Int. J. Appl. Math. Comput. Sci., 2003, Vol. 13, No. 3, 327–335

SELECTING DIFFERENTIALLY EXPRESSED GENES FOR COLON TUMOR CLASSIFICATION

KRZYSZTOF FUJAREWICZ*, MAŁGORZATA WIENCH**

* Institute of Automatic Control
Silesian University of Technology
ul. Akademicka 16, 44–100 Gliwice, Poland
e-mail: kfujarewicz@ia.polsl.gliwice.pl

** Department of Nuclear Medicine and Endocrine Oncology
Centre of Oncology, Maria Skłodowska-Curie Memorial Institute
44–101 Gliwice, Poland
e-mail: wiench@io.gliwice.pl

DNA microarrays provide a new technique of measuring gene expression, which has attracted a lot of research interest in recent years. It was suggested that gene expression data from microarrays (biochips) can be employed in many biomedical areas, e.g., in cancer classification. Although several, new and existing, methods of classification were tested, a selection of proper (optimal) set of genes, the expressions of which can serve during classification, is still an open problem. Recently we have proposed a new recursive feature replacement (RFR) algorithm for choosing a suboptimal set of genes. The algorithm uses the support vector machines (SVM) technique. In this paper we use the RFR method for finding suboptimal gene subsets for tumor/normal colon tissue classification. The obtained results are compared with the results of applying other methods recently proposed in the literature. The comparison shows that the RFR method is able to find the smallest gene subset (only six genes) that gives no misclassifications in leave-one-out cross-validation for a tumor/normal colon data set. In this sense the RFR algorithm outperforms all other investigated methods.

Keywords: colon tumor, gene expression data, microarrays, support vector machines, feature selection, classification

1. Introduction

DNA microarrays (biochips) constitute a new tool which can be used by biologists to obtain information about expression levels of thousands of genes simultaneously. Their main advantages are: the reproducibility and scalability of the obtained data, a short time of one experiment and, of course, a large number of genes whose expression 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 production. While spotted microarrays are obtained by using special spotting robots, oligonucleotide microarrays are synthetized, often using photolitographic technology (the same as used during the production of computer chips).

There are many ways of exploiting data from microarrays. One of the most frequently used manners is the classification of samples belonging to different classes. Such a classification can be applied, e.g., to medical diagnosis and choosing a proper medical therapy. One of the first papers dealing with the problem of classification was the one by Golub *et al.* (1999). In this paper samples of two types: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) were classified and clusterized. For classification purposes the authors proposed the so-called weighted voting (WV) algorithm. The AML/ALL data set (available via the Internet) was used by other scientists for testing different analysis methods. For example, the same data set was used for testing a more traditional perceptron algorithm in (Fujarewicz and Rzeszowska-Wolny, 2000; 2001). The obtained results were slightly better than those obtained using the WV algorithm. In (Furey *et al.*, 2000) a relatively new and promising method of classification and regression called the support vector machines (Boser *et al.*, 1992; Vapnik, 1995; Christianini and Shawe-Tylor, 2000) was applied to the same data set. In (Brown *et al.*, 2000) the SVM technique was tested on another microarray data set. Moreover, in this work the SVM approach was compared with other methods such as decision trees, Parzen windows,

Fisher's linear discriminant, and the conclusion was that the SVM significantly outperformed all other investigated methods. Therefore the SVM technique can be regarded as a very promising supervised learning tool dealing with microarray gene expression data.

Choosing a 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 the classification quality. One of these elements is gene selection. In (Golub *et al.*, 1999) a method called the neighborhood analysis (NA) was used, while in (Fujarewicz and Rzeszowska-Wolny, 2000; 2001) Sebestyen's criterion (1962) modified by Deuser (1971) was 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 an 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 a univariate approach to the problem is not the best way. On the other hand, in the case of microarray gene expression data, a naive approach to the problem by checking all subsets of thousands of genes is impossible due to a high computational cost.

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

In (Fujarewicz *et al.*, 2003) a new method called the recursive feature replacement (RFR) for gene selection was proposed.[1] The RFR method uses the SVM technique and iteratively optimizes the leave-one-out cross-validation error. The comparison of the RFR method with other algorithms such as the NA algorithm and those proposed in the papers (Szabo *et al.*, 2002; Chilingaryan *et al.*, 2002) showed the superiority of the RFR method.

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

One of benchmark data sets which are frequently used for testing various 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).[2] 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 were 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 the 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 ones). Nguyen and Rocke (2002) tested two methods of data selection on the colon data set: principal component analysis (PCA) and partial least squares (PLS), and two methods of classification: logistic discrimination (LD) and quadratic discriminant analysis (QDA). The best results were obtained after applying LD classification to the first 50 and 100 components (linear combinations of gene expression vectors) given by the PLS method. Unfortunately, there were still four misclassifications obtained in leave-one-out cross-validation.

