Reliability analysis of Classification of Gene Expression Data

Sujata dash, KMBB, Bhubaneswar, Orissa, India.
B.N. Patra, GIET, Gunupur, Orissa, India.

Abstract

Gene expression data usually contains a large number of genes, but a small number of samples. Feature selection for gene expression data aims at finding a set of genes that best discriminate biological samples of different types. Classification of tissue samples into tumor or normal is one of the applications of microarray technology. When classifying tissue samples, gene selection plays an important role. In this paper, we propose a two-stage selection algorithm for genomic data by combining some existing statistical gene selection techniques and ROC score of SVM and k-nn classifiers. The motivation for the use of a Support Vector Machine is that DNA microarray problems can be very high dimensional and have very few training data. This type of situation is particularly well suited for an SVM approach. The proposed approach is carried out by first grouping genes with similar expression profiles into distinct clusters, calculating the cluster quality, calculating the discriminative score for each gene by using statistical techniques, and then selecting informative genes from these clusters based on the cluster quality and discriminative score .In the second stage, the effectiveness of this technique is investigated by comparing ROC score of SVM that uses different kernel functions and k-nn classifiers. Then Leave One Out Cross Validation (LOOCV)is used to validate the techniques.

Key Words : Fisher Criterion, Golub Signal-to-Noise, Mann-Whitney Rank Sum Statistic, Leave One Out Cross Validation (LOOCV), Support Vector Machine(SVM)

1. Introduction

The problem of cancer classification has clear implications on cancer treatment. Additionally, the advent of DNA microarrays introduces a wealth of genetic expression information for many diseases including cancer. An automated or generic approach for classification of cancer or other diseases based upon the microarray expression is an important problem. A generic approach to classifying two types of acute leukemia was introduced in Golub et. al.[7]. They achieved good results on the problem of classifying acute myeloid leukemia (AML) versus acute lymphoblastic leukemia (ALL) using 50 gene expressions. Their approach to classification consisted of summing votes for each gene on the test data, and looking at the sign of the sum. In this paper, four statistical techniques include Fisher Criterion, Golub Signal-to-Noise, traditional t-test, and Mann-Whitney Rank Sum Statistic are studied. The objective is to investigate the impact and importance of the gene selection techniques to the tissue classification performance. The effectiveness of this technique is investigated by comparing ROC score of SVM that uses different kernel functions: the dot product, quadratic dot product, cubic dot product and the radial basis function and the *k*-nn classifiers. The LOOCV is applied to validate the techniques. Results show that a better classification performance can be achieved by the classifiers if genes are first selected prior to the classification task.

2. Background on cDNA Microarrays

A *gene* consists of a segment of DNA which codes for a particular *protein*, the ultimate expression of the genetic information. A *deoxyribonucleic acid* or *DNA* molecule is a double-stranded polymer composed of four basic molecular units called nucleotides. Each *nucleotide* comprises a phosphate group, a deoxyribose sugar, and one of *four nitrogen bases*. The four different bases found in DNA are adenine (A), guanine (G), cytosine (C), and thymine (T).The two chains are held together by hydrogen bonds between nitrogen bases, with base-pairing occurring according to the following rule: G pairs with C, and A pairs with T. While a DNA molecule is built from a four-letter alphabet, proteins are sequences of twenty different types of *amino acids*. The expression of the genetic information stored in the DNA molecule occurs in two stages: (i) *transcription*, during which DNA is transcribed into *messenger ribonucleic acid* or *mRNA*, a single-stranded complementary copy of the base sequence in the DNA molecule, with the base uracil (U) replacing thymine; (ii) *translation*, during which mRNA is translated to produce a protein. The correspondence between DNA's four-letter alphabet and a protein's twenty-letter alphabet is specified by the *genetic code*, which relates nucleotide triplets to amino acids. cDNA microarrays consist of thousands of individual DNA sequences printed in a high density array on a glass microscope slide. The relative abundance of these DNA sequences in two DNA or cDNA samples may be assessed by monitoring the differential hybridization of the two samples to the sequences on the array. To this end, the two DNA samples or *targets* are labeled using di_erent fluorescent dyes (*e.g.* a red-fluorescent dye Cy5 and a green-fluorescent dye Cy3), then mixed and hybridized with the arrayed DNA sequences or *probes.*After this competitive hybridization, fluorescence measurements are made separately for each dye at each spot on the array. The ratio of the fluorescence intensity for each spot is indicative of the relative abundance of the corresponding DNA sequence in the two samples (see http://rana.Stanford.EDU/software/ for more information on the measurement of fluorescence intensities). Microarrays are being applied increasingly in cancer research to study the molecular variations among tumors . This should lead to an improved classification of tumors, which in turn should result in progresses in the prevention and treatment of cancer. An important aspect of this endeavor is the ability to predict tumor types on the basis of gene expression data. We review below a number of prediction methods and assess their performance on the cancer datasets described in Section 3.

