene selection
Cancer classification
Wrapper method
Filter method
Boosting

ABSTRACT

Gene selection is one of the important issues for cancer classification based on gene expression profiles. Filter and wrapper approaches are widely used for gene selection, where the former is hard to measure the relationship between genes and the latter requires lots of computation. We present a novel method, called gene boosting, to select relevant gene subsets by integrating filter and wrapper approaches. It repeatedly selects a set of top-ranked informative genes by a filtering algorithm with respect to a temporal training dataset constructed according to the classification result for the original training dataset. Empirical results on three microarray benchmark datasets have shown that the proposed method is effective and efficient in finding a relevant and concise gene subset. It achieved competitive performance with fewer genes in a reasonable time, as well as led to the identification of some genes frequently getting selected.

1. Introduction

Promising a new insight into the mechanisms of living things, DNA microarray technology measures the expression level of thousands of genes simultaneously [1]. Some genes could be related to a particular type of cancer, but many of them are irrelevant or redundant features that affect the speed and accuracy of classification [2]. Gene selection that identifies the optimal subset of relevant genes is one of the major challenges in cancer classification based on gene expression profiles. It helps improve classification accuracy, reduce the computational cost, and gain significant insight into the inherent cancer mechanisms [1,3].

Gene selection methods can be categorized into filter and wrapper approaches [2,4]. The filter method selects the top-ranked genes according to their individual discriminative power without involving any induction algorithm. Genes are evaluated by various measures of the general characteristics of the data, and the performance of filter-based gene selection is generally determined by those measures. It is efficient for high-dimensional data owing to its linear time complexity, but it cannot discover the synergy effect or suppressibility among genes. The wrapper method, in contrast, evaluates candidate gene subsets by using an induction algorithm. Since the predictive accuracy of the induction algorithm determines the goodness of the selected subsets, it is capable of considering the correlations among genes but often computationally expensive [5–7].

In many studies on cancer classification using microarray data, filter approaches have been widely investigated. Lee et al. [8] have developed a multivariate Bayesian model for gene selection by using a combination of truncated sampling and Markov Chain Monte Carlo (MCMC), while Bae and Mallick [9] have improved the model by using a two-level hierarchical Bayesian model. Wang et al. [3] have combined gene ranking and clustering analysis, and Guan and Zhao [10] have proposed a semiparametric two-sample test to identify differentially expressed genes and to select marker genes. Li et al. [11] and Statnikov et al. [12] have compared conventional gene ranking measures such as *t*-statistics, information gain, signal-to-noise ratio, etc.

On the other hand, recent works on gene selection tend toward wrapper approaches. Li et al. [13] have introduced a multivariate approach by using the genetic algorithm and the *k*-nearest neighbor method and showed the capability of wrapper approaches, and Liu et al. [14] and Li et al. [2] have used support vector machines, instead of the *k*-nearest neighbor method, to incorporate with the genetic algorithm. Zhu et al. [1] have proposed a Markov blanket-embedded genetic algorithm that adds or deletes genes through evolution, while Banerjee et al. [4] have employed the rough set theory to represent the minimal sets of non-redundant genes in a multi-objective framework and used the multi-objective genetic algorithm to generate minimal gene subsets.

Different from the wrapper methods based on evolutionary computation, some researchers have used a recursive heuristic algorithm. Li and Yang [15] have proposed a wrapper method that recursively eliminates redundant genes according to the accuracy of classification, while Ruiz et al. [6] have presented the best incremental ranked subset (BIRS) algorithm that adds a gene according to the statistical

* Corresponding author. Tel.: +82 2 2123 3877; fax: +82 2 365 2579.
E-mail address: hjinh@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr (S.-B. Cho).

0031-3203/$ - see front matter © 2009 Published by Elsevier Ltd.
doi:10.1016/j.patcog.2009.01.006

significance for the improvement of classification. Tang et al. [16] have proposed the two-stage support vector machine-recursive feature elimination (SVM-RFE) algorithm. In each loop, it calculates the change of the margin width of SVMs after removing a gene, which signifies the weight of the gene, and ranks genes according to their weights. Genes with the smallest weight are removed from the gene subset.

In order to get a subset of non-redundant but still highly informative genes, in this paper, we propose a gene boosting technique using the combination of filter and wrapper methods. The proposed method selects a set of top-ranked informative genes by a filtering algorithm, and then classifies training samples by using an induction algorithm with the selected genes. According to the classification result, it constructs a temporal dataset for selecting other informative genes, and appends new genes into the gene subset. The method iterates the process until it satisfies a termination condition like no more improvement of classification for training samples or reaching to the target size of gene subsets. Contrary to conventional wrapper-based gene selection methods that are computationally expensive, the proposed method provides the efficiency of applying wrapper approach in high-dimensional domains and obtains better results than the filter approach. We will show the usefulness of the proposed method on three popular cancer datasets.