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

The paper is organized as follows. In Section 2 we present methods we used for colon data preprocessing and preselection. The RFR method of gene selection is described in Section 3. Finally, Sections 4 and 5 present results and conclusions. In addition, because the RFR method uses the SVM technique, the latter is briefly described in Appendix.

2. Data Preprocessing and Preselection

The tumor/normal data set contains expression levels of 2000 genes measured using Affymetrix oligonucleotide microarrays for 62 samples (40 tumor and 22 normal) of colon tissues. The data set can be expressed in the form of a matrix composed of 62 column vectors

$$\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, , \mathbf{x_N}]. \tag{1}$$

Each vector $\mathbf{x}_i$ has 2000 elements, so that the dimension of the matrix $\mathbf{X}$ is 2000×62. The data were preprocessed using the following steps. Firstly, the data were

[1] In (Fujarewicz *et al.*, 2003) the name RFR was not used.

[2] The tumor/normal colon data set is freely obtainable on the web site `http://microarray.princeton.edu/oncology/affydata/`

log transformed (base 10). Then all columns and rows were normalized. Normalization consists in the substraction of the mean and division by the standard deviation of an appropriate vector. Finally, we applied the "squashing function" suggested in (Guyon *et al.*, 2002):

$$f(x) = \operatorname{atan}\left(\frac{x}{c}\right) \tag{2}$$

in order to reduce the influence of outliers. We chose the value of $c = 0.1$.

After the preprocessing stage the data were preselected using the modified Sebestyen criterion (Deuser, 1971; Sobczak i Malina, 1978). The modified Sebestyen criterion for a given subset of genes Γ and for a more general case of L classes is given by the formula

$$G_\Gamma = \frac{1}{L}\sum_{i=1}^{L}\left\{\frac{1}{L-1}\sum_{j=1,j\neq i}^{L} S_\Gamma(i,j) - \frac{m_i}{m_i-1}S_\Gamma(i,i)\right\}, \tag{3}$$

where m_i is the number of samples representing the i-th class and

$$S_\Gamma(i,j) = \frac{1}{m_i m_j}\sum_{p\in\Omega_i}\sum_{t\in\Omega_j}\left\|\mathbf{x}_p^\Gamma - \mathbf{x}_t^\Gamma\right\|^2 \tag{4}$$

is a measure of separation between classes i and j. In (4) the symbol $\|\cdot\|$ stands for the Euclidean norm, Ω_i is the set of the indices of samples from the i-th class and $\mathbf{x}_p^\Gamma$ is the vector of expression levels of the p-th sample for the gene subset Γ. The main advantage of the criterion (3) is its additivity property:

$$G_\Gamma = \sum_{k\in\Gamma} G_k, \tag{5}$$

where G_k is calculated using (3) for only one feature. In this case Eqn. (4) takes the form

$$S_k(i,j) = \frac{1}{m_i m_j}\sum_{p\in\Omega_i}\sum_{t\in\Omega_j}(x_p^k - x_t^k)^2. \tag{6}$$

Formula (6) is a measure of separation between classes i and j along the k-th axis (x_p^k is the expression level of the k-th gene in the p-th sample).

Hence, in order to find the best subset of l genes, i.e., that with the highest value of (3), one has to take simply l genes with the highest values of G_k calculated separately for each gene. The values of G_k for all 2000 genes are presented in Fig. 1. For the future calculations, we chose the first 300 genes.

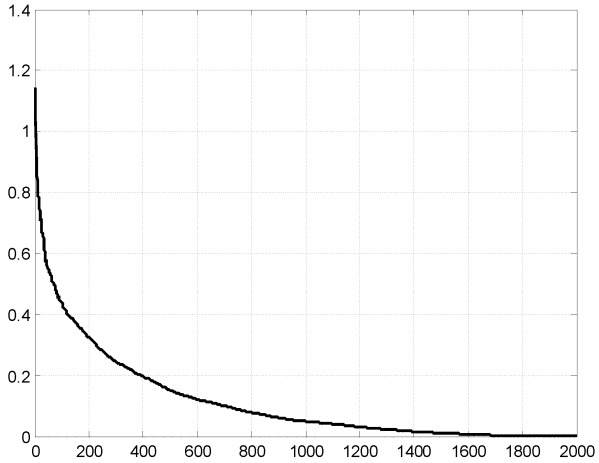

Fig. 1. Sebestyen criterion G_k calculated for all 2000 genes separately and drawn in descending order.

3. Recursive Feature Replacement (RFR) Method

In this section the RFR method (Fujarewicz *et al.*, 2003) will be described. At the beginning, two performance indices used for evaluating the classification quality for a particular gene subset will be described. Both indices are based on the result of leave-on-out cross-validation but they use different mathematical formulas.

3.1. Performance Index

3.2. Evaluation of Gene Subset Generalization Ability

The fact that is worth recalling here is that the aim of constructing a recognition system is not to perfectly separate the training set. The main aim is to find the feature set (a gene set in our application), the form of the classifying function and the learning algorithm, for which the samples not being used during the learning phase are classified correctly. In other words, the learning machine should be characterized by a good generalization ability.