3. Gene Selection Technique

3.1 The Fisher Criterion[9], *fisher*, is a measure that indicates how much the class distributions are separated. The coefficient has the following formula:

$$fisher = \frac{(\mu_1 - \mu_2)^2}{(\nu_1 + \nu_2)} \qquad (1)$$

where *μi* is the mean and *vi* is the variance of the given gene in class *i* (there are two classes in this study, the positive class i.e. the normal tissue sample and the negative class, i.e. the tumor tissue sample). It gives higher values to genes whose means differ greatly between the two classes, relative to their variances.

3.2 Golub Signal-to-Noise [7] used a measure of correlation that emphasizes the "Signal-to-Noise" ratio, *signaltonoise*, to rank the genes.

$$signaltonoise = \frac{(\mu_1 - \mu_2)}{(\sigma_1 + \sigma_2)} \qquad (2)$$

Where *μi* is the mean and *σi* is the standard deviation of the gene in class *i*.

3.3 Traditional t-test [2], *t-test* assumes that the values of the two tissues variances are equal. The formula is as

$$ttest = \frac{(\mu_1 - \mu_2)}{\sqrt{\left(\frac{\nu_p}{n_1}\right) + \left(\frac{\nu_p}{n_2}\right)}} \qquad (3)$$

where *μi* is the mean and *vp* is the pooled variance,

$$v_p = ((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(n_1 + n_2 - 2)\text{, and } s_i^2 = \sum_{i=1}^{n_i}(x_i - \overline{x})^2 \Big/ (n_i - 1) \qquad (4)$$

3.4 The Mann-Whitney Rank Sum Statistic[2], *mann*, has the following formula:

$$mann = n_1 * n_2 * \frac{(n_1 + 1)}{2} - r_1 \qquad (5)$$

Where *ni* is the sizes of sample *i*, and *r1* is the sum of the ranks in sample1.

These techniques are used because they look into the expression profiles of the genes in tumor and normal class [7]. In these techniques, each gene is measured for correlation with the class according to some measuring criteria in the formulas. The genes are ranked according to the score, *S*, and the top *T* numbers of genes are selected.

4. The Procedure for Gene Selection and Classification of Gene Expression Data

The procedure for this experiment is shown:
i. Getting the data.
ii. Setting the number of genes to be selected, *T*, the gene selection technique and the classifier. In this experiment, the number of genes to be selected is set to be from 1 to 100.
iii. Applying LOOCV technique for validation and evaluation purpose, include leaving one sample out in S3.1, selecting genes in S3.2 and S3.3 and training and testing the classifiers from S3.4 to S3.6.
iv. Calculating the ROC score based on the predicted class.
v. The process is repeated for another number of genes to be selected, another gene selection technique and another classifier until all combinations are done.

INPUT: Gene expression data matrix, X= $\{x_{11},…….,x_{np}\}$and the class label for each column, y C $\{-1,1\}$where n is the number of genes and p is the number of tissue samples.
S1. Get the data with *p* tissues (samples).
S2. Pre-set the combination: the gene selection technique, the classifier and number of genes to be selected, *T*, (the experiment run from 1 to 100 genes).

LEAVE ONE OUT CROSS VALIDATION:
S3. For i = 1 to *p*
S3.1 Leave i^{th} sample out.

GENE SELECTION:
S3.2 Calculate the discriminative score, *S,* for each gene for the remaining *p*-1samples, and rank the genes based on the score.
S3.3 Select top *T* genes based on the ranked score, *S*.