2. Proposed method

2.1. Gene boosting

Boosting, proposed by Freund and Schapire in 1996, is an ensemble method of producing a series of base classifiers, which are trained with the iteratively reweighted or re-sampled training data including more difficult cases [17,18]. In this work, we propose a novel gene selection method (named gene boosting) that combines filter and wrapper approaches based on boosting by re-sampling. Contrary to conventional approaches that apply boosting to the construction of base classifiers in ensembling, the proposed method is a novel attempt to employ boosting in gene selection.

For a given training dataset $\{(x_i, y_i) \in \chi \times \{-1, +1\},\ i=1, \ldots, m\}$, a filter method selects a set of new informative genes N_t iteratively with a re-sampled population R_t from χ, at each round $t = 1, \ldots, T$, where they help to minimize the error with respect to the distribution D_t. A base classifier $f_t(x_i) : \chi \rightarrow [-1, +1]$ is trained with a set of genes $G (G = G \cup \{N_t\})$, which is incrementally appended with the newly selected genes N_t. The boosting procedure terminates when it satisfies a condition, and the gradually built gene subset is $G = \{N_1, \ldots, N_t\}$.

In order to re-sample the population R_t, a distribution function $D_t(i)$ assigns the importance to the sample x_i. For the first round, all samples have the same importance, $D_1(i) = \frac{1}{m}$, $\forall i = 1, \ldots, m$, and in each round the importance is updated according to Eq. (1).

$$D_{t+1}(i) = \frac{D_t(i) \times \exp[-\alpha_t \cdot f_t(x_i) \cdot y_i]}{Z_t} \tag{1}$$

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t} \tag{2}$$

$$\varepsilon_t = \sum_{i=1}^{m} D_t(i) \times p(f_t(x_i) \neq y_i) \tag{3}$$

Z_t is a normalization factor that makes $\sum_{i=1}^{m} D_t(i) = 1$. This procedure leads to including more misclassified samples into the re-sampled population R_t in the next round. It has been theoretically shown that the training error of classification is bounded as follows [17].

$$\frac{1}{m} |\{i : f(x_i) \neq y_i\}| \leqslant \prod_{t=1}^{T} Z_t \tag{4}$$

2.2. Gene boosting-based cancer classification

Informative genes for cancer classification are incrementally selected by gene boosting proposed in this paper. It basically follows the wrapper approach, but directly manipulates training samples contrary to the conventional wrapper approaches that only use the accuracy of classification. A filter method, embedded in this method, selects genes with respect to the training samples reconstructed. This increases the speed of gene selection, since it does not evaluate all possible gene subsets like the conventional methods. Given n genes, the proposed method measures their usefulness n times for each iteration and finally obtains $inc \times T$ genes within the time complexity of $O(n \times T)$. On the other hand, a conventional wrapper-based method needs to evaluate n^{inc} combinations of genes per iteration to obtain $inc \times T$ genes, which might be unacceptably large like $O(n^{inc} \times T)$ when the number of genes incrementally added is large. The brief overview of the proposed and conventional wrapper-based gene selection methods is as follows.

Proposed method	Conventional wrapper-based method
For T iterations	For T iterations
Evaluate n genes individually	Evaluate all possible combinations of n genes whose size is inc
Sort them according to their ranks	Append inc genes of the best combination
Append top inc genes	

The proposed gene boosting algorithm for gene selection is as follows. For each round, classification results obtained by CL( ) are used to adjust the distribution D that affects to select informative genes by the filter method FS( ). Genes newly appended are informative to classify samples misclassified with the current gene subset, thereby it gradually improves the classification performance. In this work, kNN($k(=5)$ nearest neighbor) with Euclidean distance is used for the base classifier CL( ), and inc (# of genes appended for each loop) whose performance on the training dataset is the highest is selected among several candidates.

GeneBoost_CancerClassification(χ, inc)
  // χ: {training dataset}
  // inc: # of genes incrementally added
Initialize:
  $G := \{\phi\}$
  $R_t := \{\phi\}$
  $\forall i \in \{1, \ldots, m\}$, $D_1(i) := \frac{1}{m}$
  $N_t := FS(\chi, D_1, inc)$ // filter-based gene selection function
for $t = 1, \ldots, T$
  $G := G \cup N_t$
  $D_{t+1} := CL(\chi, D_t, G)$ // classification function
  $N_{t+1} := FS(\chi, D_{t+1}, inc)$
end
return G

In order to select a subset of informative genes, in this paper, we use a popular filter-based gene selection method that measures the similarity with a predefined ideal marker gene [19]. At first, we construct a temporal training dataset (m samples) that includes x_i in proportion of $D(i)/(\sum_{k=1}^{m} D(k))$. Assume the class label $y_i \in Y = \{-1, +1\}$, and we can define two ideal marker genes K^+ and K^- represented as