In general, in the leave-one-out cross-validation method one vector $\mathbf{x}_k$ is removed from the training set and the remaining vectors serve during the learning phase. After this it is checked how the removed vector is classified. In the RFR method the SVM technique is used for finding a linear classification rule (see Appendix). If the SVM is used, the leave-one-out cross-validation method can be formally stated as follows:

1. Remove one vector $\mathbf{x}_k$ from the training set.
2. For the remaining vectors calculate $\mathbf{w}^o$ and b^o using the SVM method (see Appendix, Eqns. (22) and (23)).
3. For the removed vector $\mathbf{x}_k$ calculate the function
$$f_{\text{norm}}(\mathbf{x}_k) = \frac{d_k}{\|\mathbf{w}^o\|}({\mathbf{w}^o}^T\mathbf{x}_k + b^o). \tag{7}$$
4. Repeat Steps 1–3 for $k = 1, 2, \ldots, N$.

In (7), d_k is the target output (14), see Appendix. The symbol $\|\cdot\|$ denotes the Euclidean norm. Owing to the division by the norm of $\mathbf{w}^o$ the absolute value of (7) is equal to the Euclidean distance between the decision surface and the vector $\mathbf{x}_k$. This is because after this normalization the norm of the gradient of the function (7) is equal to 1. The positive value of (7) indicates that the vector $\mathbf{x}_k$ is correctly classified.

As has been mentioned above, we use two different performance indices based on all values of (7) calculated for all samples.

The first index is a simple percentage index which takes into account how many samples are correctly classified in leave-one-out cross-validation:

$$J_{\mathrm{cv1}} = \frac{N_{\mathrm{corr}}}{N} \cdot 100\%, \tag{8}$$

where N_{corr} is the number of positive values of (7). The second performance index is based only on the worst (minimal) value among all values of (7):

$$J_{\mathrm{cv2}} = \frac{1}{\sqrt{n}} \min_k f_{\mathrm{norm}}(\mathbf{x}_k). \tag{9}$$

In (9), the result is divided by $\sqrt{n}$ in order to make the results comparable for training sets with different numbers of genes n. High values of (8) and (9) indicate a good generalization ability. If the performance index (9) is positive, then all samples during leave-one-out cross-validation are classified correctly.

Note that the cross-validation method evaluates the generalization ability of the whole recognition system. Since in our approach the form of the discriminant function and the learning algorithm are fixed, the outcome of the cross-validation method presented here depends only on the way of selecting the gene set. Moreover, for a fixed gene subset this outcome is unique because both the method of cross-validation and the SVM technique give unique results.

Let us denote by Ω the set of numbers of all measured genes $\Omega = \{1, 2, \ldots, N\}$, and by $\Omega^* \subset \Omega$ any of its subsets. The symbols

$$J_{\mathrm{cv1}}(\Omega^*) \tag{10}$$

and

$$J_{\mathrm{cv2}}(\Omega^*) \tag{11}$$

will respectively denote the values of the performance indices (8) and (9) calculated for the gene subset $\Omega^\star$.

3.3. Algorithm

As has been mentioned in the Introduction, due to a high computational cost, it is impossible to examine all subsets of thousands of genes the expressions of which are measured using microarrays. Therefore the RFR algorithm uses a heuristic rule, where the subset of genes $\Omega^\star$ is modified in successive iterations so that the value of the performance index increases. Since the performance index (8) takes only discrete values, the second performance index (9), which is real valued, is used. The algorithm consists in performing the following steps:

1. Read an initial subset $\Omega^* \subset \Omega$.
2. Find the single gene of the number $k \in \Omega^\star$ that maximizes $J_{\mathrm{cv2}}(\Omega^* \backslash \{k\})$.
3. Find the single gene of the number $l \in \Omega \backslash \Omega^\star$ that maximizes $J_{\mathrm{cv2}}(\Omega^* \cup \{l\})$.
4. If $J_{\mathrm{cv2}}((\Omega^\star \backslash \{k\}) \cup \{l\}) > J_{\mathrm{cv2}}(\Omega^\star)$, then $\Omega^* := (\Omega^* \backslash \{k\}) \cup \{l\}$, and go to Step 2.
5. Stop.

Note that the number of genes n in the subset $\Omega^\star$ does not change, so the algorithm has to be run for every $n = 2, 3, \ldots, M-1$, where M is the number of all genes. As a starting gene subset for n we choose an optimal gene subset obtained for $n-1$ supplemented by one of the remaining genes with the best modified Sebestyen criterion (as described in the previous section). As the first optimal one-element gene subset we choose simply the gene which maximizes (9).

4. Results

We implemented and applied to the tumor/normal colon data set the following four methods: recursive feature replacement (RFR), recursive feature elimination (RFE), neighborhood analysis (NA), and the pure Sebestyen method. The NA method and the Sebestyen criterion were applied to the entire 2000 gene data set, while RFR and RFE were applied to the set of the first (best) 300 genes previously preselected by the Sebestyen criterion as mentioned in Section 2.

