ELSEVIER

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

The linear neuron as marker selector and clinical predictor in cancer gene analysis

Michalis E. Blazadonakis[*], *Michalis Zervakis*

Technical University of Crete, Department of Electronic and Computer Engineering, University Campus, Chania Crete 73100, Greece

ARTICLE INFO

Article history:
Received 11 June 2007
Received in revised form
22 February 2008
Accepted 23 February 2008

Keywords:
Microarray analysis
Gene selection
Wrapper methods
Filter methods
RFE methods
External cross-validation
Survival analysis

ABSTRACT

Objective: The problem of gene selection has been extensively studied in a number of scientific works using various kinds of methods. However, the application of a linear neuron is a novel approach possessing several advantages. In this work we propose to study the behavior of such a linear neuron, appropriately adapted and trained to the problem of gene selection in the DNA microarray experiment.

Methods and materials: We explore the proposed approach in terms of an accuracy evaluation criterion, which is used to assess the performance of the proposed methodology, but we also evaluate the produced results in terms of cluster quality and survival prediction. Cluster quality reflects the ability of the method to select differentially expressed genes, which in turn leads to better clustering and survival prediction.

Results: We directly compare the proposed methodology with RFE-SVM, a well known and broadly accepted method demonstrating remarkable performance on various data sets of clinical interest.

Conclusions: Conducted computational experiments show that the proposed approach can be efficiently used within the field of gene selection producing high-quality results in terms of accuracy and robustness.

1. Introduction

The advent of microarray technology has given scientists a valuable tool to monitor the behavior of thousands of genes in a single experiment. The behavior of each gene is kept in a separate cell in an m by n expression matrix M, where each row corresponds to the expression levels of a single gene, while each column corresponds to a different examination (sample). The expression level (behavior) of each gene is recorded in terms of a color map varying in a range from green to red for instance, with the mid-point being represented by black. A green colored cell manifests that the specific gene has expressed itself more in the normal than in the pathological state, a red color in a cell implies exactly the opposite, while a black color means that the specific gene has expressed itself in exactly the same way in both situations. Colors are translated into numbers on a closed interval $[-3, +3]$ for instance, where -3, 0 and $+3$ indicate green, black and red, respectively. In these kinds of experiments we encounter the problem that the number of features (genes) is much larger (order of thousands) than the number of samples (order of tenths). The primary goal of gene selection is to find a set of genes with size much smaller than the initial, which is able to describe the data set of interest fairly well both in terms of classification accuracy and quality. The first attribute (accuracy) relates to the ability of the selected genes to successfully classify samples into their correct class, whereas the quality attribute reflects the ability of each gene to clearly differentiate its expression between

[*] *Corresponding author*. Tel.: +30 28210 37203; fax: +30 28210 37542.
E-mail address: mblazad@ier.forthnet.gr (M.E. Blazadonakis).

0169-2607/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2008.02.009

the states of interest. This fact has been implicitly implied in almost every gene selection study (we selectively refer to [2–8]) and has also been explicitly stated in [9–12]. The concise study of a small number of genes can help biologists to get significant insight into the genetic structure and mechanisms involved in a specific disease, which may lead to drug discovery and early diagnosis.

Feature selection methods can be roughly divided into two categories [13], i.e. filtering approaches where features are ranked in a pre-processing step according to some weight coefficient independent of the classification method and wrapper approaches where a classifier is used to generate scores to be used as the feature ranking criterion. Wrapper approaches are very much dependent on the classification outcome, since they follow a recursive process where feature weights are re-evaluated in every classification cycle. Thus, wrapper approaches focus on accuracy neglecting quality aspects, whereas filter methods consider mainly the quality aspects neglecting gene interactions. The formulation of our method aims towards merging both of those aspects into a single approach.

