Pattern Recognition Letters 60–61 (2015) 16–23

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A novel aggregate gene selection method for microarray data classification ☆

Thanh Nguyen*, Abbas Khosravi, Douglas Creighton, Saeid Nahavandi

Centre for Intelligent Systems Research, Deakin University, Geelong Waurn Ponds Campus, Victoria 3216, Australia

ARTICLE INFO

Article history:
Received 17 September 2014
Available online 21 April 2015

Keywords:
Gene selection
Analytic hierarchy process
Classification
Gene expression profiles
Microarray data

ABSTRACT

This paper introduces a novel method for gene selection based on a modification of analytic hierarchy process (AHP). The modified AHP (MAHP) is able to deal with quantitative factors that are statistics of five individual gene ranking methods: two-sample t-test, entropy test, receiver operating characteristic curve, Wilcoxon test, and signal to noise ratio. The most prominent discriminant genes serve as inputs to a range of classifiers including linear discriminant analysis, k-nearest neighbors, probabilistic neural network, support vector machine, and multilayer perceptron. Gene subsets selected by MAHP are compared with those of four competing approaches: information gain, symmetrical uncertainty, Bhattacharyya distance and ReliefF. Four benchmark microarray datasets: diffuse large B-cell lymphoma, leukemia cancer, prostate and colon are utilized for experiments. As the number of samples in microarray data datasets are limited, the leave one out cross validation strategy is applied rather than the traditional cross validation. Experimental results demonstrate the significant dominance of the proposed MAHP against the competing methods in terms of both accuracy and stability. With a benefit of inexpensive computational cost, MAHP is useful for cancer diagnosis using DNA gene expression profiles in the real clinical practice.

1. Introduction

The development of the DNA microarray technology enables researchers to analyze tens of thousands of genes simultaneously. One of the important applications of the DNA microarray is to investigate the differences between a healthy cell and a cancer cell. Cancer is basically a group of related diseases caused as a result of "genes gone bad". Many genes involve with cell division, growth and death. When those genes stop functioning properly, cell growth can spin out of control, leading to tumor formation and cancer. But different kinds of cancer are induced by different sets of genes. Therefore, to be able to better diagnose, understand, and treat cancer, it is essential to know which of the genes in cancer cells are working abnormally.

So far, various methods have been proposed to select significant genes. Diaz-Uriarte and de Andres [6] introduced a method for gene selection and microarray data classification based on random forest. The method produces small sets of genes that retain a high predictive accuracy. A family of sparse learning methods such as sparse logistic regression [18] and sparse linear discriminant analysis [23] were recommended to deal with high-dimension, low-sample gene expression data. In another approach, Tapia et al. [21] developed a sparse and stable SVM-RFE that performs gene selection at affordable computational costs in two stages, i.e. spreading and dispreading. Likewise, a general framework of sample weighting to improve the stability of feature selection method under sample variations was presented in [25].

A hybrid approach that embeds the Markov Blanket with the harmony search algorithm for gene selection was suggested by Shreem et al. [19]. The procedure works well on selected genes with higher correlation coefficients based on symmetrical uncertainty. Alternatively, Cai et al. [2] initiated a feature weighting algorithm for gene selection called LHR. LHR estimates the feature weights through local approximation based on ReliefF.

Recently, Han et al. [9] employed the gene-to-class sensitivity exploited by a single hidden layered feedforward neural network in a hybrid gene selection. The method uses k-means clustering and binary particle swarm optimization for filtering redundant genes.

For evaluating a gene selection method, in addition to the predictive ability of gene subsets, two other important aspects that need to be considered are the stability of the selected genes and the computational costs. This paper introduces a modification to the analytic hierarchy process (AHP) for a novel gene selection method. Traditional AHP often deals with qualitative factors that are derived from experts [17]. Given that the number of genes in microarray data are at around tens of thousands and the gene knowledge available to experts is always limited, completion of assessments of various genes with

☆ This paper has been recommended for acceptance by Qian Xiaoning.

* Corresponding author. Tel.: +61 3 52278281; fax: +61 3 52271046.
E-mail address: thanh.nguyen@deakin.edu.au (T. Nguyen).

http://dx.doi.org/10.1016/j.patrec.2015.03.018
0167-8655/

respect to various criteria is not always a practical proposition. The modified AHP (MAHP) is able to quantitatively integrate statistical outcomes of individual gene ranking methods via an objective ranking procedure without consulting to possibly biased and inadequate expert knowledge. Through rigorous experiments, we show that the proposed MAHP yields gene subsets that lead to a classification stability at low computational cost without sacrificing the accuracy. The arguments are organized as follows. The next section presents a background of gene selection and the MAHP method. Experimental results are presented and discussed in Section 3, followed by conclusions in Section 4.

