A Rank Sum Test Method for Informative Gene Discovery *

Lin Deng† Jian Pei‡ Jinwen Ma¶ Dik Lun Lee†

†Department of Computer Science, Hong Kong University of Science and Technology, Hong Kong, China
Email: {ldeng, dlee}@cs.ust.hk

‡Department of Computer Science and Engineering, State University of New York at Buffalo, USA
School of Computing Science, Simon Fraser University, Canada
Email: jianpei@cse.buffalo.edu

¶School of Mathematical Sciences and LMAM, Peking University, Beijing, 100871, China
Email: jwma@math.pku.edu.cn

ABSTRACT

Finding informative genes from microarray data is an important research problem in bioinformatics research and applications. Most of the existing methods rank features according to their discriminative capability and then find a subset of discriminative genes (usually top k genes). In particular, t-statistic criterion and its variants have been adopted extensively. This kind of methods rely on the statistics principle of t-test, which requires that the data follows a normal distribution. However, according to our investigation, the normality condition often cannot be met in real data sets.

To avoid the assumption of the normality condition, in this paper, we propose a rank sum test method for informative gene discovery. The method uses a rank-sum statistic as the ranking criterion. Moreover, we propose using the significance level threshold, instead of the number of informative genes, as the parameter. The significance level threshold as a parameter carries the quality specification in statistics. We follow the Pitman efficiency theory to show that the rank sum method is more accurate and more robust than the t-statistic method in theory.

To verify the effectiveness of the rank sum method, we use support vector machine (SVM) to construct classifiers based on the identified informative genes on two well known data sets, namely colon data and leukemia data. The prediction accuracy reaches 96.2% on the colon data and 100% on the leukemia data. The results are clearly better than those from the previous feature ranking methods. By experiments, we also verify that using significance level threshold is more effective than directly specifying an arbitrary k.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Algorithms, measurement

Keywords

Informative genes, ranking criterion, rank sum test, support vector machine, classification

1. INTRODUCTION

The DNA microarray technology enables rapid, large-scale screening for patterns of gene expression. A DNA microarray experiment can provide simultaneous, semi-quantitative readouts on expression levels for thousands of genes [3]. The raw microarray data is transformed into gene expression matrices. Figure 1 shows an example. Usually, a row in the matrix represents a gene and a column represents a sample. The numeric value in each cell characterizes the expression level of a specific gene in a particular sample. Many data sets are now available on the web (e.g., [19, 2, 8]).

In many microarrary data sets, the samples can be divided into several subgroups, such as tumor tissues and normal tissues (cases and controls). Each subgroup is called a *phenotype*. A critical task is to identify the *informative genes* – the genes that are discriminative among different phenotypes. To elaborate, suppose there are 8 samples in a data set. Among them, 4 samples are tumor tissues and the other 4 samples are normal tissues, as shown in Figure 2. The expression levels of 5 genes are illustrated. Genes g_1,

* This research is supported in part by Research Grant Council, Hong Kong SAR, China, under grant number HKUST6225/02E, National Science Foundation of the United States under grant number IIS-0308001, and National Natural Science Foundation of China under grant number 60071004. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies. This research was partly conducted at Peking University, when the first author studied there.

KDD'04, August 22-25, 2004, Seattle, Washington, USA.

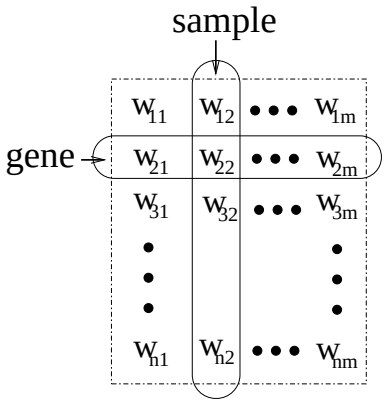


Figure 1: A gene expression matrix.

g_2 and g_3 are informative, since they are more or less consistent within each phenotype, and discriminative between the two phenotypes. On the other hand, genes g_4 and g_5 are non-informative genes, since they cannot manifest any difference between the two phenotypes.