In Fig. 2 the value of the performance index (9) calculated for the first 30 gene subset obtained using different methods is presented. It can be easily seen that the RFR and RFE methods are superior to the NA and Sebestyen methods. For small gene subsets the performance index calculated using the RFR method grows faster and starts to be positive for the subset of only six genes. The performance index reaches first a local maximum for the subset of ten genes which are listed in Table 1. The suboptimal subset of six genes, for which the performance index starts to be positive, is a subset of a larger 10-element suboptimal subset and it is listed on top in Table 1. It is not a

general rule of the RFR method but in the case of this particular colon data set it appeared to be true.

For larger gene subsets it is less than the values of the performance index in the RFE method but still remains positive. This means that all samples are classified correctly in leave-one-out cross-validation, see Fig. 3. Hence from a practical point of view this difference is not very important.

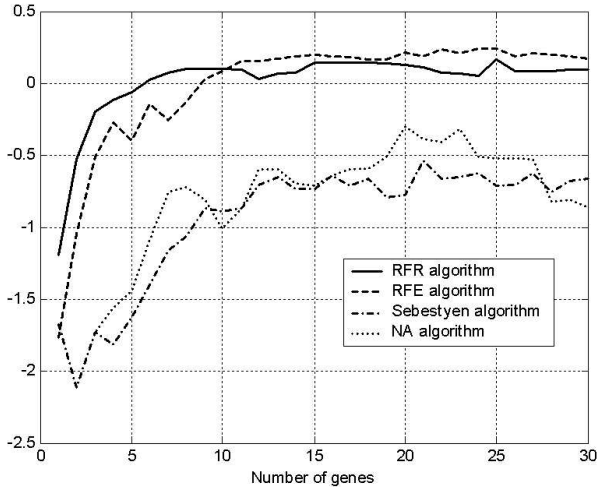

Fig. 2. Performance index J_{cv2} calculated for the first 30 gene subsets obtained using different methods of data selection.

Table 1. Set of 10 genes selected using the RFR method.

Gene number	Gene name
H06524	Gelsolin Precursor, Plasma (Human)
M82919	Human gamma amino butyric acid (GABAA) – beta-3 receptor – subunit mRNA, complete cds.
T59878	Peptidyl-Prolyl Cis-Trans Isomerase B Precursor (HUMAN)
H64807	Placental Folate Transporter (Homo sapiens)
M36634	Human vasoactive intestinal peptide (VIP) mRNA, complete cds.
X12369	Tropomyosin Alpha Chain, Smooth Muscle (HUMAN)
T50797	Deoxyridine 5'-Triphosphate Nucleotidohydrolase (Human)
X15880	Human mRNA for collagen VI alpha-1 C-terminal globular domain
R75843	Translational Initiation Factor 2 Gamma Subunit (Homo sapiens)
M58050	Human membrane cofactor protein (MCP) mRNA, complete cds.

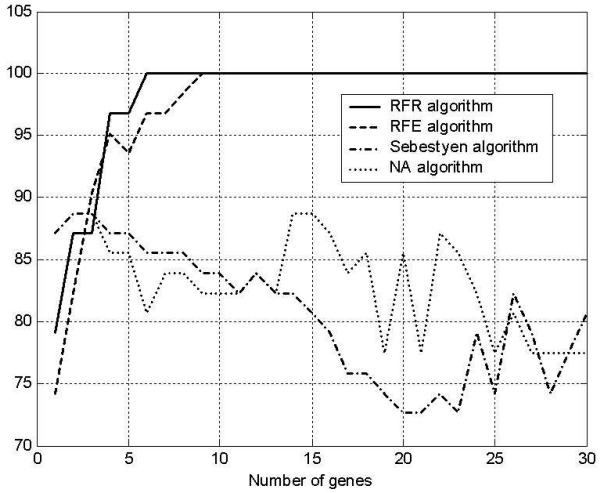

Fig. 3. Performance index J_{cv1} calculated for the first 30 gene subsets obtained using various methods of data selection.

In this article we use an approach to the evaluation of the generalization ability of a gene subset based on leave-one-out cross-validation in the standard meaning, where one sample is removed in one learning-classification cycle. Such an approach was used in all previously cited works devoted to gene selection. An approach where one sample is removed in one selection-learning-classification cycle seems to be quite reasonable, although it requires much harder computational efforts.

