# Classification of DNA microarray data with PCA based Support Vector Machines

**Tomasz Stokowy**
*Silesian University of Technology,*
*Automatic Control Department;*
*Cancer Center and Institute of Oncology,*
*Nuclear Medicine and Endocrine Oncology Department;*
*Gliwice, Poland*
(09.08.2010, prof. dr hab. inż. Andrzej Świerniak, *Silesian University of Technology*)

**Abstract**

The article includes information about the advantages of support vector machines in DNA microarray data classification. The analysis of the data was performed on the set of 108 microarrays representing different types of thyroid cancer and healthy thyroid tissue samples.

The idea of application of PCA based SVM evolved from the studies on explanation of biological variability sources in the Affymetrix microarray data.

There was presented way of classification data presentation, method of dealing with technical bias in the microarray experiment results and proves for molecular differences in thyroid cancer types.

Proposed solutions are said to be a proper solution for classification results presentation. Thanks to its application proper explanation of molecular differences in thyroid cancer types has been obtained.

## 1. Introduction

Thyroid cancer is a neoplasm developing in the patients of various age, including significant number of patients below the age of 20. Thyroid tumors represent several classes, however some of those classes are very hard to be recognized histopathologically. Improper diagnosis may lead to application of treatment dangerous for the patient.

Molecular diagnosis based on the microarray analysis can be a great supporting tool for medical diagnosis. In some cases bioinformatic analysis accuracy can be even higher than application of a classical histopathological approach.

There appears an opportunity to investigate usefulness of artificial intelligence methods in thyroid cancer diagnosis. One of such methods are principal components analysis based support vector machines [1] – method that has not been applied in cancer diagnosis yet.

Application of this technique can be powerful thanks to its nonlinearity, which can fit to biological variance of samples. Implementation of the method with open source R/Bioconductor package 'kernlab' [2] makes the analysis easy to be applied also for different types of tumors.

## 2. Data

The transcriptome of samples was analyzed by the use of oligonucleotide expression microarrays Affymetrix in the laboratory of Nuclear Medicine and Endocrine Oncology Department, Cancer Center and Institute of Oncology, Gliwice, Poland.

There were prepared set of the 108 samples described by 22280 GCRMA normalized gene expressions:

- 30 Papillary thyroid cancers (PTC),
- 30 Follicular thyroid cancers (FTC),
- 18 Medullary thyroid cancers (MTC),
- 30 Normal thyroid tissues (normal),

hybrydized on two types of microarrays: Affymetrix HG U133a and HG U133plus2 (Figure 1 and 2).

Randomly selected half of the samples from each class were used as a training set for the supervised analysis, the other half was used as the independent testing set.