CLASSIFICATION:
S3.4 Train the classifier on the remaining *p*-1 samples by using the selected genes.
S3.5 Test the trained classifier by using the left out i^{th} sample.
S3.6 Record the predicted class from S3.5, put back the i^{th} sample.

ROC CALCULATION:
S4. Calculate the ROC score based on the predicted class and save the ROC score.
S5. Go to S2 for another number of genes to be selected, another gene selection technique and another classifier, stop if all combinations are done.

OUTPUT: ROC scores for each number of genes to be selected, *T* and gene selection technique.

5. Tissue Classification

Two classifiers are proposed to evaluate the validity of the selected genes. They are the SVM [1] with different kernels and the *k*- nn [6].

5.1 Support Vector Machines for Tissue Classification

Different kernel functions, the dot product and radial basis function are used for this experiment [4][5][8][1].
The dot product has the following formula:

$$K(x\ ,\ y) = (x \cdot y + 1)^d \qquad (6)$$

where x and y are the vectors of the gene expression data. The parameter d is an integer which decides a rough shape of a separator. In the case where d equals to 1, a linear classifier is generated, and in the case where d is equal to or more than 2, a nonlinear classifier is generated. In this experiment, when d is equals to 1, it is called the SVM dot product, when d is equals to 2, it is called the SVM quadratic dot product and when d is equals to 3, it is called the SVM cubic dot product.
The radial basis kernel has the following formula:

$$K(x,y) = \exp\left(\frac{-|x-y|^2}{2\sigma^2}\right) \qquad (7)$$

where σ is the median of the Euclidean distances between the members and nonmembers of the class.
The main advantages of SVMs are that they are robust to outliers, converge quickly, and find the optimal decision boundary if the data is separable. Another advantage is that the input space can be mapped into an arbitrary high dimensional working space where the linear decision boundary can be drawn. This mapping allows for higher order interactions between the examples and can also find correlations between examples.

SVMs are also very flexible as they allow for a big variety of kernel functions.

5.2 *k*-nearest neighbor for Tissue Classification

The *k*-nn classifier is a simple classifier based on a distance metric between the testing samples and the training samples [6]. The main idea of the method is, given a testing sample *s*, and a set of training tuples *T* containing pairs of the form (*ti*, *ci*) where *ti's* are the expression values of genes and *ci* is the class label of the tuple. Find *k* training sample with most similar expression value between *t* and *s*, according to a distance measure. The class label with the top voting among the *k* training sample is assigned to *s*. The main advantage of *k*-nn is it has the ability to model very complex target functions by a collection of less complex approximations. It is easy to program and understand. No training or optimization is required for this classifier. It is robust to noisy training data.

6. Result Evaluation Method

ROC score is used to analyze the results for the experiment. ROC score is also the area under the curve (AUC). ROC score is a common way for evaluating classification performance because it takes into account both false negative and false positive errors and it reflects the robustness of the classification. A random classification has a ROC score approaching 0.5 while a perfect classification with no error has a ROC score at 1. In this experiment, for each possible combination of number of genes to be selected, gene selection technique and classifier, the performance varying the number of genes from 1 to 100are evaluated.

6.1 Results and Discussion

In this section, the impact and importance of gene selection to the classification performance is first studied. This is carried out by comparing the classification performance by using all genes and gene selected by statistical techniques which are mentioned above. After that, the classification performance for each classifier is compared. Finally, based on the classifier with the best classification performance, the effectiveness of each statistical technique to this classifier is discussed.

