An experimental comparison of Random Subspace and Random Projection Ensembles of Support Vector Machines for the classification of gene expression data.

ID - 165

Folgieri Raffaella[1], Bertoni Alberto[1], Valentini Giorgio[1]

[1]Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Milano, Italy

Motivation

Supervised classification of gene expression data is a difficult problem due to the high dimension, low cardinality and noise of gene expression data.

For these reasons classical statistical and machine learning methods may fail to correctly predict classes of samples or genes, and several algorithms based on ensemble methods have been proposed to enhance the accuracy, the robustness and the reproducibility of the results. Anyway in most cases the approaches proposed in the literature have been mutuated from other application domains, without considering the specific characteristics of gene expression data.

Classical approaches consist in the application of statistical and machine learning methods using gene selection methods to consider the curse of dimensionality problem.

Methods

We propose for gene selection problem a randomized algorithm based on an ensemble approach. In particular we apply Random Subspace (RS) ensembles of Support Vector Machines (SVMs) to the classification of gene expression data: the SVM base learners are trained on different views of the data obtained by extracting random subsets of features from the high dimensional gene space and then their decisions are combined through majority voting.

We propose also Random Projections (RP) ensembles, a generalization of Random Subspace ensembles, where linear combinations of the features (gene expression levels) are applied to perform projections to lower dimensional subspaces. Both the proposed methods exploit the high dimensionality and the redundancy of information of gene expression data, combining multiple base learners trained on different subspaces of the data domain.

Results

We compared our proposed methods with state of the art machine learning ensemble algorithms, using single Support Vector Machines as a baseline method. Cross validated results with Leukemia (Golub et al., 1999) and Colon (Alon et al., 1999), two widely analyzed DNA microarray data sets, show that RS and RP ensembles achieve equal or better accuracies than bagged and boosted ensembles of learning machines, without using any gene selection method to reduce the dimensionality of the data.

Email: folgieri@dico.unimi.it