The ten genes separated by the RFR method reflect various cellular mechanisms. Most of them were previously observed to be associated with the colon cancer. The vasoactive intestinal peptide (VIP), which plays an important role as a neuroendocrine mediator in the secretion of water and electrolytes in the gut, is suggested to promote the growth and proliferation of tumor cells. Patients with colorectal cancer occurred to have an elevated serum level of VIP and a high density of VIP receptors in cancer cells (Hejna *et al.*, 2001). It became the base of performing scintigraphy with [123I]VIP radioligand, which was concluded to be a sensitive method for radioimaging colorectal cancer (Raderer *et al.*, 1998). The increase in the dUTPase (deoxyuridine 5'-triphosphate nucleotidohydrolase) activity takes place during mitogenic stimulation and the proliferative stage of the cell. Colorectal adenomas and adenocarcinomas revealed a wide spectrum of dUTPase expressions and its high level may be a negative prognostic marker (Fleishmann *et al.*, 1999). It predicts tumor resistance to chemotherapy, a shorter time to progression and a shorter overall survival (Ladner *et al.*, 2000). Two other genes are engaged in the immune response. The peptidyl-prolyl isomerase-B precursor (cyclophilin B) is involved in T cell activation and its expression is observed

in tumor infiltrating lymphocytes in many types of cancer (e.g., colon cancer) (Gomi *et al.*, 1999; Tamura *et al.*, 2001). The membrane cofactor protein (MCP, CD46) is one of the proteins protecting cells, also tumor cells, from the lysis by an activated complement (Jurianz *et al.*, 1999). Expressions of complement-regulatory proteins are often deregulated in cancer, which results in tumors that are resistant to an attack by complement. The MCP is highly expressed by glandular epithelium of human breast and colorectal tumor tissues and represents a possible mechanism of the tumor escape (Thorsteinsson *et al.*, 1998; Schmitt *et al.*, 1999). On the other hand, the underexpression of gelsolin was observed in cancer cells. Gelsolin is a multifunctional actin-binding protein which acts as both a regulator and an effector of apoptosis (Kwiatkowski, 1999). It is downregulated in several types of tumors and its abnormal expression is among the most common defects found in human breast, gastric, bladder and colon cancer (Porter *et al.*, 1993; Winston *et al.*, 2001; Rao, 2002). Also, the loss of basement membrane components, such as type IV collagen, has been demonstrated in colorectal cancer (Galbavy *et al.*, 2002; Oka *et al.*, 2002). It is related to the loss of differentiation and the malignant potential of epithelial tumors of the colon. The tropomyosin alpha chain (smooth muscle) represents the cluster of muscle genes mentioned by Alon *et al.* (1999) as those that differentiate between tumors and normal tissues. It is due to a high muscle content in the normal samples. Similarly, the decrease in the GABAA receptor expression may be due to the lack of an innervated circular muscle strip in a tumor tissue (Grider and Makhlouf, 1992). Although there has been no evidence so far of association between cancerogenesis and translational initiation factor 2 gamma subunit (eIF-2 gamma), other translation components such as eIF-4 and eIF-2 alpha have been reported to be overexpressed in human tumors, including colorectal cancer (Lobo *et al.*, 2000; reviewed in: Dua *et al.*, 2001). The role of the gene number H64807 is not very clear and it needs further research.

5. Conclusion

In this article the problem of finding differentially expressed genes for the tumor/normal classification of colon tissues has been investigated. The data set consisted of gene expression profiles of 2000 genes measured for 62 colon tissues (40 tumor and 22 normal) using Affymetrix DNA microarrays. Four methods of gene selection: recursive feature elimination, recursive feature replacement, neighborhood analysis and the pure Sebestyen criterion were used. The comparison showed that the RFE and RFR methods worked much better than the two other investigated methods.

The results of leave-one-out cross-validation obtained for the RFE and RFR methods showed that the RFR method gives better values of the performance index (9) for a smaller gene subset while the RFE method is slightly better for larger gene subsets. This phenomenon is probably related to the nature of both the methods. In the RFR method we start with a one-element gene subset and in successive runs of the algorithm the suboptimal *k*-element gene subsets are reached starting from a previously found $(k-1)$-element suboptimal gene subset. The occurrence of local maxima is the reason why a global optimum is not reached. On the other hand, in previous iterations the RFE eliminates (for larger gene subsets) the genes which could be useful in smaller gene sets.

The ten genes selected by the RFR method for which the leave-one-out cross-validation performance index reached an optimal value were listed and analyzed. Most of them have been previously reported to be associated with colon cancer.

Acknowledgement

The work has been partly supported by the grant of the State Committee for Scientific Research in Poland (KBN) No. PBZ KBN-040/P04/08 in 2003, and partly by the NATO grant LST.CLG.977845 and the NIH grant CA 84978 during the first author's visit at the Department of Statistics, Rice University, Houston, TX. The authors wish to express their gratitude to Professor Marek Kimmel for several helpful comments and suggesting references.

References

Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D. and Levine A.J. (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.

Boser B.E., Guyon I.M. and Vapnik V. (1992): *A training algorithm for optimal margin classifiers*. — Proc. 5-th Ann. Workshop *Computational Learning Theory*, Pittsburgh, pp. 144–152.

Brown M.P.S., Groundy W.N., Lin D., Cristianini N., Sugnet C.W., Furey T.S., Ares Jr M. and Haussler D. (2000): *Knowledge based analysis of microarray gene expression data by using support vector machines*. — Proc. Nat. Acad. Sci., Vol. 97, No. 1, pp. 262–267.

