*Partial Least Squares PLS, dimension reductions,
MMulticlass Support Vector Machines MSVM,
One-Versus-One OvO, One-Versus RestOvR*

Sebastian STUDENT[*]

USING MULTICLASS SVM METHODS FOR CLASSIFICATION OF DNA MICROARRAY DATA

One important application of gene expression microarray data is classification of samples into categories, such as the type of tumor. A classifier using Multiclass SVM [4] (Support Vector Machines) is described in this article. Our classifier involves dimension reduction using Multivariate Partial Least Squares (MPLS) for classification more than two classes. We use also two methods based on binary classifications: One-Against-All [5] and One-Against-One [6]. These three methods have been tested on a data set involving 125 tumor/normal thyroid human DNA microarrays samples. There are 66 Papillary throid carcinoma, 32 follicular throid carcinoma and 27 normal tissues. The most important thing is to find small number of genes that discriminate between these three classes with good accuracy. The best genes can be selected for Q-PCR validation. Molecular markers differentiating between throid cancer and normal tissues can help in clinical diagnostics and therapy methods. For error estimation we are use the bootstrap .632 [8] technique. Major issue with bootstrap estimators is their high computational cost. That is why we use a OpenMosix with MPI (Message Passing Interface) cluster technology for this system for parallel computation. space.

1. INTRODUCTION

The most common application of microarray analysis is classification samples into different classes. Such a classification can be used for early diagnosis and for choose the best therapy methods. Support vector machines SVM [1,2,3] is one of the classification method that gives very good results. For classification into more than two classes we compare three classifications methods: Multiclass SVM (MSVM) [4], One-Against-All [5] and One-Against-One [6]. The most problem causes huge dimension of data arrays. The problem is that we have much more genes (several thousands) than number of observations (in our case 125). Traditional statistical methodology for classification does not work when there are more variables than samples. Thus, methods able to cope with the high dimensionality of the data are needed. This is very important, because a small lists of genes are very useful for understanding mechanism of cancer formation and metastasis. Significant aspect is to build a good tool for multiclass classification that uses small number of genes. In this case we can use it not only for microarrays analysis, we also can use it as a diagnosis tool that used Real-time quantitative PCR (Q-PCR). This method allows reducing costs and can be applied for general diagnostics investigation. In this paper we use multivariate partial least squares MPLS [7] for dimension reduction. It is very important to

[*] Institution: Silesian University of Technology, Automatic Control, Electronics and Computer Science; Automatic Institute; Address: Akademicka 16, 44-100 Gliwice, Poland

design the classifier from the sample data and then apply an error estimation procedure. In our case we use 0.632 bootstrap estimator [8]. It was shown that the bootstrap estimators, in particular the 0.632 estimator, gives better performance than cross-validation and resubstitution for relatively small sample microarray classification [9]. The bootstrap methodology is a general resampling strategy that can be applied to error estimation. The 0.632 bootstrap estimator

$$\hat{\varepsilon}_{b632} = (1-0.632)\hat{\varepsilon}_{resub} + 0.632\hat{\varepsilon}_0 \qquad (1)$$

where

$\hat{\varepsilon}_{resub}$ - resubstitution estimator estimates the error by directly computing the error on the training data

$\hat{\varepsilon}_0$ - zero estimator estimates the error by computing the error on the left-out (test) data points

The 0.632 bootstrap estimator uses a weighted average of the zero and resubstitution estimators. Error in every iteration for train and test data points is given by the average:

$$\varepsilon = \frac{\sum_{n=1}^{N} \varepsilon_n}{N} \qquad (2)$$

where N is number of classes.

The main weakness of classification methods with bootstrap estimators is their high computational cost. For this reason we used a OpenMosix with matlab toolbox MatlabMPI (Message Passing Interface) for parallel computation. Scheme of this classifier using support vector machines is shown in Fig1.

