Asymmetric Support Vector Machines for gene selection

ID - 244

Muselli Marco[1], Francesca Ruffino[2]

[1]Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni, Consiglio Nazionale delle Ricerche, Genova, Italy

[2]Dipartimento di Scienze dell'Informazione, Università di Milano, Milano, Italy

Motivation

When tissues belonging to two different physio-pathological states are analyzed through DNA microarray experiments, one of the main targets amounts to finding the subset of relevant genes for the functional states in exam. Gene selection methods are automatic techniques having the aim of identifying this subset of genes.

Unfortunately, a systematic approach for verifying the effectiveness of gene selection methods has not yet been established, since no real world problem exists where the subset of relevant genes is completely known. To overcome this problem a suitable mathematical model building biologically plausible gene expression data has been introduced [1]. It is able to generate artificial datasets resembling real ones, thus allowing to analyze the validity of gene selection methods, since the actual subset of relevant genes is known in advance.

In this contribution we present a new gene selection method based on the theory of reliable learning, according to which a binary label concerning reliability can be attached to each experiment with DNA microarray. Through this approach a new classification technique, called Asymmetric Support Vector Machine (ASVM), can be derived. Although its behavior is similar to that of standard Support Vector Machines (SVM) [2], it can be shown that the generalization error of ASVM converges more rapidly to zero. The application of Recursive Feature Extraction (RFE) [3] allows to adopt ASVM for gene selection purposes.

Methods

In many classification problems a subset of examples in the training set whose output is reliable can be determined. However, the standard algorithm for SVM training is not able to manage this information, which can lead to more effective classifiers. The introduction of a new mathematical framework, called reliable learning, for treating this specific situation has led to the derivation of a new classification technique, called ASVM.

Since in general the a priori information on reliability of experiments with DNA microarray cannot be available, three different steps can be devised in ASVM training: the first two aim to determining the subset of reliable patterns for the two considered classes. The last one employs a modified version of the standard training algorithm for SVM, which is able to take into account reliability information. In particular the hard margin approach is employed for reliable samples, whereas a small amount of error is allowed when the classification of the input pattern is uncertain.

A gene selection method can be derived from ASVM by adopting the RFE procedure, here schematized: 1. The ASVM is adopted, started from the original training set, to derive the directional vector w = (w1, , wn) of the generalized optimal hyperplane; the reliability of patterns in the training set are taken into account when obtaining w.

2. The components of w are used to evaluate the relevance of the n genes. In particular, the higher is the squared value of the component wj, the more relevant is the corresponding gene gj.

3. The less relevant gene g* is added at the head of a list r (initially empty) and removed from the training set.

These three steps are repeated iteratively until the list r contains all the genes in decreasing order of relevance.

Results

To evaluate the quality of the subset of genes retrieved by ASVM-RFE, two biologically plausible artificial gene expression datasets created according to a proper mathematical model [1] have been employed. In particular, the Colon cancer dataset, from Alon et al. [4], and the Leukemia dataset, from Golub et al. [5], have been considered; for each of them 50 artificial datasets with the same statistical behavior have been generated. The standard Diggle test [6] has been adopted to measure statistical similarity.

Since in every artificial dataset the subset R of relevant genes is completely known, the quality of a gene selection method can be evaluated by determining the intersection between R and the list r of relevant genes produced by the gene selection algorithm. If p is the size of this intersection, the higher is the value of p, the more effective is the corresponding method.

This procedure has been employed to compare the performances of ASVM-RFE and SVM-RFE.

For each method and each artificial dataset the value of p has been computed, analyzing the number of training sets where ASVM-RFE turns out to be more effective than SVM-RFE.
With this approach we have obtained that in 43 out of the total 50 cases resembling the Colon cancer dataset the results produced by ASVM-RFE are significantly better than those obtained with SVM-RFE. A similar conclusion has also been derived in the analysis of Leukemia dataset, where the proposed method is more effective in 41 out of 50 cases.
Although a more extensive set of trials has to be performed to obtain a reliable answer about the comparison between ASVM-RFE and SVM-RFE, these preliminary results characterize ASVM-RFE as a promising approach for gene selection.
Email: marco.muselli@ieiit.cnr.it