2. Gene selection methods

Microarray data commonly collected with the number of genes (often in tens of thousands) is much larger than the number of samples. Many standard techniques therefore find inappropriate or computationally infeasible in analyzing such data. The fact is that not all of the thousands of genes are discriminative and needed for classification. Most genes are not relevant and do not affect the classification performance. Taking such genes into account enlarges the dimension of the problem, leads to computational burden, and presents unnecessary noise in the classification practice. Therefore it is essential to select a small number of genes, called informative genes, which can suffice for good classification. However, the best subset of genes is often unknown [24].

Common gene selection approaches are filter and wrapper methods. Filter methods rank all features in terms of their goodness using the relation of each single gene with the class label based on a univariate scoring metric. The top ranked genes are chosen before classification techniques are executed. In contrast, wrapper methods require the gene selection technique to combine with a classifier to evaluate classification performance of each gene subset. The optimal subset of genes is identified based on the ranking of performance derived from implementing the classifier on all found subsets. The filter procedure is unable to measure the relationship among genes whilst the wrapper approach requires a great computational expense [10].

In this paper, to enhance the robustness and stability of microarray data classification, we introduce a novel gene selection method based on a modification of the AHP. The idea behind this approach is to assemble prominent discriminant genes from different gene selection ranking methods through a systematic hierarchy.

The next subsections scrutinize background of common gene selection filter methods, which are followed by our proposal. Note that the following gene selection methods are accomplished by ranking genes via scoring metrics. They are statistic tests based on two data samples in the binary classification problem. The sample means are denoted as μ_1 and μ_2, whereas σ_1 and σ_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes.

2.1. Two-sample t-test

The two-sample t-test is a parametric hypothesis test that is applied to compare whether the average difference between two independent data samples is really significant. The test statistic is expressed by:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{1}$$

In the application of t-test for gene selection, the test is performed on each gene by separating the expression levels based on the class variable. Note that the absolute value of t is used to evaluate the significance among genes. The higher the absolute value, the more important is the gene.

2.2. Entropy test

Relative entropy, also known as Kullback–Liebler distance or divergence is a test assuming classes are normally distributed. The entropy score for each gene is computed using the following expression:

$$e = \frac{1}{2}\left[\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 2\right) + \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)(\mu_1 - \mu_2)^2\right] \tag{2}$$

After the computation is accomplished for every gene, genes with the greatest entropy scores are selected to serve as inputs to the classification techniques.

2.3. Receiver operating characteristic (ROC) curve

Denote the distribution functions of X in the two populations as $F_1(x)$ and $F_2(x)$. The tail functions are specified respectively $T_i(x) = 1 - F_i(x)$, $i = 1, 2$. The ROC is given as follows:

$$ROC(t) = T_1\left(T_2^{-1}(t)\right), \quad t \in (0, 1) \tag{3}$$

and the area under the curve (AUC) is computed by:

$$AUC = \int_0^1 ROC(t)dt \tag{4}$$

The larger the AUC, the less is the overlap of the classes. Genes with the greatest AUC therefore are chosen to form a gene set.

2.4. Wilcoxon method

The Wilcoxon rank sum test is equivalent to the Mann–Whitney U-test, which is a test for equality of population locations (medians). The null hypothesis is that two populations enclose identical distribution functions whereas the alternative hypothesis refers to the case two distributions differ regarding the medians. The normality assumption regarding the differences between the two samples is not required. That is why this test is used instead of the two-sample t-test in many applications when the normality assumption is concerned.

The main steps of the Wilcoxon test (see [5]) are summarized below:

1) Assemble all observations of the two populations and rank them in the ascending order.
2) The Wilcoxon statistic is calculated by the sum of all the ranks associated with the observations from the smaller group.
3) The hypothesis decision is made based on the p-value, which is found from the Wilcoxon rank sum distribution table.

In the applications of the Wilcoxon test for gene selection, the absolute values of the standardized Wilcoxon statistics are utilized to rank genes.

2.5. Signal to noise ratio (SNR)

SNR defines the relative class separation metric by:

$$SNR(f_i, c) = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \tag{5}$$

where c is the class vector, f_i is the ith feature vector. By treating each gene as a feature, we transform the SNR for feature selection to gene selection problem for microarray data classification.

SNR implies that the distance between the means of two classes is a measure for separation. Furthermore, the small standard deviation favors the separation between classes. The distance between mean values is thus normalized by the standard deviation of the classes [8].