Chilingaryan A., Gevorgyan N., Vardanyan A., Jones D. and Szabo A. (2002): *A multivariate approach for selecting sets of differentially expressed genes*. — Math. Biosci., Vol. 176, pp. 59–69.

Christianini N. and Shawe-Tylor J. (2000): *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. — Cambridge: Cambridge Univ. Press.

Deuser L.M. (1971): *A hybrid multispectral feature selection criterion*. — IEEE Trans. Comp., pp. 1116–1117.

Dua K., Williams T.M. and Beretta L. (2001): *Translational control of the proteome: relevance to cancer*. — Proteomics, Vol. 1, pp. 1191–1199.

Fleishmann J., Kremmer E., Muller S., Sommer P., Kirchner T., Niedobitek G. and Grasser F.A. (1999): *Expression of deoxyuridine triphosphatase (dUTPase) in colorectal tumour*. — Int. J. Cancer, Vol. 84, pp. 614–617.

Fujarewicz K. and Rzeszowska-Wolny J. (2000): *Cancer classification based on gene expression data*. — J. Med. Inf. Technol., Vol. 5, pp. BI23–BI27.

Fujarewicz K. and Rzeszowska-Wolny J. (2001): *Neural network approach to cancer classification based on gene expression levels*. — Proc. IASTED Int. Conf. *Modelling Identification and Control*, Innsbruck, Austria, pp. 564–568.

Fujarewicz K., Kimmel M., Rzeszowska-Wolny J. and Swierniak A. (2003): *A note on classification of gene expression data using support vector machines*. — J. Biol. Syst., Vol. 11, No. 1, pp. 43–56.

Furey T.S., Christianini N., Duffy N., Bednarski D.W., Schummer M. and Haussler D. (2000): *Support vector machine classification and validation of cancer tissue samples using microarray expression data*. — Bioinformatics, Vol. 16, No. 10, pp. 906–914.

Galbavy S., Lukac L., Porubsky Y., Cerna M., Labuda M., Kmet'ova J., Papincak J., Durdik S. and Jakubowsky J. (2002): *Collagen type IV in epithelial tumours of colon*. — Acta Histochem 2002, Vol. 104, pp. 331–334.

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

Gomi S., Nakao M., Niija F., Imamura Y., Kawano K., Nishizaka S., Hayashi A., Sobao Y., Oizumi K. and Itoh K. (1999): *A cyclophilin B gene encodes antigenic epitopes recognized by HLA-A24-restricted and tumor-specific CTLs*. — J. Immunol., Vol. 163, pp. 4994–5004.

Grider J.R. and Makhlouf G.M. (1992): *Enteric GABAA: Mode of action and role in the regulation of the peristaltic reflex*. — Am. J. Physiol., Vol. 262, pp. G690–694.

Guyon I., Weston J., Barnhill S. and Vapnik V. (2002): *Gene selection for cancer classification using support vector machines*. — Mach. Learn., Vol. 64, pp. 389–422.

Haykin S. (1999): *Neural Networks—A Comprehensive Foundation (2nd Ed.)*. — Upper Saddle River, NJ: Prentice-Hall.

Hejna M., Hamilton G., Brodowicz T., Haberl I., Fiebiger W.C., Scheithauer W., Virgolin I., Kostler W.J., Oberhuber G. and Raderer M. (2001): *Serum levels of vasoactive intestinal peptide (VIP) in patients with adenocarcinoma of the gastrointestinal tract*. — Anticancer. Res., Vol. 21, pp. 1183–1187.

Jurianz K., Ziegler S., Garcia-Schuler H., Kraus S., Bohana-Kashtan O., Fishelson Z. and Kirschfink M. (1999): *Complement resistance of tumor cells: Basal and induced mechanisms*. — Mol. Immunol., Vol. 36, pp. 929–939.

Kwiatkowski D.J. (1999): *Functions of gelsolin: Motility, signaling, apoptosis, cancer*. — Curr. Opin. Cell. Biol., Vol. 11, pp. 103–108.

Ladner R.D., Lynch F.J., Groshen S., Xiong Y.P., Sherrod A., Caradonna S.J., Stoehlmacher J. and Lenz H.J. (2000): *dUTP nucleotidohydrolase isoform expression in normal and neoplastic tissues: Association with survival and response to 5-fluorourcil in colorectal cancer*. — Cancer Res., Vol. 60, pp. 3493–3503.

Li L., Weinberg C.R., Darden T.A. and Pedersen L.G. (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.

Lobo M.V., Martin M.E., Perez M.I., Alonso F.J., Redondo C., Alvarez M.I. and Salinas M. (2000): *Levels, phosphorylation status and cellular localization of translational factor eIF2 in gastrointestinal carcinomas*. — Histochem J., Vol. 32, pp. 139–150.

Nguyen D.V. and Rocke D.M. (2002): *Tumor classification by partial least squares using microarray gene expression data*. — Bioinformatics, Vol. 18, No. 1, pp. 39–50.

