Incremental classification models and characterization for gene expression data analysis

ID - 146

Guarracino Mario Rosario[1,2], Feminiano Davide[1], Del Vecchio Blanco Francesca[2], Cuciniello Salvatore[1]
[1]High Performance Computing and Networking Institute v Italian Research Council
[2]Lab-GPT

Motivation

Supervised learning refers to the capability of a system to learn from a set of examples, which is a set of input/output couples. This set is called the training set. The trained system is able to provide an answer (output) for a new question (input). The term supervised originates from the fact that the desired output for the training set of points is provided by an external teacher. In the simples case, when the answer is limited between two possibilities, we refer to binary classification.

Support Vector Machine (SVMs) are the state-of-the-art for the existing classification methods. These methods classify the points from two linearly separable sets in two classes by solving a quadratic optimization problem, to find the optimal separating hyperplane between these two classes. The computed hyperplane maximizes the distance from the convex hulls of each class. These techniques can be extended to the nonlinear cases by embedding the data in a nonlinear space using kernel functions.

In case of a large number of experiments, those methods can provide classification models that do not generalize. Furthermore, in case of training set updates, existing methods need start over, to take into account the effects of new points. Finally, when applied to data with many features, as in the case of gene expression analysis, they can be slow.

Methods

In this study, we present Incremental Learning and Decremented Characterization of Regularized Generalized Eigenvalue Classification (ILDC- ReGEC), a novel algorithm derived from SVM. It is capable to train a classifier with a substantially smaller subset of points and features of the original data. The proposed method provides a constructive way to understand the influence of new training data on an existing classification model and the grouping of features that determine the class of samples.

Results

Results show that the proposed method i) has a classification accuracy comparable to other methods, ii) has a computational performance lower then most methods, and iii) provides small subsets of points and features. We show, through numerical experiments on publicly available microarray data sets, that this technique has comparable accuracy with respect to other methods. Furthermore, experiments show it is possible to obtain a classification model with about 30% of the training samples and less then 5% of initial features. Results are discussed and compared to other studies.

Email: mario.guarracino@hotmail.it