STUDIA INFORMATICA 2009
Volume 30 Number 2A (83)

Weronika PIĄTKOWSKA, Jerzy MARTYNA
Jagiellonian University, Institute of Computer Science

A HYBRID CLASSIFIER BASED ON SVM METHOD FOR CANCER CLASSIFICATION

Summary. In this paper, we proposed a new method of applying Support Vector Machines (SVMs) for cancer classification. We proposed a hybrid classifier that considers the degree of a membership function of each class with the help of Fuzzy Naive Bayes (FNB) and then organizes one-versus-rest (OVR) SVMs as the architecture classifying into the corresponding class. In this method, we used a novel system of ordering the recognized expression profiles by means of using FNB and genering SVMs with the OVR scheme. The results show that our hybrid classifier is comparable to the conventional methods.

Keywords: SVM method, Fuzzy Naive Bayes, cancer classification

HYBRYDOWY KLASYFIKATOR OPARTY NA METODZIE SVM DLA KLASYFIKACJI CHORÓB ONKOLOGICZNYCH

Streszczenie. W artykule zaproponowano nową metodę klasyfikacji chorób onkologicznych. Użyto w niej m.in. naiwnego, rozmytego klasyfikatora bayesowskiego (ang. *Fuzzy Naive Bayes*) oraz maszyny z wektorami wspierającymi (ang. *Support Vector Machines*) jako systemu klasyfikującego. Tak powstały hybrydowy klasyfikator klasyfikuje choroby onkologiczne porównywalnie z konwencjonalnymi metodami.

Słowa kluczowe: metoda SVM, naiwny rozmyty bayes, klasyfikacja chorób onkologicznych

1. Introduction

Support Vector Machines (SVMs) are adaptive learning systems which receive labeled training data and transform these problems into optimization problems [12]. SVMs are usually solved by finding solutions to quadratic programming problems. Originally the SVMs were used for binary pattern classification problems where data were linearly separable, but the algorithm has been extended to handle data that are not separable by introducing slack variables [3] and to use nonlinear decision regions via kernel functions [9]. Therefore, a solution to the SVMs working with suitable kernel functions can be found by solving the quadratic programming problem in the dual observation space rather than in the primal feature space, thereby reducing overall computations.

DNA microarrays contain information about the gene expression variations of cells in different tissues [1]. The microarrays allow to understand the activities of genes underlying different cancers. Thus, the obtained information can in turn be used to identify types or subtypes of cancers

Microarrays allows to understand the activities of genes underlying different cancers. Thus, the obtained information can in turn be used to identyfy types or subtypes of cancers. Are in use currently two types of DNA microarrays: the *spotted cDNA* [4] developed at Stanford University and digonucleotide chips [6] developed by *Affymetrix*. Spotted microarrays are made of a solid surface onto which miniscule amounts (spots) of single strands of nucleotide sequences are placed which are deposited by an automated process called *contact spotting* in a grid-like arrangement. Each spot defines a specific gene and serves as a probe against which a sample RNA is hybridized. With digonucletide chips the probes are synthetized on the array on the basis of the sequences of existing or hypothetical genes using photolithographic technology. *Affymetrix* also uses multiple probes to represent the genes.

In most computational experiments with microarrays the raw data developed from these arrays must be computationally collected, processed, and integrated. This process of data preparation is called pre-processing. It allows for compensating systematic measurement errors due to array equipment imperfection and also for obtaining a single expression level for each gene. As a result, the data from different microarrays are integrated into a single data matrix. Each row of this matrix of gene expression corresponds to a different gene. Each column corresponds to a different sample of time instant of which the expression data were measuremed.

In this paper, we propose a new modified SVM method for cancer classification. The Fuzzy Naïve Bayes method described by Randon and Lawry [11] and used in pattern recognition and data analysis relies on the use of some distance function. In the proposed method, the selection stage by the Bayesian likelihood fitness function are added to conventional SVM method.

The ramainder of this paper is organized as follows. In section 2, we give basic concepts of cancer classification with the use of the SVMs method. In section 3, we overview the FNB method that was proposed to resolve unclassifiable regions in multiclass problems. In section 4, we give several experiments results to show the validity of our proposed method. Finally section 5 gives the conclusions.

2. Basic concepts of cancer classification using SVMs

In this section we give basic concepts of cancer classification with the use of the SVMs method.

With the help of the microarray technologies a large volume of gene expression profiles is produced. Microarray techniques lead to a complete understanding of the molecular variations among diseases. These gene expressions provide information about illness including some types of cancers. Several data mining methods have been developed which involve classification of gene expressions [8].