Oka Y., Naito I., Manabe K., Sado Y., Matsushima H., Ninomiya Y., Mizuno M. and Tsuji T. (2002): *Distribution of collagen type IV alpha 1–6 chains in human normal colorectum and colorectal cancer demonstrated by immunofluorescence staining using chain-specific apitope-defined monoclonal antibodies*. — J. Gastroenterol. Hepatol., Vol. 17, pp. 980–986.

Porter R.M., Holme T.C., Newman E.L., Hopwood D., Wilkinson J.M. and Cuschieri A. (1993): *Monoclonal antibodies to cytoskeletal proteins: an immunohistochemical investigation of human colon cancer*. — J. Pathol., Vol. 170, pp. 435–440.

Raderer M., Kurtaran A., Hejna M., Vorbeck F., Angelberger P., Scheithauer W. and Virgolini I. (1998): *123I-labelled vasoactive intestinal peptide receptor scintigraphy in patients with colorectal cancer*. — Br. J. Cancer, Vol. 78, pp. 1–5.

Rao J. (2002): *Targeting actin remodeling profiles for the detection and management of urothelial cancers—A perspective for bladder cancer research*. — Front. Biosci., Vol. 7, pp. e1–8.

Schmitt C.A., Schwaeble W., Wittig, B.M., Meyer zum Buschenfelde K.H. and Dippold W.G. (1999): *Expression and regulation by interferon-gamma of the membrane-bound complement regulators CD46 (MCP), CD55 (DAF), and CD59 in gastrointestinal tumours*. — Eur. J. Cancer, Vol. 35, pp. 117–124.

Sebestyen G.S. (1962): *Decision Making Processes in Pattern Recognition*. — New York: Macmillan.

Sobczak W. and Malina W. (1978): *Methods of Data Selection*. — Warsaw: WNT, (in Polish).

Szabo A., Boucher K., Carroll W.L., Klebanov L.B., Tsodikov A.D. and Yakovlev A.Y. (2002): *Variable selection and pattern recognition with gene expression data generated by the microarray technology*. — Math. Biosci., Vol. 176, pp. 71–98.

Tamura M., Nishizaka S., Maeda Y., Ito M., Harashima N., Harada M., Shichijo S. and Itoh K. (2001): *Identification of cyclophilin B-derived peptides capable of inducing histocompatibility leukocyte antigen-A2-restricted and tumor-specific cytotoxic T lymphocytes*. — Jpn. J. Cancer Res., Vol. 92, pp. 762–767.

Thorsteinsson L., O'Dowd G.M., Harrington P.M. and Johnson P.M. (1998): *The complement regulatory proteins CD46 and CD59, but not CD55, are highly expressed by glandular epithelium of human breast and colorectal tumour tissues*. — APMIS, Vol. 106, pp. 869–878.

Vapnik V. (1995): *The Nature of Statistical Learning Theory*. — New-York: Springer-Verlag.

Winston J.S., Asch H.L., Zhang P.J., Edge S.B., Hyland A. and Asch B.B. (2001): *Downregulation of gelsolin correlates with the progression to breast carcinoma*. — Breast Cancer Res. Treat, Vol. 65, pp. 11–21.

Appendix – Support Vector Machines

Here the reader who is not familiar with the SVM method will find a very brief introduction to this area. For more details, we refer the reader to the books by Christianini and Shawe-Tylor (2000), Haykin (1999) and Vapnik (1995).

Although SVM is a universal tool for nonlinear classification and regression, the basic idea of the method will be explained using the linear classification of a linearly separable training set. Consider a set of N vectors $\{\mathbf{x}_i\}_{i=1}^N, \mathbf{x}_i \in \mathbb{R}^n$. Each vector represents one and only one class ω_1 or ω_2. In a standard linear classification problem we look for a weight vector $\mathbf{w} \in \mathbb{R}^n$ and a scalar bias b of the linear classifying (discriminant) function

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b, \tag{12}$$

which satisfies the following set of inequalities:

$$\begin{cases} \mathbf{w}^T\mathbf{x}_i + b > 0 & \text{for} \quad \mathbf{x}_i \in \omega_1, \\ \mathbf{w}^T\mathbf{x}_i + b < 0 & \text{for} \quad \mathbf{x}_i \in \omega_2. \end{cases} \tag{13}$$

If such a function exists, the training set is called *linearly separable.*

For notational simplicity, introduce the set of desired responses (target outputs) $\{d_i\}_{i=1}^N$:

$$d_i = \begin{cases} +1 \ \text{ when } \ \mathbf{x}_i \in \omega_1, \\ -1 \ \text{ when } \ \mathbf{x}_i \in \omega_2. \end{cases} \tag{14}$$

The discriminant function (12) determines a hyperplane in an *n*-dimensional input space which is called the *decision surface*. The equation of this surface is as follows:

$$\mathbf{w}^T\mathbf{x} + b = 0. \tag{15}$$

Several possible positions of the decision surface for $n = 2$ are presented in Fig. 4. The crosses and circles indicate

