Chinese Science Bulletin 2004 Vol. 49 No. 15 1—6

Rank sum method for related gene selection and its application to tumor diagnosis

DENG Lin[1], MA Jinwen[1] & PEI Jian[2]

1. Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, China;
2. Department of Computer Science and Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260-2000, USA

Correspondence should be addressed to Ma Jinwen (e-mail:jwma@math.pku.edu.cn)

Abstract **Tumor diagnosis by analyzing gene expression profiles becomes an interesting topic in bioinformatics and the main problem is to identify the genes related to a tumor. This paper proposes a rank sum method to identify the related genes based on the rank sum test theory in statistics. The tumor diagnosis system is constructed by the support vector machine (SVM) trained on the set of the related gene expression profiles. The experiments demonstrate that the constructed tumor diagnosis system with the rank sum method and SVM can reach an accuracy level of 96.2% on the colon data and 100% on the leukemia data.**

Keywords: gene expression profiles, rank sum method, support vector machines, tumor diagnosis, gene selection.

DOI: 10.1360/03te0143

With the rapid development of DNA micro-array technology, more and more gene expression profiles become available and accurate. These biological data can be used to analyze human health status and disease factors. Nowadays, how to reveal valuable information from gene expression profiles has become an important topic in the bioinformatics community.

Gene expression data is often expressed by a matrix $W = (w_{ij})_{n\times m}$ which can be described by Fig. 1. Usually, a row in the matrix represents a gene and a column represents a sample (i.e. an instance). The numeric value w_{ij} characterizes the mRNA expression level of a specific gene "*i*" in a particular sample "*j*". Through analyzing gene expression data, biologists can obtain abundant valuable biological information. In recent years, gene expression analysis methods have been applied to broad areas, such as tumor classification, diagnosis as well as gene function analysis. The basic techniques for gene expression analysis include clustering, classification and principle component analysis (PCA). In particular, tumor diagnosis based on gene expression data has become an important research topic in bioinformatics[1—6]. In 1999, Golub et al.[1] first used a nearest neighbor analysis method to classify leukemia based on gene expression data. In his paper, he used a simplified formula of t-statistic as the tumor discrimination criteria to select related genes. In the same year, Alon et al.[2] clustered the colon gene expression to find a relationship between genes and tumors with the t-statistic to identify related genes. In 2000, Brown et al.[3] applied several classification techniques to tumor classification and compared the results. It was reported that support vector machine (SVM) is the best. Similar results are also verified by the studies of Dudoit et al.[4], Furey et al.[5], and Guyon et al.[6].

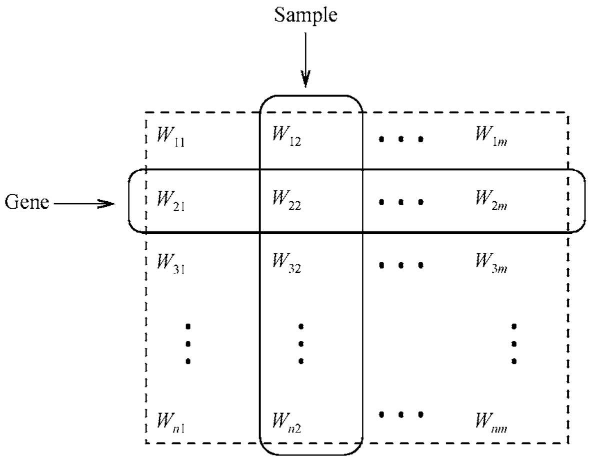

Fig. 1. A gene expression matrix**.**

All these studies have shown that tumor classification and diagnosis on gene expression profiles is feasible and reliable. However, if the gene expression data is not preprocessed to leave out noises before being inputted into classifiers, the result is often unsatisfactory. In that case, the generalization ability of the tumor classifier is bad (testing error is high), even for the SVM. The reason is that the gene expression profiles are usually containing too much noise. There are often thousands of genes in a typical gene expression profile, but probably only a small part of genes are highly related to the tumor phenotype under investigation. If the unrelated genes are not filtered out, the huge number of data dimensions would make classification difficult; furthermore, the unrelated genes would become noise and affect the classifier. Some gene selection methods have already been proposed to identify the related genes[1,2,5—8]. The most extensively applied methods are the t-statistic method and its variants. However, the t-statistic method is based on the theory of t-test in statistics. As is well-known, t-test is a parametric testing method that assumes that the samples follow Gaussian (normal) distribution. Consequently, the t-statistic method and its variants all explicitly or implicitly assume that the normality condition holds. However, in our investigation, this assumption is usually invalid in real applications.