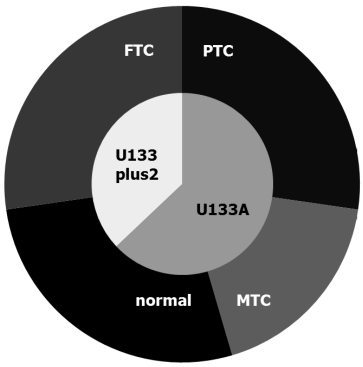


**Fig.1. Distribution of the samples in the data set**

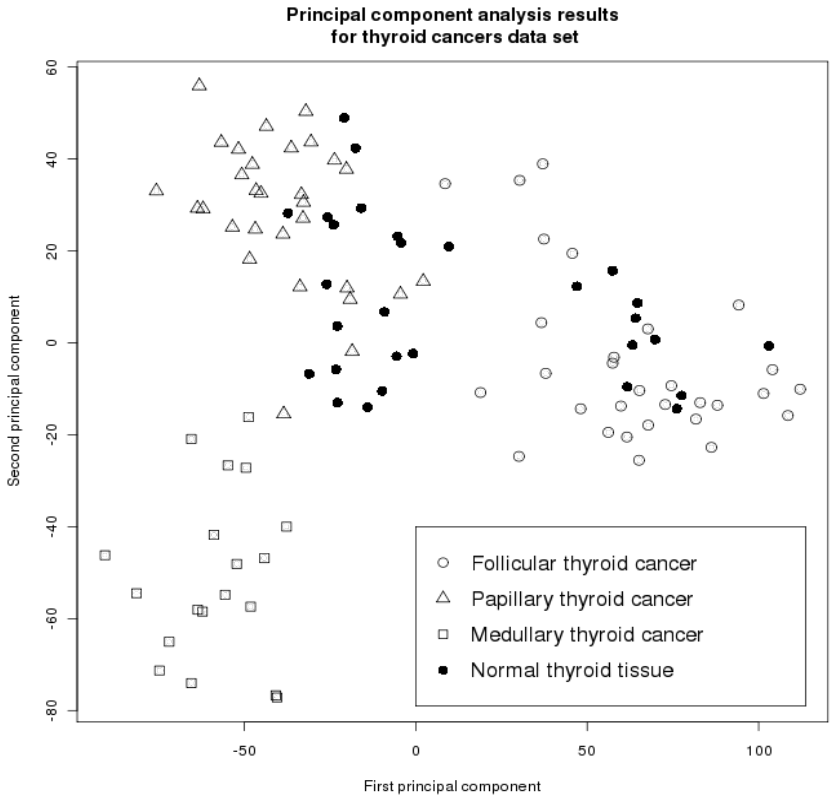


**Fig.2. Distribution of the data on the PCA plane**

## 3. Results

There were proposed a computer program in the R/Bioconductor environment to estimate PCA based SVM predictions for DNA microarray data.

The program was applied to the data set describing thyroid cancer and normal thyroid tissue samples. The data show that it is possible to explain biological differences in cancer filtering out some technical effects. It is possible to propose valid microarray analysis for technically biased data set, however it is necessary to test the results on the independent samples and validate the results with the quantitative PCR or other molecular method.

Thanks to the method classification and its accuracy measure can be now well explained by biological and medical background of the experiment.

## 4. Discussion

PCA based SVM results in a class membership prediction $y_p$ of testing samples across the whole PCA components plane. Values estimated by SVM are represented as numeric values included in the $y_p \in \langle -1,1 \rangle$ interval. Higher value of $|y_p|$ indicates stronger prediction of class membership for testing sample. Plot of the $y_p$ values is the main and the most important result of the method because its shape and properties can be interpreted biologically.

The first and second components describe in majority of cases the most significant difference between classes. The next components refer to respectively less important factors of samples variability.

The prediction plots give also an opportunity to observe the outliers and verify the calculated classification accuracy. The accuracy close to 100% as obtained in the Figures 4 and 5 is a proof for molecular difference between investigated types of thyroid tumors.