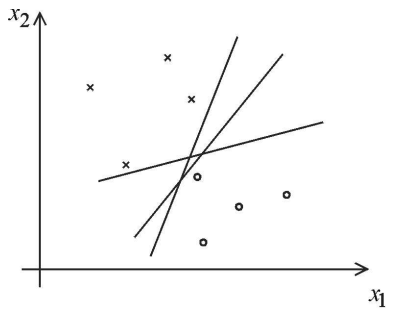

Fig. 4. Three examples of the decision surface perfectly separating the training set.

members of the ω_1 and ω_2 classess, respectively. For a linearly separable case there are an infinite number of "good" discriminant hyperplanes, i.e., those satisfying inequalities (13), but only one is optimal in the SVM sense, see Fig. 5.

The optimal hyperplane P^o satisfies inequalities (13), but it also maximizes the *margin of separation* γ, which indicates the Euclidean distance ρ between the hyperplane P and the closest vector. Hence the problem can be formulated mathematically as follows:

Problem 1. *Find optimal* $\mathbf{w}^o$ *and* b^o *that maximize*

$$\gamma = \min_i \rho(P, \mathbf{x}_i), \quad i = 1, 2, \ldots, N \tag{16}$$

subject to the constraints (13).

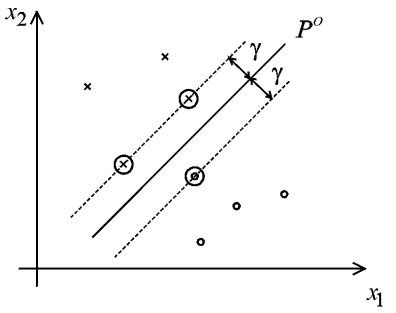

Fig. 5. Optimal decision surface in the SVM sense.

The vectors for which $\rho(\cdot)$ takes on a minimal value are called the *support vectors*. In Fig. 5 they are marked with circles.

It can be easily shown (Boser *et al.*, 1992) that Problem 1 can be transformed into the following quadratic programming problem:

Problem 2. *Find optimal* $\mathbf{w}^o$ *and* b^o *that minimize the cost function*

$$J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} \tag{17}$$

subject to the constraints

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \ldots, N. \tag{18}$$

In practice, the following dual problem is solved:

Problem 3. *Find optimal Lagrange multipliers* $\{\alpha_i^o\}_{i=1}^N$ *that maximize the cost function*

$$Q(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j d_i d_j \mathbf{x}_i^T\mathbf{x}_j \tag{19}$$

subject to the constraints

$$\sum_{i=1}^{N}\alpha_i d_i = 0, \tag{20}$$

$$\alpha_i \geq 0, \quad i = 1, 2, \ldots, N. \tag{21}$$

A non-zero value of α_i indicates that $\mathbf{x}_i$ is one of the support vectors. Optimal $\mathbf{w}^o$ and b^o can be calculated using equations

$$\mathbf{w}^o = \sum_{i=1}^{N}\alpha_i^o d_i \mathbf{x}_i, \tag{22}$$

$$b^o = d_s - (\mathbf{w}^o)^T\mathbf{x}_s, \tag{23}$$

where $\mathbf{x}_s$ is any support vector, i.e., a vector for which $\alpha_s > 0$.

Now let us assume that the training set is not linearly separable. Then, of course, the sets of constraints (13) and (18) are inconsistent. In order to make this problem tractable, a set of nonnegative scalar variables $\{\xi_i\}_{i=1}^N$ called the *slack variables* is introduced to the inequalities (18):

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \quad i = 1, 2, \ldots, N. \tag{24}$$

The cost function (17) is also modified:

$$J(\mathbf{w}, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i, \tag{25}$$

where C is a user-specified positive constant.

The problem of finding an optimal discriminant function in the non-separable case is stated as follows:

Problem 4. *Find optimal* $\mathbf{w}^o$ *and* b^o *minimizing the cost function (25) subject to the constraints (24) and*

$$\xi_i \geq 0, \quad i = 1, 2, \ldots, N. \tag{26}$$

In much the same way as in the separable case, instead of Problem 4, the following dual problem is usually solved:

Problem 5. *Find optimal Lagrange multipliers* $\{\alpha_i\}_{i=1}^N$ *that maximize the cost function (19) subject to the constraints (20) and*

$$0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, N. \tag{27}$$

Note that the only difference between Problems 3 and 5 (for the separable and non-separable cases) is in the constraints (21) and (27).

The optimum value of the vector $\mathbf{w}$ is again given by (22). The optimum value of the bias can be calculated using (23), where $\mathbf{x}_s$ is any support vector with the slack variable equal to zero. It can be shown that for such a vector the inequality has to be satisfied.

Linear SVM is a special case of more general nonlinear SVM constructed by introducing an additional set of nonlinear functions, see books (Christianini and Shawe-Tylor, 2000; Haykin, 1999; Vapnik, 1995).