6.2 Importance of Gene Selection Technique Prior to Tissue Classification

Figure-1 shows the classification performance by using all genes and gene selected by using statistical techniques. The ROC scores recorded for the gene selection techniques in the figure are the average ROC scores for number of genes selected from 1 to 100. From the figure, by using all genes, the best performance is obtained by using SVMs with radial basis function while *1*-nn, *2*-nn and *5*-nn have worst performance. *3*-nn and *4*-nn are comparable to each other when all genes are used. The performances of the classifiers are improved after genes are selected by gene selection techniques especially for *k*-nn classifier. This shows the importance of applying gene selection techniques to select informative genes prior to the classification task. Applying gene selection techniques in selecting genes helps in removing a large number of irrelevant genes which improves the classification performance. Since one of the advantages of SVMs is, it is robust to outliers and allows nonlinear classification to be done, gene selection techniques does not give big impact to its performance, but, a better performance still can be obtained after applying gene selection techniques, which can be seen from the figure. One might ask why there is still a need to do gene selection if the classification performance using SVM has little difference while using all the genes in the dataset compare to the selected subset of genes. One reason for this is that selecting subset of genes not only can help biologists to identify the potential genes rather than swimming in the huge dataset, it helps the classifier to build a better and simple rule for classifying future unknown data.

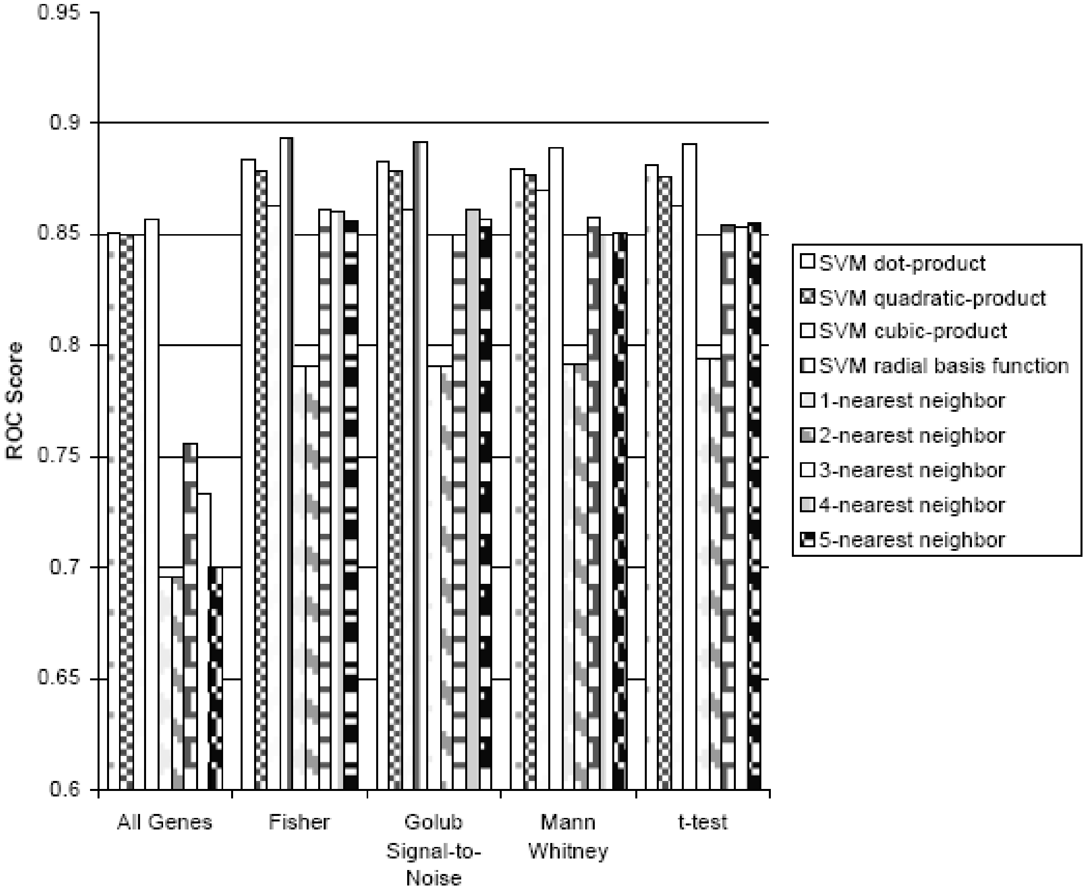

Figure 1: Classification performance by using all genes and genes selected by statistical techniques

.

This figure shows that a better classification performance can be achieved if genes are first selected by the gene selection techniques. However, which combination of statistical techniques and classifier and how many genes are needed for the best performance? Next section answers this question.

6.3 Classification Performance between Different Classifiers

Table-1 summarizes the performance for each SVM classifier. The ROC scores recorded in the table are the average ROC score over all trials with the number of selected genes from 1 to 100.