To avoid the normality condition, we propose a rank sum method to identify the related genes, based on the non-parametric statistical testing theory. Then, we use SVM trained on identified related genes to construct the tumor diagnosis system. Verified by experiments on two data sets, our rank sum method can provide SVM with a very good generalization ability.

1 Gene selection and tumor diagnosis

(ⅰ) Statistics analysis on related gene selection.

Researches on related gene discovery have been carried out for a long time. But before DNA micro-array technology emerged, previous researches were based on biological characteristics. In the past five years, new techniques based on statistical analyzing gene expression data have been proposed. Most of the statistical methods introduce some discrimination criteria to select genes with large discrimination criterion value.

In particular, t-statistic and its variants are currently the most extensively used discrimination criterion. The expression of t-statistic is $T = \frac{m_{i,+} - m_{i,-}}{S_w\sqrt{\frac{1}{n_+}+\frac{1}{n_-}}}$, Where $S_w^2 = \frac{(n_+ - 1)S_{i,+}^2 + (n_- - 1)S_{i,-}^2}{n_+ + n_- - 2}$, , and $m_{i,+}$, $m_{i,-}$, $S_{i,+}$ and $S_{i,-}$ are the mean and standard deviation of gene "*i*" on the positive and negative samples, respectively; n_+ and n_- are the number of samples in the positive and negative class respectively. Actually, the t-statistic is used in a two-sample t-test to measure how large difference two Gaussian populations have. So, the larger is the absolute value of t-statistic, the more significantly the gene expression varies in different phenotypes. That means that the gene with a large t-statistic value, is highly related to the tumor factor. According to this intuition, the t-statistic method selects the top K genes with the highest absolute t-statistic values.

Some researchers also proposed several simplified expressions of t-statistic. For example, Golub et al.[1] used $W_i = \frac{m_{i,+} - m_{i,-}}{S_{i,+} + S_{i,-}}$ as his discrimination criterion, and selected the same number of genes with positive and negative W_i values. Furthermore, Furey et al.[5] used the absolute value of W_i and Pavlidis et al.[7] used the quadratic formula of W_i as their discrimination criterion. The t-statistic and its variants are essentially the same, because all their numerators are the difference in means between different phenotypes, and the denominators are expressions of variances for normalization.

Ding[8] proposed the F-statistic discrimination criterion as a general formula of t-statistic to deal with the multi-class (more than two tumor classes) gene selection problem. Suppose $g = (g_1, g_2, \cdots, g_n)$ is the expression of a gene, *K* is the number of tumor classes, the *F*-statistic expression is $F = \left[\sum_k n_k(\bar{g}_k - \bar{g})^2/(K-1)\right]/S^2$, where $\bar{g}$ and $\bar{g}_k$ are the means of expression level of the entire sample set and the K_{th} class sample set respectively, n_k and S_k are the means and standard deviations of the K_{th} class sample set, $S^2 = \left[\sum_k (n_k - 1)S_k^2\right]/(n-K)$. When $K = 2$, $F = t^2, t = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\frac{\bar{g}_1 - \bar{g}_2}{S}$, the *F*-statistic degenerates to the t-statistic.

As mentioned before, the t-statistic method and its variants are based on the t-test in statistics. However, the t-test is a parametric testing method and requires samples to follow the Gaussian distribution. Can we still use the t-statistic to select related genes if the gene expression does not follow a Gaussian distribution? From previous studies as well as our investigation, generally the t-statistic method still works, but it does not work as well as the normality condition holds. In that case, the t-statistic can loosely reflect how large is the difference in distributions between different phenotypes.