The gene expressions allow for obtaining some information which is useful for the classifier building. The irrelevant or redundant data can decrease the accuracy of classification. Therefore, a classifier which is sufficiently resistant to inaccuracy must be provided. The SVMs method represents one of the most important classifiers. We recall that the SVM maps an input sample on a high-dimensional space and minimizes the number of misclassified objects in the training set and maximizes the margin between the bounding planes.

For training set $\{(x_i, y_i)\}_{i=1}^{N}$ with the input data $x_i \in R^n$ and the output data $y_i \in R$ with the class label $y_i \in \{-1,1\}$, the SVM calculates the linear classifier

$$y(x) = sign[w^T x + b] \tag{1}$$

When the data of the two classes are separable we have the original SVM classifier [12], [13], [14] that satisfiies the following conditions.

$$\begin{cases} w^T \phi(x_i) + b \geq +1 & if \quad y_i = 1 \\ w^T \phi(x_i) + b \leq -1 & if \quad y_i = -1 \end{cases} \tag{2}$$

These two sets of inequalities can be combined into one single set as follows:

$$y_i[w^T \phi(x_i) + b] - 1 \geq 0, \qquad i = 1,2, \dots, N \tag{3}$$

where $\phi_i :\ R^n \to R^m$ is the feature mapping the input space to a usually high dimensional feature space. The data points are linearly separable by a hyperplane defined by the pair $(w \in R^m,\ b \in R)$. Thus, the classification function is given by

$$f(x) = sign\{w^T \phi(x) + b\} \tag{4}$$

Instead of estimating with the help of the feature map we work with a kernel function in the original space given by

$$K(x_i, y_i) = \phi(x_i)^T \cdot \phi(y_i) \tag{5}$$

We introduce slack variable ξ_i such that

$$y_i[w^T \phi(x) + b] \geq 1 - \xi_i, \ \ \xi_i > 0, \ \ i = 1,2, \ ... \ , N \tag{6}$$

The following minimization problem is accounted for as follows:

$$\min_{w,b,\xi} \ J(w,b,\xi) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i \tag{7}$$

subject to

$$y_i[w^T \phi(x_i) + b] \geq 1 - \xi_i, \ \ \xi_i > 0, \ \ i = 1,2, \ ... \ N, \ \ C > 0 \tag{8}$$

where C is a positive constant parameter used to control the tradeoff between the training error and the margin.

The dual problem of the system (8), obtained as a result of Karush-Kuhn-Tucker (KKT) condition, leads to a well-known convex quadratic programming (QP).

3. A hybrid classifier based on SVMs for cancer classification

In this section, we present our hybrid classifier for cancer classification which is based on SVMs and Fuzzy Naive Bayes (FNB).

The overview of our hybrid classifier is given in Fig. 1. Fuzzy Naive Bayes (FNB) are used to estimate the probability for classes $prob = \{p_1, p_2, \ ... \ , p_m\}$, while SVMs classify samples by using the original training data set of gene expression profiles. The proposed SVMs allows for a probabilistic ordering of cancer classes which, further, is used by our FNB after its estimation. The Fuzzy Naive Bayes are generally based on the Bayesian theorem. We assume that a focal set F for each attribute j is given. Let attribute x_j be numeric with universe Ω_j, then the likelihood of x_j given C_k can be represented by a density function $p(x_j \mid C_k)$ determined from the gene expression profiles D_k and a prior density according to Jeffrey's rule [5], namely

$$p(x_j \mid C_k) = \sum_{F \in F_j} p(x_j \mid F) P(F \mid C_k) \tag{9}$$

From Bayes theorem, we can obtain

$$p(x_j \mid F) = \frac{P(F \mid x_j)p(x_j)}{P(F)} = \frac{m_{x_j}(F)p(x_j)}{pm(F)} \tag{10}$$

where

$$pm(F) = \int_{\Omega_j} P(F \mid x_j)p(x_j)dx_j = \frac{\sum_{x\in D} m_{x_j}(F)}{|D|} \tag{11}$$

Substituting Eq. (10) in Eq. (11) and re-arranging gives:

$$p(x_j \mid C_k) = p(x_j)\sum_{f\in F_j} m_{x_j}(F)\frac{P(F \mid C_k)}{pm(F)} \tag{12}$$