There are often thousands of genes in a data set, but probably only a small subset of genes are highly related to the phenotypes under investigation. Therefore, the discovery of informative genes as feature selection from an extremely high dimensional space has broad applications. For example, to construct an effective classifier for tumor diagnosis, the informative genes should be used. The non-informative genes should be filtered out so that the classification can avoid noises. Moreover, informative genes can also provide insights into the causes of phenotypes. The investigation on the informative genes may further indicate the inherent factors that lead to the differences between the cases (e.g., the tumor tissues) and the controls (e.g., the normal tissues), or between different subtypes of tumor tissues.

One may think, given a set of labeled samples, finding the informative genes may be straightforward. We only need to find the genes that are consistent within each phenotype and very different between the phenotypes. Unfortunately, the cases are far from that simple. The real microarray data sets are always very noisy. Almost no gene is perfectly consistent in each phenotype and sharply different between phenotypes. Instead, from the very noisy data, we have to find out the genes that are very likely to be informative.

The problem of informative gene discovery has been studied extensively in the last 5 years. It has become increasingly clear that, for many data analysis tasks on microarray data such as classification and clustering, instead of considering many genes, it is effective to consider a small number of informative genes [15]. To select informative genes, a ranking criterion is often introduced to quantify the discriminative capability of individual genes. Then, a subset of genes with the highest ranking criterion values are selected as informative genes. This kind of methods are known as *feature ranking methods*.

Although several ranking criteria have been proposed, there exist two serious problems. First, many methods require a user-specified threshold on the number of informative genes. That is, they select the top k genes as the informative ones. However, it is often hard for a user to specify such a parameter. Second, some methods use the t-statistic or its variations as the selection criteria. The t-statistic requires that the data follows the Gaussian (normal) distribution. According to our investigation (see Section 3), the assumption about the data distribution often does not hold in gene expression data.

In this paper, we propose a novel ranking criterion for informative gene discovery based on non-parametric testing. In particular, we use rank sum test. By non-parametric testing, we do not assume any specific data distribution in gene expression data. Moreover, we use the significance level to select informative genes and thus provide the quality guarantee in statistics. To verify the effectiveness of our new method for informative gene discovery, we construct the support vector machine (SVM) classifiers using the identified informative genes. Our experimental results on two well known real data sets, the colon data set and the leukemia data set, show that the rank sum test method is effective: the SVM-based classifiers using the informative genes so identified are clearly more accurate than the previous methods.

The rest of the paper is arranged as follows. In Section 2, we describe the informative gene discovery problem concisely and review the related work. The rank sum test method is introduced in Section 3. The experimental results are reported in Section 4. Section 5 discusses related issues and Section 6 concludes the paper.

2. PROBLEM DESCRIPTION AND RELATED WORK

In this section, we describe the problem of informative gene discovery, and present a brief review of the related work.

2.1 Informative Gene Discovery

In general, for a set of genes G and a set of samples S, a *microarray data set* can be modeled as a two dimensional matrix $M = (m_{i,j})_{n\times m}$, where $m_{i,j}$ is the expression level of gene g_i on sample s_j.

Often, the microarray tests are conducted on controlled groups of samples. That is, the samples can be divided into certain groups (called phenotypes), such as the groups of normal tissues and tumor tissues. Typically, a sample has thousands of genes and there are only less than 100 samples in a microarray data set. Typical data analysis tasks on microarray data sets include classification, clustering and pattern discovery.

If we treat each sample as an object and each gene as a feature/dimension, the data analysis is in a very high dimensional space with thousands of dimensions. As indicated by previous studies, data analysis in such high dimensional spaces is usually ineffective and deficient. Therefore, it is desirable to select a small subset of genes such that each gene is discriminative among the subgroups of samples. Such genes are called *informative genes*.

Previous studies strongly indicate that the number of informative genes is usually much smaller than the total num-

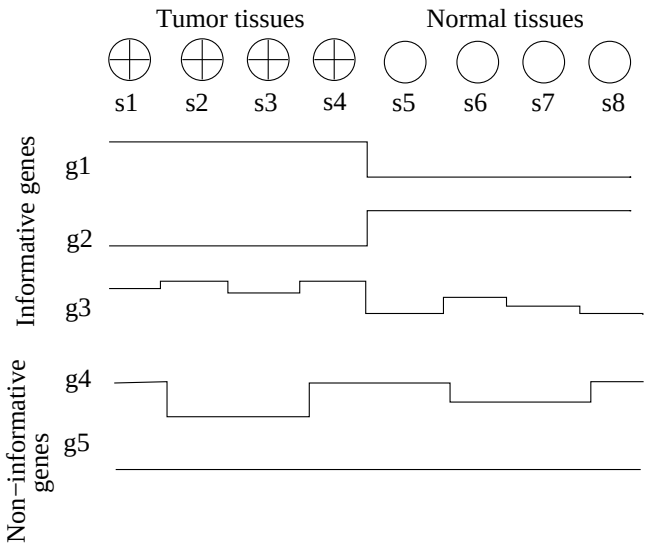


Figure 2: Informative genes

ber of genes. Data analysis using informative genes is effective and efficient.