Two problems arise when the normality condition is violated. On the one hand, the order of genes in t-statistic may not reflect their true capability of discriminating phenotypes. For instance, suppose a gene A follows the normal distribution and a gene B follows a uniform distribution within an interval. Then, gene A's t-statistic value can be larger than gene B's. Consequently gene A ranks higher than gene B in related gene selection. This may lead to a wrong order of genes. The key is that gene B's p-value[9], which reflects the true discriminating capability of gene B, should be calculated according to uniform distribution instead of normal distribution. Then, gene B may rank higher than gene A according to their p-values. In short, blindly applying t-statistic to gene expression data that does not follow a Gaussian distribution may lead to errors. On the other hand, if the normality condition is violated, the t-statistic will not follow the t-distribution any more. So we cannot get the p-value of a gene from the t-distribution table, which means that we cannot use the significance level to select the related genes. Therefore, the users have to specify the related gene number directly, which is difficult to them. As the theoretical analysis, the normality condition limits the applicable area of the t-statistic method and its variants.

To test whether gene expression data often follows the Gaussian distribution in practice, we use the Skewness and Kurtosis statistics[9] to conduct the normality test on three well-known real data sets. They are Colon data set1),

1) Colon data, 40 tumor samples, 22 normal samples, 2000 genes, from http://www.molbio.princetion.edu/colondata.

Breast data set[2)] and Leukemia data set[3)]. The null hypothesis is that a gene satisfies the normality condition. We choose a significant level 0.05. That is, the error rate of mis-rejecting a gene that actually satisfies the normality condition is smaller than 5%. The results are shown in Table 1. As can be seen, nearly half of the genes' null hypotheses are rejected. Even for those genes whose null hypotheses are not rejected, they still may not follow the Gaussian distribution. It is clear that the three data sets do not satisfy the normality condition. In general, the gene expression data may not satisfy the normality condition. When the normality condition does not hold, we need a better related gene selection method to substitute the t-statistic method. The rank sum method is such a kind of method.

(ii) Rank sum gene selection method. To avoid the normality condition, we introduce a rank sum method for related gene selection based on the rank sum test theory[10] in non-parametric statistics. The non-parametric statistical methods have a distribution free property, so the normality condition is not necessary for them. The theory of non-parametric statistics also prove that the Pitman efficiency of rank sum test is much higher than the t-test, when the normality condition is violated[11]. For the gene selection problem, a higher Pitman efficiency means that the discrimination criterion of rank sum method is more reliable in terms of reflecting the discriminating capability of genes.

The general idea of rank sum method is that we conduct rank sum test on each gene, then select the genes whose null hypotheses are rejected, which means the genes are highly related to the tumor factors. The idea of rank sum test is that, instead of using the original observed data, we can list the data in the value ascending order, and assign each data item a "rank", which is the place of the item in the sorted list. Then, the ranks are used in the analysis. Using the ranks instead of the original observed data makes the rank sum method much less sensitive to outliers and noises than the t-statistic method. An outlier will change the t-statistic value greatly, but not much to the ranks. A gene expression data set often has many outliers and noises. Thus, the rank sum method is expected to be better than the t-statistic method for gene selection.

For different number of phenotypes, there are two types of rank sum test: Wilcoxon rank sum test and Kruskal-Wallis rank sum test. The former one is used to solve the two phenotypes related gene selection problem, and the latter one is for multi-phenotypes related gene selection. In this paper, we focus on two phenotypes problem. The major step of the Wilcoxon rank sum test is described as follows.

(1) Building hypotheses. Build the hypotheses as: null hypothesis H_0: the distributions between different phenotypes are the same; alternative hypothesis H_1: the distributions between different phenotypes are different. The significance level threshold a is specified.

(2) Combining and ranking. Combine all observations from the two populations and rank them in value ascending order. If some observations have tied values, we assign each observation in a tie their average rank. For instance, there are k observations having the same value. They are ranked at the positions from n+1 to n+k, then each of then will be assigned $n+\dfrac{k+1}{2}$ as their ranks.

(3) Computing Wilcoxon statistics. Add all the ranks associated with the observations from the smaller group (with the sample size n_1). This gives the Wilcoxon statistics W. If the null hypothesis holds, the expectation value of W should be $\dfrac{(n_1+n_2+1)}{2}n_1$, where n_2 is the sample size of the other phenotype. Therefore, if the value of W is greatly different from the expected value, then the probability of null hypothesis is small. If the value of W is out of a certain bound, the null hypothesis will be rejected.

(4) Testing. After computing the Wilcoxon statistics, we can use the Wilcoxon rank sum distribution table or a statistics toolkit, such as Matlab or SAS to get the associated p-value. Then, we compare the p-value with the specified significance level threshold. If the p-value is smaller than the significance level α, the gene will be selected.