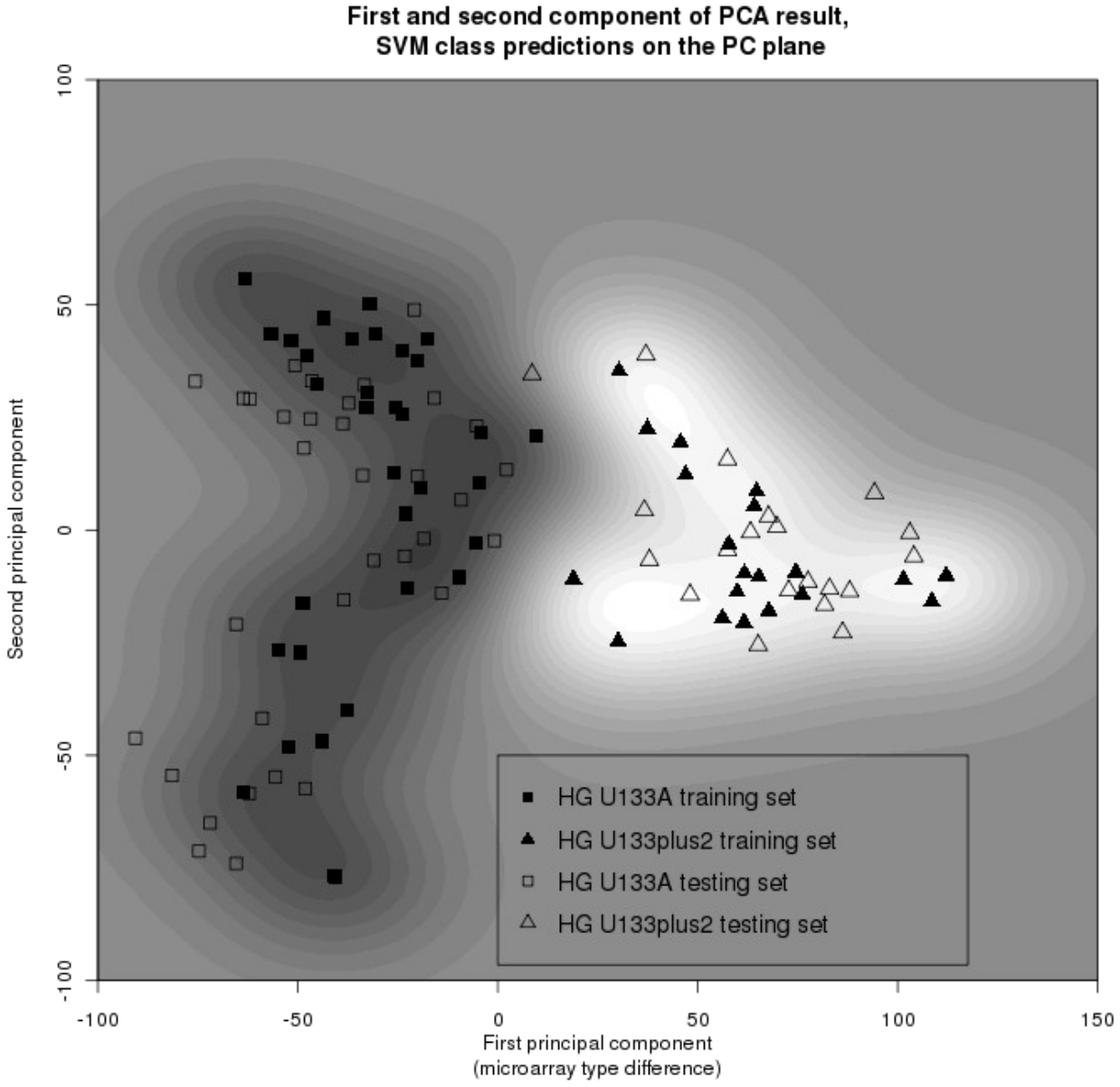


**Fig.3. Technical bias in experiment result**

Accuracy close to 100% has been also obtained in the first principal component, presented in the Figure 3.

The difference between classes could be mistaken with FTC versus other samples difference. In fact, 10 normal samples hybrydized on Affymetrix HGU133plus2 (which can be observed in FTC cluster in the Figure 2) are proof that the difference in the first principal component is the technical bias which have to be taken into account. [3]

Two close generations of the Affymetrix arrays, analyzed on the same set of genes give significant bias. On the other hand this technical factor can be filtered out by extraction of the first principal component. Application of Singular Value Decomposition results in list of genes that we should remove from the analysis. Elimination of appropriate genes in the microarray preprocessing makes the results valid.

The second principal component has been recognized as the difference between medullary thyroid cancer and other samples (Figure 4) [4].

