Support Vector Machine Classification of Microarray Data

Sayan Mukherjee & Ryan Rifkin

Artificial Intelligence Laboratory and
The Center for Biological and Computational Learning
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Use the learning from examples paradigm to make class predictions and infer genes involved in these predictions from DNA microarray expression data. Specifically, we use a Support Vector Machine (SVM) classifier [6] to predict cancer morphologies and treatment success and determine the relevant genes in the inference.

Motivation:

Previous Work: A generic approach to classifying two types of acute leukemias was introduced in Golub et. al. [3]. SVM's have been applied to this problem [5] and also to the problem of predicting functional roles of uncharacterized yeast ORF's [1].

Approach: We used a SVM classifier to discriminate between two types of leukemia. The output of classical SVM's is a class designation ± 1. In this particular application it is important to be able to reject points for which the classifier is not confident enough. We introduced a confidence interval on the output of the SVM that allows us to reject points with low confidence values. It is also important in this application to infer which genes are important for the classification. We have preliminary results for a feature selection algorithm for SVM classifiers.

The SVM was trained on the 38 points in the training set and tested on the 34 points in the test set. Our results (see table 2 and figure (1)) are the best reported so far for this dataset.

genes	rejects	errors	confidence level	$\|d\|$
7129	3	0	$\sim 93\%$	.1
40	0	0	$\sim 93\%$	.1
5	3	0	$\sim 92\%$	.1

Table 2: Number of errors, rejects, confidence level, and the $|d|$ corresponds to the cutoff for rejection. The first case is with no feature selection the next two are with feature selection.

Confidence levels were computed as follows. We use the leave-one-out estimator on the training data to get 38 $|d|$ values, d is the real valued output of the SVM before the sign is taken. We then estimate the distribution function, $\hat{F}(|d|)$ from the $|d|$ values. The confidence level $C(|d|)$ is simply

$$C(|d|) = 1 - \hat{F}(|d|).$$

Feature selection is performed by iteratively minimizing the following two functionals. First the standard SVM functional is minimized:

$$-\sum_{i=1}^{\ell} \alpha_i + \frac{1}{2} \sum_{i,j}^{\ell} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{3}$$

subject to

$$C \geq \alpha_i \geq 0\,, \quad \sum_{i=1}^{\ell} \alpha_i y_i = 0.$$

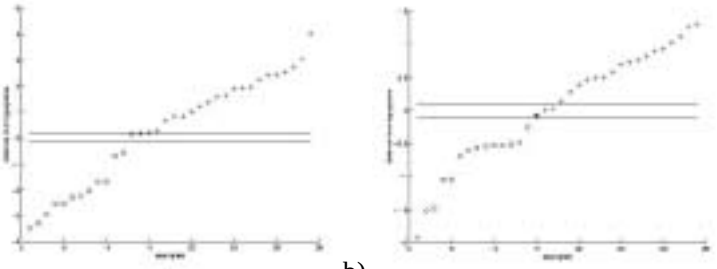
Then the following functional is minimized with respect to the diagonal matrix $\mathbf{P}$ (with elements p_f)

$$\frac{1}{2} \sum_{i,j}^{\ell} y_i y_j \alpha_i \alpha_j K(\mathbf{P}\mathbf{x}_i, \mathbf{x}_j) \tag{4}$$

subject to

$$p_f \geq 0\,, \quad \sum_{f=1}^{h} g(p_f) = N,$$

where N can be interpreted as the number of expected features and imposes a constant volume constraint. The function is determined by the properties of the mapping from input space to feature space [2].



a) b)

Figure 1: Plots of the distance from the hyperplane for test points (a) feature vector of 49 (b) feature vector of 7129. The $+$ are for class ALL, the o for class AML, the $*$ are mistakes, and the line indicates the decision boundary.

Impact: The problem of cancer classification has clear implications on cancer treatment. Additionally, the advent of DNA microarrays introduces a wealth of genetic expression information for many diseases including cancer. An automated or generic approach for classification of cancer or other diseases based upon the microarray expression will have a strong impact on disease treatment and diagnosis.

Future Work: Apply the SVM approach to other DNA microarray classification problems. Examine feature selection algorithms. Look at probabilistic kernels such as those proposed in [4].

Research Support: This report describes research done within the Center for Biological & Computational Learning in the Department of Brain & Cognitive Sciences and in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology.

This research was sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032

Additional support was provided by: Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., Komatsu Ltd., Merrill-Lynch, NEC Fund, Nippon Telegraph & Telephone, Siemens Corporate Research, Inc., Toyota Motor Corporation and The Whitaker Foundation.

References:

[1] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, Jr M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS*, 97(1):262–267, 2000.

[2] C.J.C Burges. *Geometry and Invariance in Kernel Based Methods.* M.I.T. Press, Cambridge, MA, 1999.

[3] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. 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*, 286:531–537, 1999.

[4] T. Jaakkola, M. Meila, and T. Jebara. Maximum entropy discrimination. In *Advances in Neural Information Processing Systems 12*. 2000. to appear.

[5] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. G olub, J.P. Mesirov, and T. Poggio. Support vector machine classification of microarray data. AI Memo 1677, Massachusetts Institute of Technology, 1999.

[6] V. N. Vapnik. *Statistical learning theory*. J. Wiley, 1998.