Generally, for a multi-class gene selection problem (more than two phenotypes), the Kruskal-Wallis rank sum test can be used, which is an alternative to the F-statistic method proposed by Ding[8]. We omit the details here.

(iii) Tumor classification method. To verify the effectiveness of rank sum gene selection method, we build classifiers using support vector machines (SVM)[12]. As indicated by the previous studies, SVM is one of the best

Table 1 The results on normality test on three real data sets

	Colon cancer		Breast cancer		Leukemia	
	normal	tumor	normal	tumor	aLL	AML
Total genes	2000	2000	5776	5776	7129	7129
Rejected genes	730	1483	2250	2474	4542	2558

2) Leukemia data, 47 acute lymphoblastic leukemia (ALL), 25 acute myeloid leukemia (AML), 7129 genes, from http://www.genome.wi.mit.edu/MPR/data_set_ALL_AML.html.

3) Breast data, 13 normal samples, 14 tumor samples, 5776 genes, from http://genome-www.stanford.edu/sbcmp.

classifiers for the gene expression data. Here we review some basic ideas of SVM.

Figure 2 shows an example of SVM in a two dimension case. The hollow circles and the filled circles represent positive samples and negative samples, respectively, while H is the classification boundary. H_1 and H_2 are two lines parallel to H and pass through the samples which are nearest to H. The distance between H_1 and H_2 is called the margin. The optimal classification boundary is the line which can not only separate all samples of different classes correctly but also maximize the margin. When H is the optimal classification boundary, the samples on H_1 and H_2 are called support vectors. The process of building SVM is just the process of finding the optimal classification boundary.

For a general case, let us suppose the classification boundary is $\omega x + b = 0$. After normalization, the problem becomes: for a linear separable sample set $\{(x_i, y_i),$ $i = 1, 2, \cdots, N, x_i \in R^d, y_i \in \{\pm 1\}\}$, with the constrain $y_i(\omega x_i + b) - 1 \geqslant 0, i = 1, 2, \cdots, N$, we want to maximize the margin 2/||w||. In order to do so, we minimize $f(w) = \frac{1}{2} \| w \|^2$ under the above constrain.

By utilizing the Lagrange multiplier technique, we can convert the original problem into a dual problem. That is, maximizing the objective function $Q(a) = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i,j=1}^{N} a_i a_j y_i y_j (x_i x_j)$, with the constrain $\sum_{i=1}^{N} y_i a_i = 0,$ $a_i \geqslant 0, i = 1, 2, \cdots, N$. This becomes a quadratic optimization problem, and has a unique solution. It can be solved that the optimal classification boundary is $f(x) = Sgn(\sum_{i=1}^{N} a_i^* y_i (x_i x) + b^*)$. It is also easy to prove

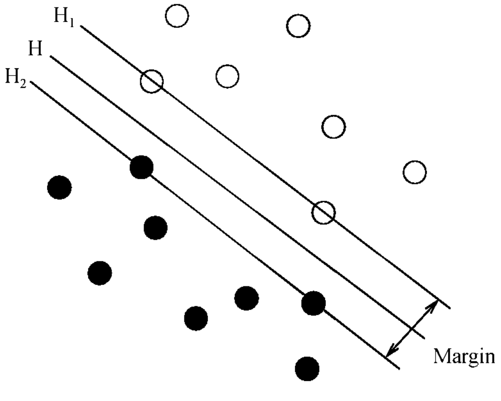

Fig. 2. Support vector machines.

that only a small part of a_i^* is not equal to zero. The corresponding samples actually are support vectors. b^* is a threshold, which can be derived from the support vectors.

For the case of not linear separable, we can use some non-linear mapping to transfer the original problem to a linear separable problem in a high-dimensional feature space, and find the optimal classification boundary in the feature space. As we only use the inner products in the original problem, we can perform the transformation by replacing the inner product with a kernel function $K(x, x')$. After the transformation, the objective function becomes $Q(a) = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i,j=1}^{N} a_i a_j y_i y_j K(x_i, x_j)$ and the optimal classification boundary becomes $f(x) =$ $\mathrm{Sgn}(\sum_{i=1}^{N} a_i^* y_i K(x_i, x) + b^*)$. There are some common used kernel functions including: the polynomial kernel $K(x, x_i) = [(xx_i) + 1]^q$, the radial basis function kernel $K(x, x_i) = e^{-\frac{|x - x_i|^2}{S^2}}$, and the sigmoid function kernel K $(x, x_i) = \tanh(v(xx_i) + c)$.
For our tumor diagnosis task, the tumor diagnosis system is constructed by SVM trained on the set of the related genes selected by the rank sum method. And the trained SVM is used to predict tumors on the testing data.