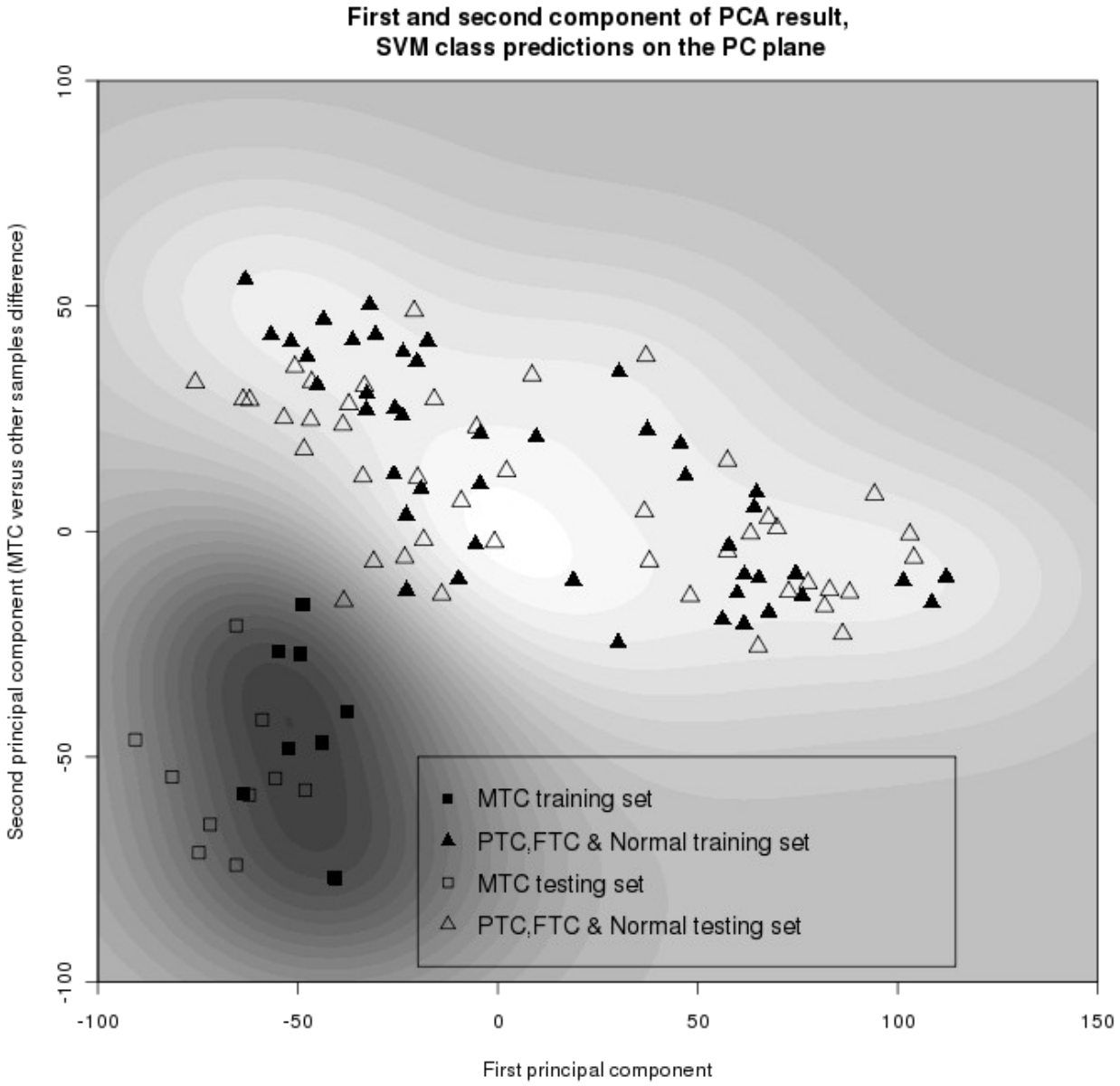


**Fig.4. Molecular difference between medullary thyroid cancer(MTC) and other samples from the data set**

Medullary thyroid cancer originates from parafollicular thyroid cells (C cells). That makes the difference with other thyroid tumors, originating from epithelial thyroid cells. This difference in the origin causes the difference in molecular profile of tumors, presented as the second dimension of principal components matrix.

The third component presented in the Figure 5 represents the papillary thyroid cancer versus unchanged cancerously tissue [5].

In the further dimensions of proximity matrix there appears as well the difference between follicular thyroid cancer and papillary thyroid cancer which are well-differentiated thyroid cancers.

The differences between classes in DNA microarray data can be observed by PCA based SVM, however it is necessary to pay attention during interpretation of sources of biological variability in further dimensions of PCA.

It remains to be shown that the microarray experiments are related to the other molecular platforms like micro RNA chips, exon arrays or next generation sequencing. Thanks to that the integrative analysis will help us to understand well the tasks of particular parts of a human genome.