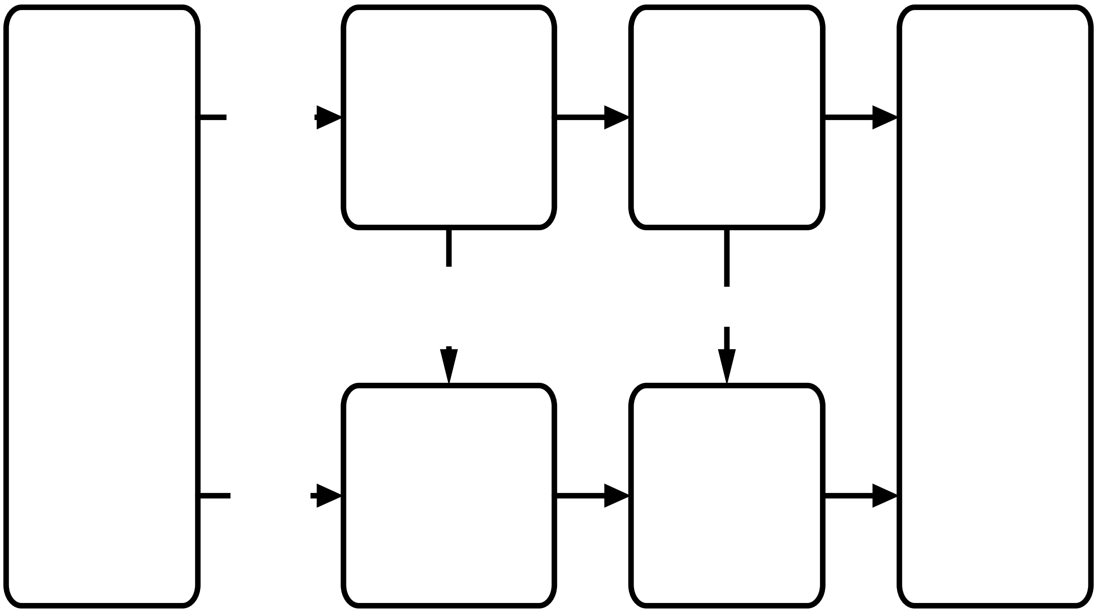

Fig.1. Scheme of classifier using support vector machines SVM

2. DATA STRUCTURE

We have conducted calculations based on real patient data. Our dataset include 125 probes of thyroid carcinoma with multiple classes: papillary thyroid cancer (PTC), follicular thyroid cancer with follicular adenoma (FTC+FA) and normal thyroid tissue (NmHy). The most important task is to find small number of genes that discriminate between our classes with the smallest error rate. That means that we search for molecular markers to understand the molecular basis of metastasis of various cancer. The number of samples in each class is: PTC 66 samples, FTC+FA 32 samples, NmHy 27 samples.

3. MULTIVARIATE PLS ANALYSIS

For gene expression data the number of tissue samples is much smaller than the number of genes, that is why the dimension reduction is needed. The goal of dimension reduction methods is to reduce the high dimensional predictor space (in our case genes space) to a lower dimensional space. There are some techniques such as Principal Components Analysis PCA and Partial Least Squares PLS. PCA procedure reduce the high dimensional data without regard to response variation. In contrast to PCA, PLS components are chosen so that the sample covariance between the response and a linear combination of the predictors is maximum. For that reason PCA is worse than PLS in prediction [10] and is not implement in our classifier.

To describe PLS some notations are required. Let $\boldsymbol{X}$ be an N x p matrix of N samples and p genes. y denote the N x 1 vector of response values, in our case indicator of tissue class. In PLS the components are constructed to maximize the objective criterion based on the sample covariance between $\boldsymbol{y}$ and linear combinations of genes $\boldsymbol{Xc}$. Thus, we find the weight vector $\boldsymbol{w}$ satisfying criterion:

$$\boldsymbol{w_k} = \arg\max_{\boldsymbol{w}'\boldsymbol{w}=1} \operatorname{cov}^2(\boldsymbol{Xw}, \boldsymbol{y}) \tag{3}$$

Subject to the orthogonality constraint

$$\boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{w}_j = 0 \; \textit{for all} \; 1 \le \boldsymbol{j} \le \boldsymbol{k} \tag{4}$$

Because we can find more than one component in our program every next vector $\boldsymbol{w}$ is considered with smaller weight corresponding to variation that this component explain. The weights of each components equal to normalized covariance $\operatorname{cov}(\boldsymbol{Xw}, \boldsymbol{y})$. For classification with SVM we use only genes that obtain the maximal summarized weights. We don’t use calculated components for classification, but we use weighted components for build genes ranking and the best genes are used in our classifier. The best genes have the highest summarized weights and we take only 30 best genes.

