Gene Selection on Micro Array Data through Support Vector Machines

Sepp Hochreiter and Klaus Obermayer
Department of Electrical Engineering and Computer Science
Technische Universität Berlin, 10587 Berlin, Germany

Introduction. Gene expression data sets typically consist of about 100 tissue samples each of which is described by the expression values of few 10,000 genes. Machine learning techniques require more examples (tissue samples) than features (genes) to guarantee sufficient generalization capability [5]. Therefore the selection of relevant genes is mandatory for processing micro array data. The quality of the final results strongly depend on the feature selection algorithms [1]. We present a new feature selection method where the selected feature vectors correspond to support vectors. Our approach is tested on three gene expression data sets where the outcome of a cancer therapy must be predicted. For all three cancer types (brain tumor, lymphoma, and breast cancer) we were able to considerably improve the prediction results compared to the results of standard methods published in NATURE and NATURE MEDICINE. The improvement of the results is important because for a negative therapy outcome prediction two policies are possible: an alternative treatment or a closer observation of the patient. Positive prediction can indicate a lower dosis of medication and, therefore, lower risk of side effects of the treatment.

Basic Idea. We consider the classification task for datasets which are described by matrices. Rows and columns of these matrices correspond to objects where row and column objects may be from different sets and column objects are labeled. Data matrix entries express relationships between row and column objects and are produced by an unknown kernel. This kernel represents a dot product in some (unknown) feature space. Therefore the unknown kernel defines two mappings: a mapping of the labeled column objects into the feature space and a mapping of the row objects (features) into the feature space. In this feature space a linear column object classifier should be constructed. However the dot products between column objects are not available. Therefore standard support vector techniques cannot be utilized. We derive a new objective function for model selection in such a feature space according to the principle of structural risk minimization. The normal vector is expanded with respect to row objects, i.e. the feature objects. For further details see [3].

Because row objects can be interpreted as features and our method assigns support vector weights to the row objects it can be used for feature selection. An additional constraint, which imposes sparseness on the row objects, enforces few selected features. An SMO technique allows us to implement an efficient algorithm to solve the optimization problem which consists of a quadratic matrix with dimension equal to the gene number (few 10.000).

Our method is different from the feature selection methods in [2, 8] which also relay on SVMs. We do not use a kernel (it is implicit given), support vectors do not correspond to labeled objects, and we use a different objective function in contrast to the usual classification margin.

Experiments and Results. We predicted the outcome of tumor treatment with chemo- or radiation therapy based on DNA microarray data published in [4, 6, 7]. Column objects are samples from tumors and row objects correspond to genes. For every sample-gene pair a snapshot of the level of gene expression was measured.

For the 60 brain tumor (medulloblastoma) examples 12 leave-one-out (LOO) errors [4] could be reduced by our method to 4-5 errors. We extracted 40-50 relevant genes. The lymphoma (diffuse large B-cell) data set contains 58 examples and the error rate was 14 LOO errors [6] using 13 genes. We obtained 12 LOO errors with 15 genes. The breast cancer data consists of gene expression data of 78 patients and 19 separate test patients. The classification performance is described by an ROC-curve because the results in [7] are given for a range of classification thresholds (no model selection for the threshold). The results in [7] lead to 2 errors on the test set, 0.77 area under the ROC-curve, a minimal LOO error of 20, and 70 selected genes. Our feature selection method resulted in 2 errors on the test set, 0.88 area under the ROC-curve, a minimal LOO error number of 12, and 30-40 selected genes. Our method clearly outperformed the standard feature selection approaches.

[1] R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[2] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[3] S. Hochreiter and K. Obermayer. Feature selection and classification on matrix data: From large margins to small covering numbers. In *NIPS 15*, 2003.

[4] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.

[5] B. Schölkopf and A. J. Smola. *Learning with kernels — Support Vector Machines, Reglarization, Optimization, and Beyond*. MIT Press, Cambridge, 2002.

[6] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, R. C. T. Aguiar J. L. Kutok, M. Gaasenbeek, M. Angelo, M. Reich, T. S. Ray G. S. Pinkus, M. A. Koval, K. W. Last, A. Norton, J. Mesirov T. A. Lister, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002.

[7] L. J. van't Veer, H, Dai, M. J. van de Vijver, A. A. M. Hart Y. D. He, M. Mao, H. L. Peterse, K. van der Kooy, A. T. Witteveen M. J. Marton, G. J. Schreiber, R. M. Kerkhoven, P. S. Linsley C. Roberts, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[8] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for SVMs. In *NIPS 12*, 2000.