

# Using Support Vector Machines for analysis of gene expression data from DNA microarrays

K. Fajarewicz, Silesian University of Technology

A. Swierniak, Silesian University of Technology, Gliwice, Poland, currently c

B. Jarzab, M. Wiench, Centre of Oncology - Institute of Oncology, Branch Gliwice, Poland

M. Kimmel, Rice University, Houston, Texas

## Abstract

DNA microarrays (biochips) are a new tool that biologists can use to obtain information about expression levels of thousands of genes simultaneously. Their main advantages are: reproducibility and scalability of obtained data, short time of one experiment and, of course, the large number of genes, the expression of which is measured. The technique of producing DNA microarrays is improving continuously.

In general, there are two different types of DNA microarrays: spotted microarrays and oligonucleotide microarrays. There are several important differences between these two types of microarrays. One of them is the technology of the production. While spotted microarrays are obtained by using special spotting robots, oligonucleotide microarrays are synthesized, often using photolithographic technology – the same as used during production of computer chips.

There are many ways of exploiting a data from microarrays. One of the most frequently used is the classification of samples belonging to different classes. Such classification can be applied for example in medical diagnosis and choosing proper medical therapy. One of the first paper dealing with classification was the article by Golub et al. (1999). In this paper samples of two types: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) were classified and clustered. For classification purposes the authors proposed so called weighted voting (WV) algorithm. The AML/ALL data set (available via Internet) was used by other scientists for testing different methods of analysis. For example, the same data set has been used for testing a more traditional perceptron algorithm in (Fajarewicz et al., 2000, 2001). Obtained results were slightly better than using WV algorithm. In (Furey et al., 2000) relatively new and promising method of classification and regression called support vector machines (Boser et al., 1992; Vapnik, 1995; Christianini et al., 2000) has been applied to the same

data set. In (Brown et al., 2000) the SVM technique has been tested on another microarray data set. Moreover, in this work SVM have been compared to other methods like: decision trees, Parzen windows, Fisher's linear discriminant and the conclusion was that SVM significantly outperformed all other investigated method. Therefore, the SVM technique can be regarded as a very promising supervised learning tool dealing with microarray gene expression data.

Choosing proper learning and classification method is a final and very important element in the recognition process when dealing with gene expression data. However there are other earlier stages of data processing, which are also very important because of their significant influence on classification quality. One of these elements is the gene selection. In (Golub et al., 1999) the method called neighborhood analysis (NA) has been used while in (Fujarewicz et al., 2000, 2001) the Sebestyen criterion (1962) modified by Deuser (1971) has been applied. In both methods a performance index evaluating discriminant ability is calculated separately for each gene. After this, a set of  $n$  genes with the highest index value is chosen for learning and classification purposes. Such approach seems reasonable. However, it may not be the best way of choosing a working gene set. This is due to the fact that expression levels of different genes are strongly correlated and univariate approach to the problem is not the best way. On the other hand, in case of microarray gene expression data, a naive approach to the problem by checking all subsets of thousands of genes is impossible due to the high computational cost.

Recently several new multivariate methods of choosing optimal (or suboptimal) gene subset have been proposed. Szabo et al. (2002) proposed a method that uses so called  $v$ -fold cross-validation combined with arbitrary chosen method of feature selection. In an approach formulated in (Chilingaryan et al., 2002) the Mahalanobis distance between vectors of gene expression is used to iterative improvement of actual gene subset. Another algorithm, combining genetic algorithms with  $k$ -nearest neighbor, has been proposed by Li et al. (2001).

In (Fujarewicz et al., 2003) a new method, called recursive feature replacement (RFR) for gene selection has been proposed. The RFR method uses SVM technique and iteratively optimizes the leave-one-out cross-validation error. The comparison of the RFR method to other algorithms such as: NA algorithm and proposed in papers (Szabo et al., 2002) and (Chilingaryan et al., 2002) showed a supremacy of the RFR method.

Recently a new method for gene selection, also based on SVM, has been published in (Guyon et al., 2002). The method, called recursive feature elimination (RFE), also outperformed other investigated methods.