where $P(F \mid C_k)$ can be derived from D_k according to

$$P(F \mid C_k) = \frac{\sum_{x\in D_k} m_{x_j}(F)}{|D_k|} \tag{13}$$

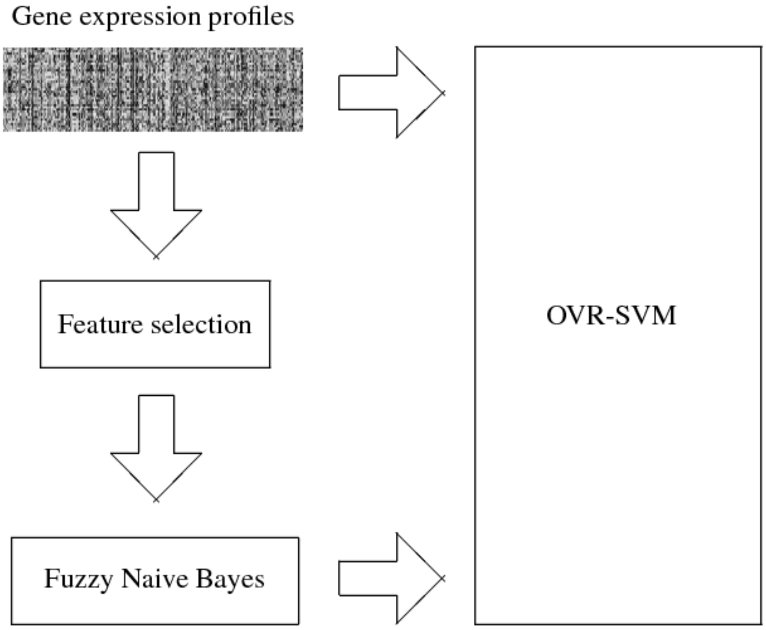

Fig. 1. Structure of hybrid classifier for cancer classification
Rys. 1. Struktura hybrydowego klasyfikatora dla klasyfikacji chorób onkologicznych

This model, called Fuzzy Naive Bayes (FNB), can provide some measures. The probability of each class x_j can be calculated with the use of Bayes theorem [7], namely:

$$p(x_j \mid C_k) = p(x_j)\sum_{F\in F_j} m_{x_j}(F_{mar\ker})\frac{P(F_{mar\ker} | C_k)}{pm(F_{mar\ker})} \tag{14}$$

where $P(F_{mar\ker} \mid C_k) = \frac{\sum_{x\in D_k} m_x(F_{mar\ker})}{|D|}$ and $F_{mar\ker}$ is a feature of the marker gene.

To improve the classification performance we used a Pearson correlation as measure of the similarity between an ideal marker and gene g_i. The Pearson correlation [2] is used here as follows:

$$C_{Pear} = \frac{\sum_{i=1}^{n}(ideal_i \times g_i) - ((\sum_{i=1}^{n} ideal_i \times \sum_{i=1}^{n} g_i)/n)}{\sqrt{(\sum_{i=1}^{n} ideal_i^2 - (\sum_{i=1}^{n} ideal_i)^2/n)(\sum_{i=1}^{n} g_i^2 - (\sum_{i=1}^{n} g_i)^2/n)}} \tag{15}$$

where n is the number of genes in the microarray data set and $ideal_i$ is a *i-th* gene in the microarray selected as the ideal marker.

Table 1

Confusion matrix

Cancer type	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Breast	65										35			
2. Prostate		86										14		
3. Lung			100											
4. Colorectal				100										
5. Lymphoma			10		90									
6. Bladder			20			80								
7. Melanoma							78			22				
8. Uterus_adeno								100						
9. Leukemia							10		90					
10. Renal										67				
11. Pancreas			33			33					34			
12. Ovary			25			25						50		
13. Mesothelioma													100	
14. CNS														100

We assumed that a gene is an informative gene if the distance given by the Pearson correlation C_{Pear} is small, while the gene is not an informative gene if the distance is large.

4. An example of analysis

To evaluate our proposed method, we used the GCM data set published by Ramaawamy et al. (2001) [10]. It consists of 144 training samples and 54 testing samples of 14 cancer classes. Each sample possesses 16063 gene expression levels. The mentioned GCM data set is available at: htpp://www.genome.wi.mit.edu/MPR/GCM.

Eight metastatic samples from the testing samples were dropped, therefore the used testing samples consisted of 46 testing samples and 14 cancer classes.

According to our method we selected 140 genes for learning FNB based on the Pearson correlation. We used the linear kernel function of SVMs. The features of samples are normalized from 0 to 1.

The obtained confusion matrix for the given 14 cancer classes is given in Table 1. As the coding strategy we used the winner-takes-all method.

Table 2

The accurracy of used methods

Method	Accurracy (%)
OVR-SVM	72
FNB	68
Hybrid classifier	80

Our programs are written in the MATLAB language. Additionally, we used the software package for the SVM algorithm which is available at http://www.kernel-machines.org.

In Table 2 we compare the accurracy of used methods. SVMs with the one-versus-rest strategy gave 72% classification accuracy. The FNB achieved 68%. The hybrid method of the OVR-SVM and the FNB produced the accuracy equal to 80%. It has been shown that our method has classified better than the OVR-SVM and the FNB treated separetely.