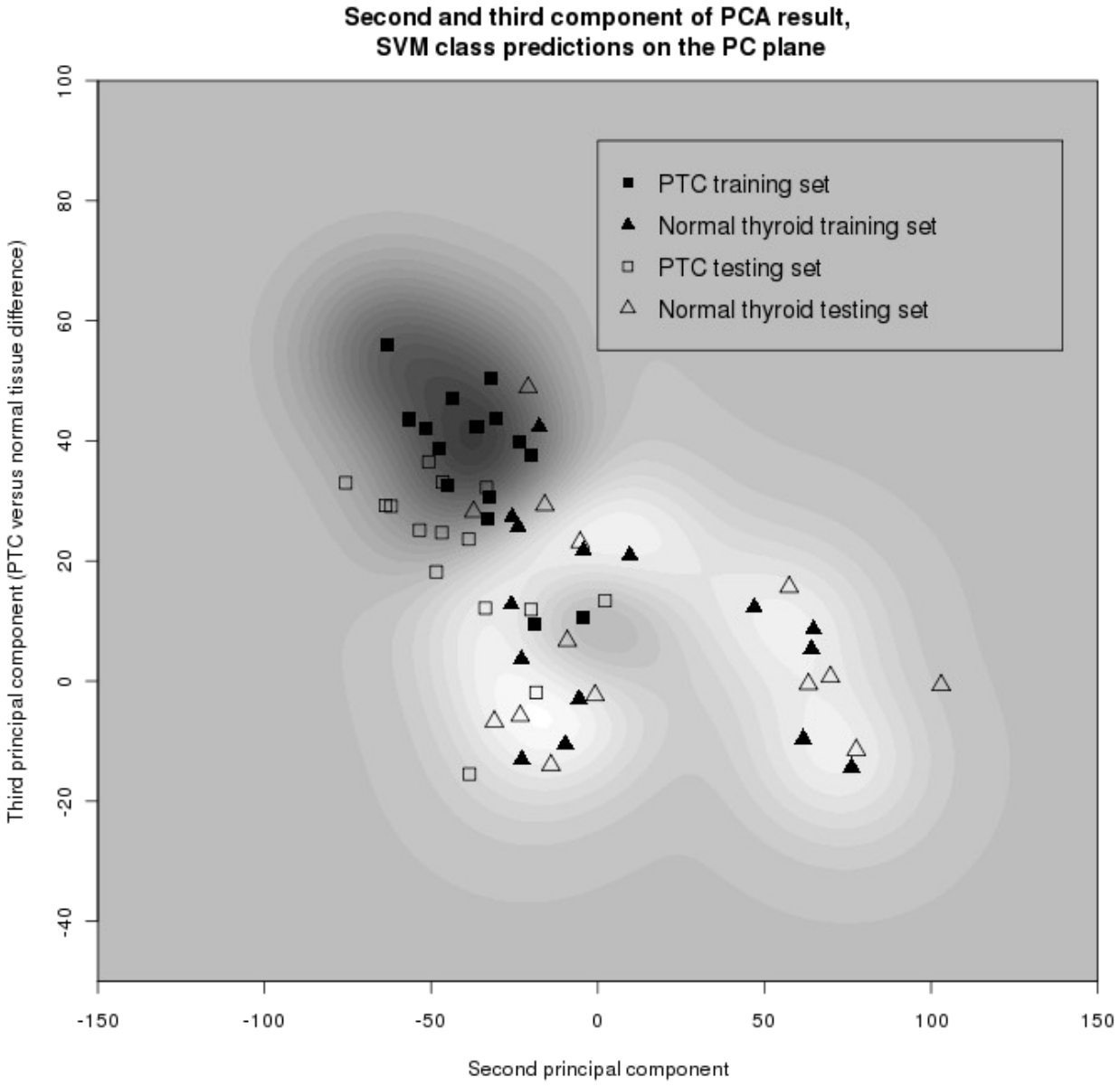


**Fig.5. Molecular difference between papillary thyroid cancer(PTC) and thyroid tissue without cancerous changes**

## 5. Conclusions

Thyroid cancer data set has been classified with accuracy close to 100% on the independent testing samples.

Predictions of nonlinear SVM C-svc classifier were plotted on the principal components plane to explain the biological difference in the samples.

The technical bias caused by the microarray platform has been neglected.

The field of molecular thyroid cancer experiments has been opened for more detailed investigations on malignity and oncocity in thyroid tumors.

## 6. Bibliography


[1] Guo-Zheng Li, Hua-Long Bu, Mary Qu Yang, Xue-Qiang Zeng, and Jack Y Yang: *Selecting subsets of newly extracted features from PCA and PLS in microarray data analysis*, BMC Genomics. 2008; 9(Suppl 2): S24.

[2] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, Achim Zeileis, *kernlab - An S4 Package for Kernel Methods in R*. Journal of Statistical Software 11(9), 1-20, (2004). URL http://www.jstatsoft.org/v11/i09/

[3] Gautier Laurent, Cope Leslie, Bolstad Benjamin M, Irizarry Rafael A.: *Affy--analysis of Affymetrix GeneChip data at the probe level*. Bioinformatics. 2004 Feb 12; 20(3):307–315.

[4] Małgorzata Oczko-Wojciechowska, Jan Włoch, Małgorzata Wiench, Krzysztof Fujarewicz, Krzysztof Simek, Grzegorz Gala, Elzbieta Gubała, Sylwia Ulczok-Szpak, Barbara Jarząb: *Gene expression profile of medullary thyroid carcinoma — preliminary results,* Polish Journal of Endocrinology, Volume 57; Number 4/2006, ISSN 0423–104X

[5] Barbara Jarząb et al.: *Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications.* Cancer research 2005;65(4):1587-97.



**Author:**

MSc. Tomasz Stokowy
Silesian University of Technology
Automatic Control Department
ul. Akademicka 10
44-100 Gliwice
tel. (032) 237 28 99
email: *tomasz.stokowy@polsl.pl*