PLS technique can be applied to dimension reduction problems on different ways: multiclass - we use PLS ones to solving multiclass problem, One-Against-All approaches in which we search so much times genes as we have class to separate members of that class form members of other classes, and One-Against-One similarly to separate members of one class from members of the other. For our needs we introduce notation PLS+MCLASS for multiclass approach and similarly PLS+OvO, PLS+OvR.

4. SUPPORT VECTOR MACHINES SVM

Originally support vector machines was designed for two class problems. There are two approaches for multiclass classification. In the first case we can combine more two-class classifier, and in the second case we can compare all class in one optimization problem. In our program we use One-Against-One (OvO), One-Against-Rest and multiclass approach (MSVM). In OvR we separate with two class SVM members of each class from members of other classes. In OvR we separate members of one class from every other class separately. Let *S* represent set of *N* points $\boldsymbol{x}_i \in \boldsymbol{R}^d$ for i=1,2,…,*N*. The principle of the SVM is to determine optimal hyperplane that split S into two half spaces which correspond to the two distinct classes and reserve the maximum geometric margin between corresponding class and the hyperplane. Determination of the membership to class -1 or 1 is done by assigning $\boldsymbol{x}_i$ to relevant. When the data set is linearly separable, that means satisfy condition of existence $\boldsymbol{w} \in \boldsymbol{R}^d$ and $\boldsymbol{b} \in \boldsymbol{R}$ such that:

$$\boldsymbol{y}_i(<\boldsymbol{w},\boldsymbol{x}>+\boldsymbol{b}) \geq 1 \tag{5}$$

for i=1,2,…,N.

The pair (w,b) defines hyperplane:

$$<\boldsymbol{w},\boldsymbol{x}>+\boldsymbol{b}=0 \tag{6}$$

To find optimal solution it is necessary to find hyperplane which maximizes the distance between (w,b) and the nearest point. Hence, the solution (w,b) maximize expression (primary problem):

$$\boldsymbol{L}(\boldsymbol{w},\boldsymbol{b},\alpha)=\frac{1}{2}<\boldsymbol{w},\boldsymbol{w}>-\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}}\alpha_i\left[\boldsymbol{y}_i(<\boldsymbol{w},\boldsymbol{x}_i>+\boldsymbol{b})-1\right] \tag{7}$$

where $\alpha \geq 0$.

Dual problem is:

$$\boldsymbol{L}(\alpha)=\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}}\alpha_i-\frac{1}{2}\sum_{\boldsymbol{i},\boldsymbol{j}=1}^{\boldsymbol{N}}\alpha_i\alpha_j\boldsymbol{y}_i\boldsymbol{y}_j<\boldsymbol{x}_i,\boldsymbol{x}_j> \tag{8}$$

Optimisation solution is

$$\boldsymbol{w}^*=\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}}\boldsymbol{y}_i\alpha_i^*\boldsymbol{x}_i \tag{9}$$

Solutions $\alpha^*, (\boldsymbol{w}^*, \boldsymbol{b}^*)$ satisfies condition:

$$\alpha^*\left[\boldsymbol{y}_i\left(< \boldsymbol{w}^*, \boldsymbol{x}_i > + \boldsymbol{b}^*\right) - 1\right] = 0 \tag{10}$$

Finally the classifier can be written as follows:

$$\boldsymbol{f}(\boldsymbol{x}) = \left(\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \boldsymbol{y}_i \alpha_i^* < \boldsymbol{x}_i, \boldsymbol{x} > + \boldsymbol{b}^*\right) \tag{11}$$

Of course solution is valid only for linearly separated cases and linear kernel. In our case the best results were obtained for linear kernel in all cases: OvR, OvO and MSVM.

5. RESULTS

Numerical experiment include all three cases OvO, OvR and multiclass. For each approaches we executed 500 bootstrap iterations for our thyroid tissue set. In every case we perform dimension reduction using PLS procedure choosing 300 best genes according to approaches PLS+OvO, PLS+OvR and PLS+MCLASS. To make gene ranking every gene, which was in first 300 genes receives one point in each iteration. In Tab.1 we show accuracy results acc and 95% confidence interval (accL, accH) estimated with percentile method.