One of benchmark data sets, frequently used for testing different methods of gene expression data processing, is the tumor/normal colon data set. This data set was presented and analyzed (clustered) in the paper (Alon et al., 1999). Expression levels of about 6500 genes were measured for 62 samples: 40 tumor

and 22 normal colon tissues. 2000 of them were selected by the authors for clustering/classification purposes. The main result of the paper (Alon et al., 1999) was the clustering experiment of the data. The data was grouped into two clusters with 8 wrong assignments: three normal tissues were assigned to the "tumor" cluster and five tumor tissues were assigned to the "normal" cluster. In (Furey et al., 2000) the SVM technique was used to classify the same data set. The classification was performed twice: for whole data set (2000 genes) and for top 1000 genes. In both cases the result of leave-one-out cross-validation was six misclassifications (3 tumor and 3 normal). Nguyen et al. (2002) tested on the colon data set two methods of data selection: principal component analysis (PCA) and partial least squares (PLS) and two methods of classification: logistic discrimination (LD) and quadratic discriminant analysis (QDA). Best results were obtained after applying LD classification to first 50 and 100 components (linear combinations of gene expression vectors) given by PLS method. Unfortunately there were still four misclassifications obtained in leave-one-out cross-validation.

In our work we have compared RFR, RFE, NA and pure Sebestyen methods to the tumor/normal colon and thyroid data sets. The comparison of obtained results shows that RFR method finds the smallest gene subset that gives no misclassifications in leave-one-out cross-validation.

### ***Acknowledgement***

*This work has been supported by KBN (Polish Scientific Committee) under a grant no. PBZ KBN-040/P04/08 and partially by the National Science Foundation under Agreement No. 0112050.*

### **References**

- [1] Alon U., N. Barai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine (1999): Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. — Proc. Natl. Acad. Sci., Vol.96, pp.6745–6750.
- [2] Boser B. E., I. M. Guyon, V. Vapnik (1992): A training algorithm for optimal margin classifiers. — Fifth Annual Workshop on Computational Learning Theory, Pittsburgh.
- [3] Brown M. P .S., W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares Jr, D. Haussler (2000): Knowledge based analysis of microarray gene expression data by using support vector machines. — Proc. of the National Academy of Sciences, Vol.97, no.1, pp. 262–267.
- [4] Chilingaryan A. , N. Gevorgyan, A. Vardanyan, D. Jones, A. Szabo (2002): A multivariate approach for selecting sets of differentially expressed genes. — Mathematical Biosciences, Vol. 176, pp. 59–69.

- [5] Christianini N., J. Shawe-Taylor (2000): An introduction to support vector machines and other kernel-based learning methods. — Cambridge Univ. Press.
- [6] Deuser L. M. (1971): A hybrid multispectral feature selection criterion. — IEEE Trans. on Comp., pp.1116–1117.
- [7] Fujarewicz K., J. Rzeszowska-Wolny (2000): Cancer classification based on gene expression data. — Journal of Medical Informatics and Technologies, Vol.5, pp.BI23–BI27.
- [8] Fujarewicz K., J. Rzeszowska-Wolny (2001): Neural network approach to cancer classification based on gene expression levels. — Proc. IASTED Int. Conf. Modelling Identification and Control, Innsbruck pp.564–568.
- [9] Fujarewicz K., M. Kimmel, J. Rzeszowska-Wolny, A. Swierniak (2003): A note on classification of gene expression data using support vector machines. — Journal of Biological Systems, Vol.11, No.1, pp.43–56.
- [10] Furey T.S., N. Christianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler (2000): Support vector machine classification and validation of cancer tissue samples using microarray expression data. — Bioinformatics, Vol.16, no.10, pp.906–914.
- [11] Golub T. R., T.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, J.R. Downing, M.A. Caliguri, C.D. Bloomfield, E.S. Lander (1999): Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. — Science, Vol.286, pp.531–537.
- [12] Guyon I., J. Weston, S. Barnhill, V. Vapnik (1999): Gene Selection for Cancer Classification using Support Vector Machines. — Machine Learning, Vol. 64, pp. 389–422.
- [13] Li L., C.R. Weinberg, T.A. Darden, L.G. Pedersen (2001): Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. — Bioinformatics, Vol.17, pp.1131–1142.
- [14] Nguyen D.V., D.M. Rocke (2002): Tumor classification by partial least squares using microarray gene expression data. — Bioinformatics, Vol.18, no.1, pp.39–50.
- [15] Sebestyen G. S. (1962): Decision making processes in pattern recognition. — Macmillan, New York.
- [16] Szabo A., K. Boucher, W.L. Carroll, L.B. Klebanov, A.D. Tsodikov, A.Y. Yakovlev (2002): Variable selection and pattern recognition with gene expression data generated by the microarray technology. — Mathematical Biosciences, Vol.176, pp.71–98.
- [17] Vapnik V. (1995): The Nature of Statistical Learning Theory. — Springer-Verlag, New-York.