SVMs	Fisher	Golub	Mann	t-test
SVM_dot	0.88	0.88	0.88	0.88
SVM_quadratic	0.88	0.88	0.88	0.88
SVM_cubic	0.86	0.86	0.87	0.86
SVM_RBF	0.89	0.89	0.89	0.89

Table-1: Summary for classification performance by using SVMs with different kernels after gene selection by using statistical techniques

Table-1 show that, SVM radial basis function performs the best. Of the three, product kernels, dot-product and quadratic product have better ROC score than cubic-product. These results indicate that over-fitting causes the misclassification for the cubic-product kernel. If more samples are obtained and they are not separable linearly, nonlinear classification may perform well [3].

Table-2 summarizes the performance for each *k*-nn classifier. The ROC scores recorded in the table are the average ROC score over all trials with the number of selected genes from 1 to 100.

k-nn	Fisher	Golub	Mann	t-test
1-nn	0.79	0.79	0.79	0.79
2-nn	0.79	0.79	0.79	0.79
3-nn	0.86	0.85	0.86	0.85
4-nn	0.86	0.86	0.85	0.85
5-nn	0.86	0.86	0.85	0.85

Table-2: Summary for classification performance by using different *k*-nn after gene selection using statistical techniques

Table-2 show that *k*-nn with *k* more than 2 outperform *k* which is equals to 1 and 2. One of the reasons for this to happen is that in the case of mislabeled training samples, it will have much greater effect on the classification result of *1*-nn since one mislabel will result in misclassifying the test sample. *3*-nn and *4*-nn is less prone to bias in the data and more tolerable to noise since it makes use of several training samples to determine the class of a test sample.