Among the various feature selection methods proposed, RFE-SVM [1] is an approach that has shown remarkable results in leukemia [3] and colon cancer [14] datasets. However, depending on the data distribution and the complexity of the classification problem, the algorithmic design based on the philosophy of RFE-SVM may lead to ill-defined and ill-distinctive clusters of selected gene signatures, as it is demonstrated in Section 4. This issue has been implicitly addressed in [15], showing that wrapper methods do not provide sufficient focus for further investigation (of their result) because many genes may be included by chance. In this study we go one step further and demonstrate that such a lack of focus could lead to the production of ill-distinctive clusters. To overcome this inadequacy we propose a new criterion for the RFE recursion based on an appropriately designed linear neuron. The proposed network acts as a linear filter for classification, but in contrast to SVMs it employs an appropriate learning procedure taking under consideration the quality aspect of selected genes, thus producing more compact and distinct clusters of markers. We refer to this new approach as recursive feature elimination based on linear neuron weights (RFE-LNW). Both RFE-SVM and RFE-LNW are wrapper feature selection methods, but the latter can embed and exploit issues from the philosophy of filter approaches.

The idea of applying a linear neuron to the problem of gene selection is a novel approach, introduced in our initial studies [11,12]. It is based on the ability of a single neuron to approximate any linear function. This idea absolutely complies with the philosophy of linear methods such as the RFE-SVM, which is based on linear support vector machines. In this study, we expand the proposed methodology by introducing an appropriately adopted learning procedure incorporating a variation of Fisher's ratio, with the aim of enriching wrapper methods with filter criteria or vice versa.

Besides methodology, another important issue concerning gene selection relates to the measures used for validation of the classification performance (prediction rule) of the selected markers. It is a common practice to assess the performance of a method by its leave one out cross-validation (LOOCV) error. Two types of LOOCV schemes are generally considered and both are assessed in this study. The first one addresses the removal of the left-out sample before the selection of differentially expressed genes and the application of the prediction rule, while the second approach handles the removal of the left-out sample after the selection process but before the application of the prediction rule. The first is usually referred to as the external LOOCV (ELOOCV) while the second is referred to as the internal LOOCV (ILOOCV) [8,16].

It is obvious that ELOOCV is a more unbiased estimator of the error rate since it is totally independent of the selection process. However, ILOOCV provides a measure that cannot be neglected, as it expresses the training ability of a selection rule within the training set. In other words, ILOOCV indicates the selection rule(s) that can learn or generalize better on the training set. The ILOOCV scheme in combination with an independent test set was used to assess the performance of RFE-SVM in [1]. Towards a fair and statistically sound comparison of the methods in this study, we assess both of these measures in combination with a 10-fold cross-validation process in all evaluation steps. Proceeding one step further than classification accuracy and cluster quality, we test the result derived through the ELOOCV procedure as a set of selected genes that can provide high classification accuracy on the independent test published by Van't Veer et al. [7]. In addition, this set of markers is proved to be an efficient survival predictor for the 234 new cases published by Van De Vijver et al. [17].

To reveal the basic differences of the two underlined methodologies (RFE-SVM and RFE-LNW) we examine their learning ability on three data sets: (a) the data set of diffuse large B-cell lymphoma published in [18], (b) the colon cancer data set in [14] and (c) the breast cancer data set published by Van't Veer et al. [7].

2. Background

2.1. Support vector machines (SVMs)

An SVM [19] attempts to find the best separating hyperplane to distinguish between the two classes of interest, i.e. positive (+1) and negative (−1). This is done by maximizing the distance $2/||\boldsymbol{w}||$ between the two parallel lines $(\boldsymbol{w}\cdot\boldsymbol{x})+b=1$ and $(\boldsymbol{w}\cdot\boldsymbol{x})+b=-1$, which form the margin of separation as shown in Fig. 1, where $\boldsymbol{w}$ represents the direction vector of the hyperplane, $\boldsymbol{x}$ is the sample on which a decision has to be taken, and b is the hyperplane intersect with the vertical axis. The final separating hyperplane passes through the middle of this margin with equation $(\boldsymbol{w}\cdot\boldsymbol{x})+b=0$. The decision function then, is a function of the form:

$$f(\boldsymbol{x}) = sgn((\boldsymbol{w}\cdot\boldsymbol{x})+b) \tag{1}$$

The sign of the value returned by Eq. (1) indicates the predicted class associated with example $\boldsymbol{x}$, while $|f(\boldsymbol{x})|$ indicates the confidence level of the resulting decision; for a more detailed description on SVM the interested reader may refer to [19]. Towards the solution of this problem, we obtain the following