Table 1. Classification accuracy for OvR, OvO and MSVM

Classification method	**Dimension reduction method**								
	PLS+OvR			**PLS+MCLASS**			**PLS+OvO**		
	acc	**accL**	**accH**	**acc**	**accL**	**accH**	**acc**	**accL**	**accH**
MSVM	0,918	0,858	0,981	0,884	0,818	0,951	**0,923**	0,853	0,989
OvR	0,911	0,851	0,964	0,885	0,815	0,940	0,919	0,824	0,978933
OvO	0,909	0,843	0,968	0,885	0,811	0,979	0,903	0,831	0,970

As we can see the best results we obtain for multiclass classifier MSVM, but in case when the PLS+OvO approach was used. For this case accuracy for different genes number is shown in Fig.2.

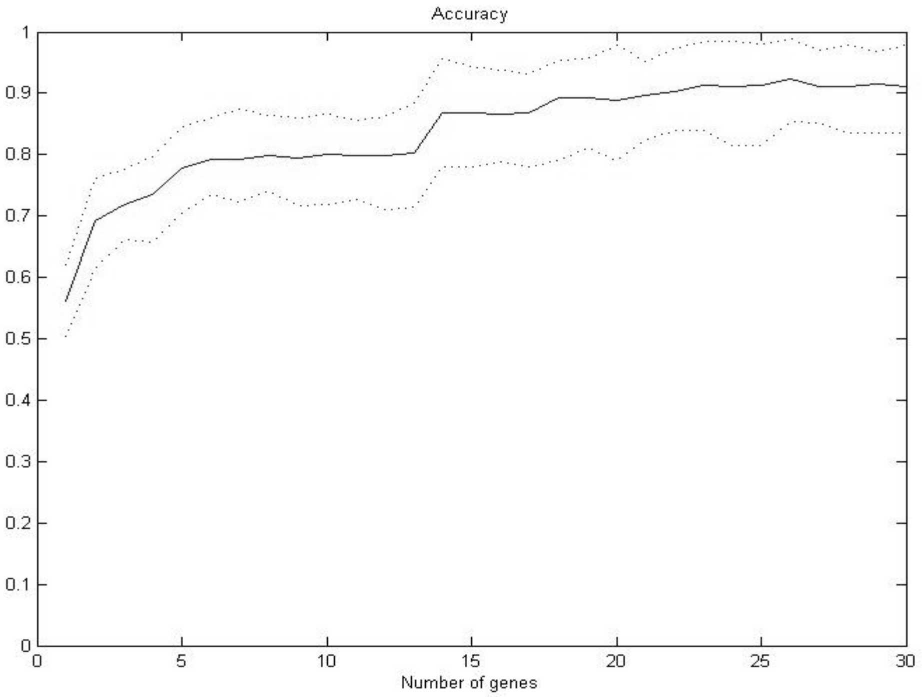

Fig.2. Classification accuracy for MSVM as function of the genes

Analysis of genes ranking figure shows that there is relatively small number of genes that was at least one times in firs 300 best genes in whole 500 iterations. In case of classifier with PLS+OvO dimension reduction we have 341 such genes.