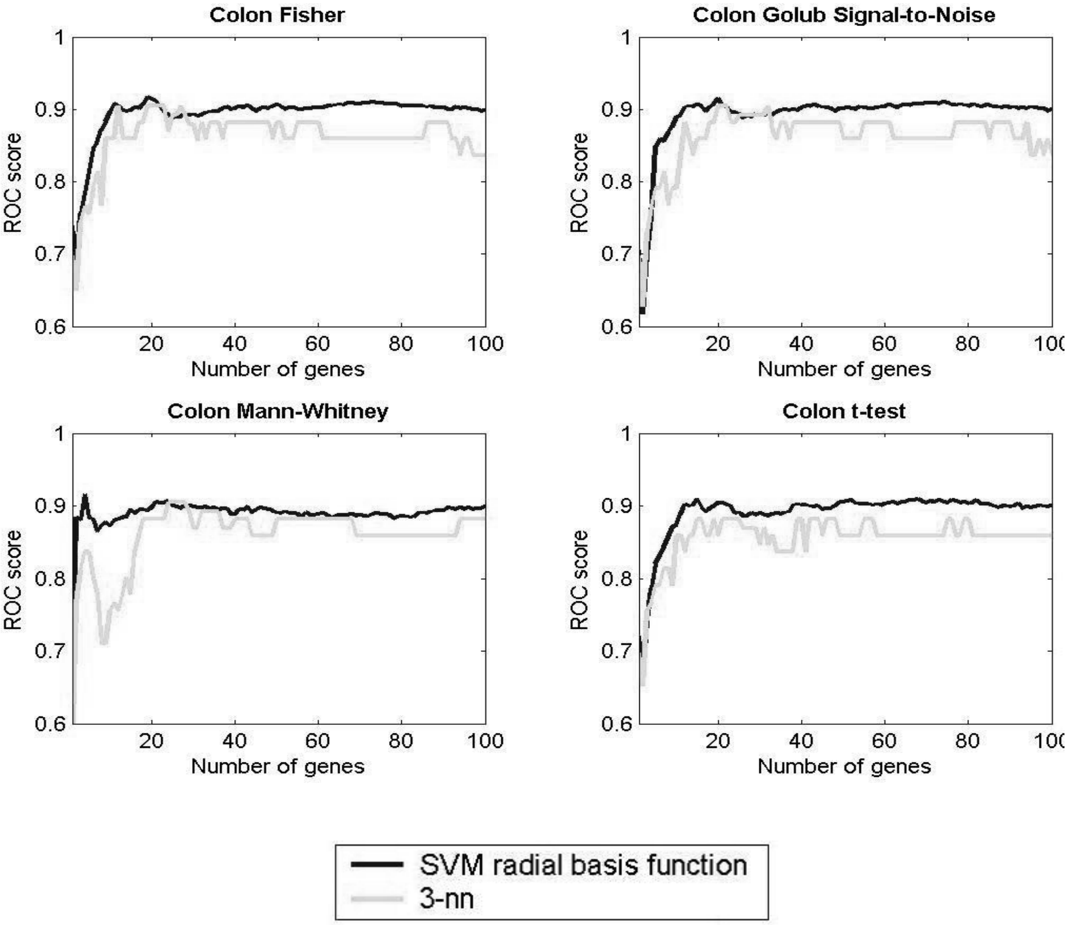

Figure-2: Classification performance between different classifiers after gene selection using statistical techniques (the best classifier is selected from SVM and *k*nn)

Figure-2 shows that SVM with radial basis function as the kernel function always produced higher ROC score than *3*-nn. Generally, the results have lower ROC score with fewer genes for both classifiers. Lowest scores always drop between the numbers of genes from 1 to 15 except for Mann-

Whitney Rank Sum Statistic. One reason for the lower scores might due to the characteristic of genes itself where genes do not act alone, but they interact with other genes for certain functions. For example, if Gene A and Gene B are in the same function it could be that they have similar regulation and therefore similar expression profiles. If Gene A has a good discriminative score it is highly likely that Gene B will, as well.

Hence the statistical techniques are likely to include both genes in a classifier, yet the pair of genes provides little additional information compared to either gene alone. If there are 5 functions in the dataset, 10 genes for each function, and if the genes in first function have the highest scores, so these 10 genes might be selected for the classification task. In this case, the genes being selected are highly redundant and thus provide little additional information. The peak performance for SVMs and *k*-nn always drop from the number of genes between 15 and 30. When the number of genes increase from 30 to 80 generally, the ROC score for SVMs and *k*-nn becomes more stable, because the possibility to select meaningful genes increase.

7. Summary

This paper reports the application of different statistical techniques to the colon dataset. These techniques include Fisher Criterion, Golub Signal-to-Noise, traditional t-test, and Mann-Whitney Rank Sum Statistic. By using these techniques, the data is rank based on the discriminative score and top *T* numbers of genes are selected. In conjunction with these gene selection techniques, several SVMs and *k*-nn classifiers are applied. Based on the genes selected by the gene selection techniques, ROC score of different combination of gene selection techniques and classifiers are obtained for analysis. The main objective of this experiment is to study the impact and importance of applying gene selection techniques prior to the classification task. Results show that a better classification performance is achieved by the classifiers if informative genes are first selected. However, finding a way to reduce redundant genes being selected in order to obtain a better classification performance is important.

8. Reference

[1] Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines And Other Kernel-Based Learning Methods.* New York: Cambridge University Press.

[2] Devore, J.L. (1995). *Probability and Statistics for Engineering and the Sciences.* 4th edition, California: Duxbury Press.

[3] Domura, D., Nakamura, H., Tsutsumi, S., Aburatani, H. and Ihara, S. (2002). Characteristics of Support Vector Machines in Gene Expression Analysis. *Genome Informatics.* (13):264 – 265.

[4] Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison Of Discrimination Methods For The Classification Of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*. 97(576): 77 – 87.

[5] Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002). Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments. *Statistica Sinica.* (12):111 – 139.

[6] Friedman, M. and Kandel, A. (1999). *Introduction to Pattern Recognition.* London: Imperial College Press.

[7] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caliguiri, M.A., Bloomfield, C.D. and Lander, E.S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science.*(286):531 – 537.

[8] Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J.P. and Poggio, T. (1999). Support Vector Machine Classification of Microarray Data. *S. Technical Report 182.* AI Memo 1676, CBCL.

[9] Smiatacz, M., Malina, W., Versatile Pattern Recognition System Based On Fisher Criterion. ul. G. Narutowicza 11/12, 80-952 Gdańsk