2 Experiments

(i) Evaluation of rank sum method. To evaluate the effectiveness of the rank sum method for gene selection, we conduct tumor diagnosis experiments on two real data sets: the colon data set and the leukemia data set, using three kinds of SVM. The programs are implemented with Matlab6.5 and SVM^{light} toolkit. The SVM^{light} toolkit is free SVM software downloadable at http://svmlight.joachims.org. We firstly select the related genes under some commonly used significance levels. The results are shown in Table 2.

After selecting related genes, we compare the accuracy of the SVM classifiers with and without gene selection. We use three kinds of SVM: linear SVM, cubic polynomial SVM (3-poly SVM) and radial basis function SVM (RBF SVM). To make the test more robust, we conduct the 4-fold cross-validation experiments. In particular, we randomly divide the colon data set that includes 40 tumor samples and 22 normal samples into 4 folds: each fold contains 10 tumor samples and 5 or 6 normal samples. Similarly, we randomly divide the leukemia data set with 47 ALL samples and 25 AML samples into 4 folds: each fold contains 18 samples—11 or 12 ALL samples and 6 or 7 AML samples. Then, we try our

Table 2 Related gene numbers under different significance level

	Original	a =0.1	a =0.05	a =0.01	a =0.001
Colon data	2000	210	109	34	8
Leukemia data	7129	1837	1425	844	398

experiment four times, each time we use three folds as the training data set and one fold as the testing data. Finally, we compute the average accuracy for the 4 results as our evaluation result. We use the prediction accuracy as our evaluation metric.

Furthermore, we make the SVM^{light} toolkit accept all testing patterns. That is, no testing patterns are rejected without labels. Also before plugging the data into SVM^{light}, we normalize the original expression data such that the mean is zero and the standard deviation is one. The prediction results are shown in Table 3 and Table 4, respectively.

From these tables, we can see that the effectiveness of the related gene selection is significant: using the related genes selected by rank sum method improves the accuracy dramatically. On the colon data, the best accuracy achieved by using the related genes is 98.3%, where only one prediction error happens in the 4-fold cross-validation experiment. On the leukemia data, the accuracy of the related gene based method is even 100%. The accuracy indicates that the tumor diagnosis based on rank sum method is applicable.

From the result, it can also be seen that specifying the significance level is critical to get a high accuracy. On the one hand, a too large significance level will not filter out all unrelated genes, which are noises to the classifier; on the other hand, a too small significance level may filter out some useful information, and the classifier also may not achieve a high accuracy. From the results of the colon data and leukemia data, we can see that 0.01 is a proper significance level to get high accuracy, which can be referenced by other data sets.

(ⅱ) Comparison with t-statistic method. As analyzed theoretically in the previous section, if gene expression data does not follow the Gaussian distribution, the rank sum method is more reasonable and reliable than the t-statistic method for related gene selection. Here, we conduct experiments to compare the rank sum method with the t-statistic method. For comparison, we select the same number of related genes as the rank sum method does at different significance levels for t-statistic method. All settings of this experiment are the same as those of the evaluation experiment of rank sum method. The comparison results of prediction accuracy (averaged among 4-fold cross validation and three kinds of SVM) are shown in Table 5.