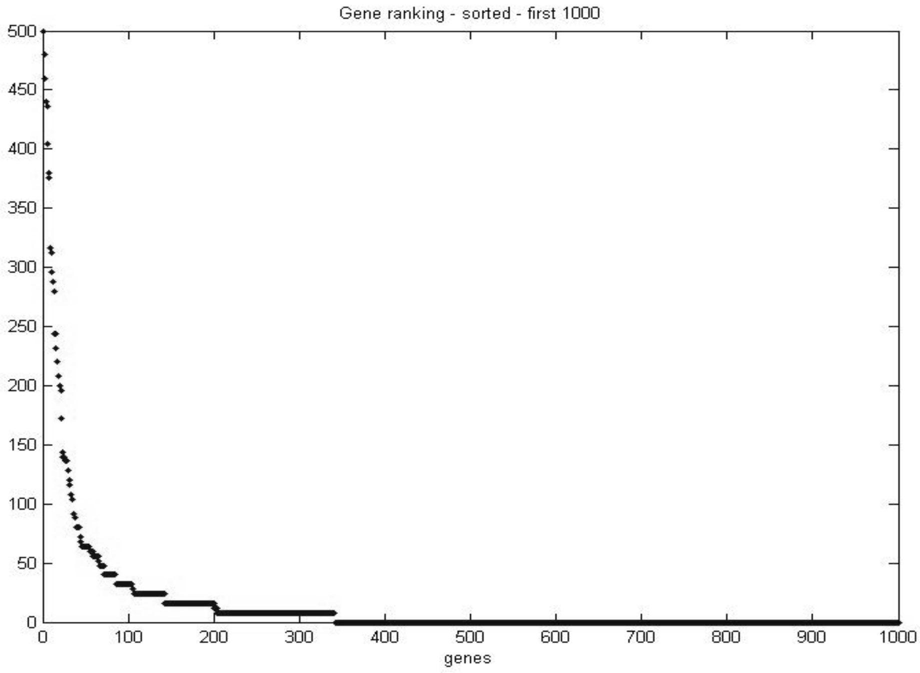

Fig.3. Genes ranking

We also build classifier that use 30 PLS components to compare the results. That means we don't use PLS procedure to find the best genes, but to choose the best variation directions ("super genes"). In this case we calculate weighted genes components and use it for classification. In Tab.2 we show accuracy results acc and 95% confidence interval (accL, accH) estimated with percentile method for this approach.

Table 2. Classification accuracy for OvR, OvO and MSVM for classifier with the best 30 "super genes"

	acc	accL	accH
MSVM	0,924	0,865	0,965
OvR	0,925	0,866	0,965
OvO	0,921	0,865	0,966

As we can see the result is better than classifier with PLS dimension reduction, but we need all genes to build our PLS components. Because we need all genes we can't reduce the costs of diagnostics investigation with Real-time quantitative PCR (Q-PCR). This approach can be useful only for microarrays dataset.

6. CONCLUSION

In this work we construct classifier that compares three multiclass methods. PLS modification permit use is for find best genes list. We need to validate presented in this article results with biological knowledge. That means we need to proof our genes list witch QPCR method. Then we can question of usefulness our system for diagnostics. We must underline that employed parallel computation makes possible work with bigger data sets. This makes possible to obtain more reliable classifier with less classifier error and can help to find new genes that take part in neoplasia. In this article we take pressure to find relatively small numbers of genes and not to find classifier witch smallest error rate. We use PLS methods not in conventional adaptation but as procedure to find the best genes.

7. ACKNOWLEDGEMENTS

I thank dr. K. Fujarewicz for help and technical assistance and prof. B.Jarząb and dr. M. Jarząb for facilitate microarrays data.

This work was supported by a grant 1/0-PBZ-MNII-2/1/2005.

BIBLIOGRAPHY

[1] BOSER B. E., I.M. GUYON, V. VAPNIK, A training algorithm for optimal margin classifiers. Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, 1992

[2] GUYON I., J. WESTON, S. BARNHILL, V. VAPNIK, Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, Vol. 64, pp. 389–422, 1999

[3] BROWN M. P .S.., W.N. Groundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares Jr, D. Haussler, 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, 2000

[4] J. WESTON AND C. WATKINS, MultiClass Support Vector Machines. In M. Verleysen, editor Proceedings of ESANN99, Brussels. D. Facto Press, 1999.

[5] L. BOTTOU, C. CORTES, J. DENKER, H. DRUCKER, I. GUYON, L. JACKEL, Y. LECUN, U. MULLER, E. SACKINGER, P. SIMARD, AND V. VAPNIK,: Comparison of classifier methods: A case study in handwriting digit recognition, in Proc. Int. Conf. Pattern Recognition. , pp. 77–87, 1994.

[6] J. FRIEDMAN.: Another Approach to Polychotomous Classification. Dept. Statist., Stanford Univ., Stanford, CA, 1996.

[7] HÖSKULDSSON A.: PLS regression methods. J. Chemometrics., 2(3) 211-228, 1988.

[8] B EFRON: Estimating the error rate of a prediction rule: improvement on cross-validation, JASA 78, pp. 316–331, 1983.

[9] ULISSES BRAGA-NETO, EDWARD R. Dougherty: Is cross-validation valid for small-sample microarray classification? Bioinformatics 20(3): 374-380, 2004.

[10] NGUYEN DV, ROCKE DM.: Tumor classification by partial least squares using microarray gene expression data. Bioinformatics. 2002 Jan;18(1):39-50, 2001.