

# Classification of Cancer Tissue Types by Support Vector Machines Using Microarray Gene Expression Data

**Jinsong Cai<sup>1</sup>, Aynur Dayanik<sup>3</sup>, Hong Yu<sup>1</sup>, Naveed Hasan<sup>2</sup>, Tachio Terauchi<sup>2</sup>, William Noble Grundy<sup>2</sup>**

**<sup>1</sup>Department of Medical Informatics, <sup>2</sup> Department of Computer Science, Columbia University, New York, New York 10027**

**<sup>3</sup> Department of Computer Science, Rutgers University, Piscataway, NJ 08855  
*jic7001@flux.cpmc.columbia.edu, bgrundy@cs.columbia.edu***

## ABSTRACT

DNA microarray technology allows detection of the expression levels of thousands of genes at a time, and thus provides a new way to understand the differences in the global gene expression patterns between cancer and normal cells. In this report, we describe a method to classify cancer and normal tissues based on the gene expression patterns obtained from DNA microarray experiments, and to discover the genes whose expression levels had the most discrepancies between cancer and normal tissues. We compared two supervised machine learning techniques, the support vector machines (SVMs) and a decision tree algorithm (C4.5), in classification of cancer versus non-cancer tissues. In both leave-one-out and 10-fold cross-validation experiments, SVMs correctly classified ~99% of the tissue samples, while C4.5 was able to classify ~81% of tissue samples. Using the Fisher linear discriminant criterion, we identified a list of genes whose expression profiles had the best correlations with tissue types (cancer versus non-cancer). With only 100 of the total 4026 genes, SVMs were able to classify the cancer tissues with no loss of performance. Thus these genes were sufficient in classification of cancer tissues for this dataset, and can be used in microarray experiments for prediction of unknown tissue types.

## METHODS

A previous published dataset (Alizadeh et al., 2000) was used in this study (Figure 1). A total of 96 tissue samples (72 cancer and 24 non-cancer) were analyzed for gene expression patterns using 4026 different genes.

The SVMs and C4.5 algorithms were as previously described (Brown et al., 2000 and Quinlan, 1993, respectively). Because the tissue types were found to be linearly separable, the SVMs employed a first-degree dot product kernel function. Both SVMs and C4.5 were trained to discriminate between two classes of tissues, cancer and non-cancer, based on their gene expression profiles. The performance was

measured by error rate (false positives plus false negatives) in both leave-one-out and 10-fold cross-validation experiments.

Genes used in the microarray experiments were ranked based on their Fisher linear discriminant scores (Bishop, 1995). A subset of the genes with the highest scores was then used in place of the complete set of genes for classification by SVMs.

## RESULTS

We examined whether SVM and C4.5 were able to correctly classify tissue types based on gene expression data. In the leave-one-out experiments, SVMs correctly classified 95 out of the 96 samples, while C4.5 only correctly classified 78 out of 96 samples. Ten-fold cross-validation experiments yielded similar results.

After ranking the genes using the Fisher criterion, we selected 3 subsets that contained 100, 50, and 20 genes with the highest scores. When using the 100 genes obtained from the Fisher criterion as the input vector, SVMs had the same performance as with the original 4026 genes. The performance deteriorates slightly as the number of genes used was decreased. In each case, the performance of SVMs with genes selected using the Fisher criterion was significantly better than that with randomly selected genes.

## REFERENCES

Alizadeh A.A., et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature* 403: 503-11.

Bishop C. (1995) *Neural Networks for Pattern Recognition*. Oxford UP, Oxford, UK.

Brown, M.P.S., et al. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* 97 (1): 262-267.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. MK, San Mateo, CA.