The results clearly show that the rank sum method is consistently better than the t-statistic method. This concurs with our theoretical analysis. The improvement in accuracy of the rank sum method against the t-statistic method

Table 3 The result on the colon data, a is significance level

	Original(2000)	a = 0.1(210)	a = 0.05(109)	a = 0.01(34)	a = 0.001(8)
Linear SVM	56.4%	90.3%	90.3%	95.1%	88.8%
3-poly SVM	31.3%	61.3%	90.3%	95.1%	88.8%
RBF SVM	45.3%	87.2%	93.6%	98.3%	88.8%
Average	44.3%	79.6%	91.4%	96.2%	88.8%

Table 4 The result on the leukemia data, a is significance level

	Original(7129)	a =0.1(1837)	a =0.05(1425)	a =0.01(844)	a =0.001(398)
Linear SVM	73.6%	94.4%	94.4%	100%	100%
3-poly SVM	55.6%	91.7%	91.7%	100%	95.8%
RBF SVM	52.8%	94.4%	100%	100%	100%
Average	60.7%	93.5%	95.4%	100%	98.6%

Table 5 Comparison of rank sum method and t-statistic method

Data set	Method	210 genes	109 genes	34 genes	8 genes
Colon data	Rank sum	79.6%	91.4%	96.2%	88.8%
	t-statistic	75.9%	90.3%	93.6%	88.8%
		1837 genes	1425 genes	844 genes	398 genes
Leukemia data	Rank sum	93.5%	95.4%	100%	98.6%
	t-statistic	92.6%	91.7%	95.4%	94.4%

is about 1%—5%, except one case where only eight related genes are selected on the colon data with significance level of 0.001. (The reason may be too few genes to compare.) The improvement is non-trivial, considering the t-statistic method also achieves accuracy above 90%. According to our normality test, both the colon data and leukemia data have many genes violating the normality condition. So the statistics theory guarantees that the rank sum method outperforms the t-statistic method in these two data sets.

3 Conclusions

We have investigated the related gene selection problem for tumor diagnosis. The t-statistic method and its variants are the state-of-the-art gene selection methods. However, this kind of methods requires that the gene expression data follows the Gaussian distribution, which is often violated in real data sets according to our investigation. In this paper, we propose the rank sum method for related gene selection, which does not require the normality condition, and therefore can be applied to any gene expression profiles. Moreover, we use SVM trained on the identified related genes to construct the tumor diagnosis system. The experiment results show that the rank sum method and the tumor diagnosis system are effective. The constructed tumor diagnosis system with the rank sum method and SVM can reach an accuracy of 96.2% on the colon data and 100% on the leukemia data. It is also demonstrated by the experiment that the rank sum method is more effective than previous t-statistic method. All the results show that the rank sum method and the tumor diagnosis system is applicable in practice.

Acknowledgements This work was supported by the Natural Science Foundation of China (Grant No. 60071004).

References

1. Golub, T. R., Slonim, D. K., Tamayo, P. et al., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, Science, 1999, 286: 531—537.
2. Alon, U., Barkai, N., Notterman, D. A. et al., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Nat'l Acad. Sci.USA, 1999, 96: 6745—6750.
3. Brown, M. P. S., Grundy, W. N., Lin D. et al., Knowledge-based analysis of microarray gene expression data by using support vector machines, Proc. Nat'l Acad. Sci., 2000, 97(1): 262—267.
4. Dudoit, S., Fridyand, J., Speed T. P., Comparison of discrimination methods for the classification of tumor using gene expression data, Journal of American Statistical Association, 2002, 97(457): 77—87.
5. Furey, T., Cristianini, N., Duffy, N. et al., Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics, 2000, 16(10): 909—914.
6. Guyon, I., Weston, J., Barnhill, S.et al., Gene selection for cancer classification using support vector machine, Machine Learning, 2002, 46(1/3): 389—422.
7. Pavlidis, P., Weston, J., Cai, J.et al., Gene functional analysis from heterogeneous data, Proc. Fifth Int. Conf. on Computational Molecular Biology, ACM Press, 2001, 249—255.
8. Ding, H. Q., Analysis of gene expression profiles: class discovery and leaf ordering, Proc. RECOMB, 2002, 127—136.
9. Goulden, C. H., Methods of Statistical Analysis, 2nd ed., New York: John Wiley & Sons, 1956.
10. Hettmansperger, T. P., Statistical Inference Based on Ranks, John Wiley & Sons, Inc., 1984.
11. Nikitin, Y., Asymptotic efficiency of non-parametric tests, Cambridge University Press, 1995.
12. Vapnik, V., Statistical Learning Theory, Wiley, 1998.
13. Joachims, T., Making large-scale SVM learning practical, Advances in Kernel Methods-Support Vector Learning (eds. Schölkopf, B. et al.), MIT-Press, 1999.

(Received July 13, 2003; accepted April 29, 2004)