5. Conclusions

The hybrid classifier based on SVMs to multiclass microarray classification has been investigated for cancer recognition. The proposed method integrates SVMs and the FNB learned with the help of the OVR scheme. To verify our method we have applied the GCM cancer dataset. To reduce the dimensionality of the coding matrix we have used the Pearson correlation. The suggested method has a comparable performance to other methods but has a better performance than the method working individually.

It has been shown that further improvement of the performance of the output process depends on the output-coding strategies. Therefore, we will find the algorithm to improve the accuracy of the multiclass classification especially when the class size is small. Some algorithms like the heuristic algorithm could be considered.

BIBLIOGRAPHY

1. Brown P. O., Brotstein D.: Exploring the New World of the Genome with DNA Microarrays. Nat. Genet. Suppl., 21, 1999, p. 33÷37.
2. Cho S. -B., Ryu J.: Classifying Gene Expression Data of Cancer Using Classifier Ensemble with Mutually Exclusive Features. Proc. IEE 90 (11), 2002, p. 1744÷1753.
3. Cortes C., Vapnik V. N.: Support Vector Networks. Machine Learning, 20, 1995, p. 273÷297.

4. Duggan D. J., Bittner M., Chen Y., Melter P., Trent J.: Expression Profiling Using cDNA Microarrays. Nature Genetics, 21, 1999, p. 10÷14.
5. Jeffrey R. C.: The Logic of Decision. Gordon and Brench Inc., New York 1965.
6. Lipschutz R. J., Fodor S. P. A., Gingeras T. R., Lockhart D. J.: High Density Synthetic Eigenuclectide Arrays. Nature Genetics, 21, 1999, p. 20÷24.
7. Liu J., et al.: An Improved Naive Bayesian Classifier Technique Coupled with a Novel Input Solution Method. IEEE Trans. on Systems, Man, and Cybernetics – Part C: Appl. Rev. 31, No. 2, 2001, p. 249÷256.
8. McLachlan G. J., Do K. -A., Ambroise Ch.: Analyzing Microarray Gene Expression Data. John Wiley and Sons, 2004.
9. Müller K. R., Mike S., Rätsch G., Tsuda K., Schölkopf B.: An Introduction to Kernel-Based Learning Algorithms. IEEE Trans. On Neural Networks, Vol. 12, No. 2, 2001, p. 181÷201.
10. Ramaswamy S., et al.: Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures. Proc. Nat. Acad. Sci., Vol. 98, No. 26, 2001, p. 15149÷15154.
11. Randon J., Lawry J.: Classification and Query Evaluation Using Modeling with Words. Information Sciences. Special Issue – Computing with Words: Models and Applications, Vol. 176, 2006, p. 438÷464.
12. Vapnik V. N.: The Nature of Statistical Learning Theory. Springer-Verlag, Berlin, Heidelberg, New York 1995.
13. Vapnik V. N.: Statistical Learning Theory. John Wiley and Sons, 1998.
14. Vapnik V. N.: The Support Vector Method of Function Estimation. in: J. A. K. Suykens, J. Vandewolle (eds.). Nonlinear Modeling: Advanced Black-box Techniques, Kluwer Academic Publishers, Boston 1998, p. 55÷85.

Recenzent: Prof. dr hab. inż. Andrzej Polański

Wpłynęło do Redakcji 5 marca 2009 r.

Omówienie

Mikroszeregi DNA pozwalają na analizę występowania okogenów. Przy użyciu specjalnie skonstruowanego hybrydowego klasyfikatora zbadano występowanie chorób onkologicznych. Do budowy tego klasyfikatora użyto metody wektorów podpierających (ang. *Support Vector Machines*) oraz naiwny, rozmyty klasyfikator bayesowski (ang. *Fuzzy*

Naive Bayes). Metodę SVM użyto w postaci architektury typu „jeden przeciw reszcie" (ang. *one-versus-rest*), co umożliwia oddzielną klasyfikację każdej klasy odnoszącej się do choroby onkologicznej. Wykazano, że tak opracowany hybrydowy klasyfikator posiada lepsze możliwości klasyfikacji niż obecnie stosowane konwencjonalne metody.

Addresses

Weronika PIĄTKOWSKA: Uniwersytet Jagielloński, Instytut Informatyki Stosowanej,
ul. Reymonta 4, 30-059 Kraków, Polska.

Jerzy MARTYNA: Uniwersytet Jagielloński, Instytut Informatyki, ul. Łojasiewicza 4,
30-348 Kraków, Polska, martyna@softlab.ii.uj.edu.pl .