Poster M-18

A Simple Formula for the High Regularisation Limit of SVMs & Ridge Regression with Application to Gene Selection and Classification of Microarrays

Authors:
Justin Bedo (*ANU*)
Tiberio Caetano (*NICTA*)
Brian Parker (*NICTA*)
Conrad Sanderson (*NICTA*)
Peter Sunehag (*NICTA*)
Adam Kowalczyk (*NICTA*)

Short Abstract: Microarray experiments typically deliver datasets with few samples of high dimensionality. For Support Vector Machines (SVMs) it has recently been shown that the high amounts of regularisation can lead to better performance. We show that in the high-regularisation limit C->0, SVMS converge to a centroid classifier.

Long Abstract:
Introduction:

Typical microarray datasets are characterised by high dimensionality and small number of samples. Classification can be difficult in this context as there are insufficient samples to properly characterise the distributions of the classes representing the various tumour or tissue types. As such, regularisation and feature selection become critical for obtaining accurate and reliable classifiers.

Within the framework of Support Vector Machines (SVMs) there is the Recursive Feature Elimination (RFE-SVM) algorithm for gene selection. Recent studies have shown that a smaller value of C (i.e. greater regularisation) leads to lower error rates.

We will prove that under the high regularisation limit C->0, a linear SVM is equivalent to an centroid classifier. Furthermore, in the limit the weights of the features are independent of the other features and are explicitly known. As such, the need for multiple iterations in RFE disappears and the whole process of the recursive feature selection becomes trivial.

Experiments and Results:

The first experiment was to evaluate the performance of the relevance based feature selection. For the centroid algorithm, features were selected by ranking the features with respect to the relevance (absolute value of the weight) and choosing the required number of top features. This is equivalent to the feature selection method of the RFE-SVM as the weight vector of the centroid classifier is not affected by the features that are chosen.

For the first experiment, the centroid classifier is used with various different feature selection

methods (t-test, SNR, and relevance based feature selection). Each classifier is run with various numbers of features, ranging from 2 to all, in a logarithmic fashion.

We will present the results for this experiment on the colon cancer dataset. These results show that with the relevance based feature selection, performed the best at 64 genes with a minimum error of 11.7. The accuracy using all markers was 11.9, and SNR and t-test feature selection failed to improve this value. However, when using very small numbers of features (8 or less), SNR and t-test outperformed the relevance feature selection, with SNR performing the best.

We will also present results of the same experiment setup but with the lymphoma dataset. For this dataset, all feature selection achieved a minimal error rate of 0, however the SNR t-test feature selection achieved 0 with a minimum of 512 genes, and relevance feature selection achieved 0 with 256 genes (see Table). Again, at smaller numbers of features the SNR and t-test feature selection methods start to perform better than the relevance feature selection method.

The second experiment was to compare the centroid classifier with the RFE-SVM. Again, the same 50 permutation train/test split design as the previous experiment was used. The results of this experiment show that as C decreases, the performance of the RFE-SVM increases and approaches the performance of the centroid classifier. A similar result is observable on the same experiment run on the lymphoma dataset.

The results show that the nearest centroid classifier is not only faster but it has an accuracy that is at least as good as the competition. It is also apparent from the results that it has almost no overfitting problems when we let the number of features increase dramatically, which is a very attractive feature.

All of this adds up to the conclusion that the simplest form of a nearest centroid classifier should be considered as a simple baseline gene selection and classification method for cancer classification; it is extremely fast, easy to implement and its performance is strong and reliable.