Combining dissimilarity based classifiers for cancer prediction using gene expression profiles

Ángela Blanco, Manuel Martín-Merino and Javier de las Rivas
ablancogo@upsa.es

DNA Microarrays allow us to monitor the expression level of thousands of genes simultaneously across a collection of related samples. This technology has been applied to the prediction of cancer considering the gene expression profiles in both normal and cancerous samples.

Support Vector Machines (SVM) have been applied to identify cancerous tissues considering the gene expression levels with encouraging results. This kind of techniques are able to deal with high dimensional and noisy data which is an important requirement in our practical problem. However, common SVM algorithms rely on the use of the Euclidean distance which doesn't reflect accurately the proximities among the sample profiles [2]. This feature favors the misclassification of cancerous tissues (false negative errors) which is a serious drawback in our application. The SVM has been extended to incorporate non-Euclidean dissimilarities [4]. Nevertheless no dissimilarity can be considered superior to the others because each one reflects just different features of the data and misclassify a different set of patterns.

The false negative errors of individual classifiers can be reduced by combining non-optimal classifiers [3]. To this aim, different versions of the classifier are usually built by bootstrap sampling the patterns or the features. However, resampling techniques reduce the size of the training set increasing the bias of individual classifiers and consequently the error of the resulting combination [5].

In this paper we propose a combination strategy that builds the diversity of classifiers considering a set of dissimilarities that reflect different features of the data. In order to incorporate the dissimilarities into the SVM they are first embedded in an Euclidean space such that the inter-pattern distances reflect the original dissimilarity matrix. Next, for each dissimilarity a C-SVM is trained. Finally, the resulting classifiers are properly combined using a voting strategy. Our method is able to work directly from a dissimilarity matrix.

The algorithm proposed has been tested using two benchmark datasets, Leukemia [6] and Breast Cancer [1]. The experimental results show that our combination strategy reduces significantly the false negative errors of the best SVM classifier based on a single dissimilarity. Our method compares favorably with a well known combination strategy such as Bagging [5]. Finally, the combination strategy has been applied to the widely used k-Nearest Neighbor algorithm. False negative errors of the best single classifier are reduced and Bagging is particularly improved.

References

[1] T. R. Golub et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 286(15), 531-537, October 1999.

[2] D. Jiang, C. Tang and A. Zhang. Cluster Analysis for Gene Expression Data: A Survey. IEEE Transactions on Knowledge and Data Engineering, 16(11), 1370-1386, November 2004.

[3] J. Kittler, M. Hatef, P. W. Duin and J. Matas. On Combining Classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3), 226-239, March 1998.

[4] E. Pekalska, P. Paclick, and R. Duin, A generalized kernel approach to dissimilarity-based classification, Journal of Machine Learning Research, vol. 2, 175-211, 2001.

[5] G. Valentini and T. Dietterich, Bias-variance analysis of support vector machines for the development of svm-based ensemble methods, Journal of Machine Learning Research, vol. 5, 725-775, 2004.

[6] M. West et al. Predicting the Clinical Status of Human Breast Cancer by using Gene Expression Profiles. PNAS, 98(20), 11462-11467, September 2001.