Now, the problem becomes *how we can discover the informative genes effectively.* This is called the *informative gene discovery problem.*

2.2 Related Work

Most of the existing informative gene selection methods are based on feature ranking. A typical approach is in two steps. First, a ranking criterion is defined to measure the variance of the expression values in different phenotypes for each gene. Then, the top-k genes with the best ranking measure values are selected as the informative genes. In particular, the t-statistic ranking criterion and its variants are frequently used [1, 4, 7, 9, 17, 18].

The t-statistic [1] is defined as

$$T = \frac{\mu_{i,+} - \mu_{i,-}}{S_w \sqrt{\frac{1}{n_+} + \frac{1}{n_-}}},$$

where

$$S_w^2 = \frac{(n_+ - 1)\sigma_{i,+}^2 + (n_- - 1)\sigma_{i,-}^2}{n_+ + n_- - 2},$$

and $\mu_{i,+}$, $\mu_{i,-}$, $\sigma_{i,+}$ and $\sigma_{i,-}$ are the mean and standard deviation of gene g_i on the positive and negative samples, respectively; n_+ and n_- are the number of samples in the positive and negative classes, respectively.

Intuitively, the t-statistic measures the difference of means between different phenotypes, and the difference is normalized by an expression of variances. Actually, the t-statistic is used to measure the difference between two Gaussian distributions. Based on the t-test theory, when the data follows Gaussian distribution, we can also compute the p-values to know how significant the difference is. Then, we can use the significance level, which is a threshold of p-values, to determine a set of informative genes [4].

Some simplified forms of t-statistic are also used as ranking criteria. For example, Golub et al. [9] propose the following measure

$$w_i = \frac{\mu_{i,+} - \mu_{i,-}}{\sigma_{i,+} + \sigma_{i,-}},$$

where $\mu_{i,+}$, $\mu_{i,-}$, $\sigma_{i,+}$ and $\sigma_{i,-}$ are the mean and standard deviation of gene g_i on the positive and negative samples, respectively. This ranking criterion is also adopted in other studies [7, 18]. As another example, Pavlidis et al. [17] adapt the Fisher's discriminant criterion (FDC) [5], and define a criterion as

$$\frac{(\mu_{i,+} - \mu_{i,-})^2}{\sigma_{i,+}^2 + \sigma_{i,-}^2}.$$

These two criteria have similar formulas as the t-statistic. They are considered as variants of the t-statistic.

The t-statistic like methods take the assumption that *the data follows the Gaussian distribution.* However, if the assumption does not hold, two defections of the t-statistic criterion may happen. First, the ranks of the genes may not concur with their discriminative capabilities over phenotypes. Second, using significance level to determine the informative genes may become meaningless in statistics. A detailed discussion will follow in Section 3.1.

Some methods other than the t-statistic exist. For example, the square ratio is used by Dudoit et al. [6]. Moreover, Guyon et al. [11] propose a method that selects informative genes according to their utilization to SVM classifier. The method conducts the recursive feature elimination (RFE) based on a linear SVM classifier. A recent work by Yu et al. extends the RFE method to a polynomial-SVM [21]. However, selecting informative genes by a specific classifier may "overfit" the classifier and thus may not be helpful for other applications. A gene selection method independent from specific analysis will have more extensive applications. Moreover, the accuracy of the SVM classifier may not be an intuitive and direct indicator for the quality of informative genes.

3. THE RANK SUM TEST METHOD

In this section, we develop our rank sum test method. We first discuss why the rank sum test is introduced. Then, we revisit the general idea of rank sum test, and explain how it can be used in informative gene discovery.

3.1 Why Non-parametric Testing?

A *statistical test* is used to determine the statistical significance of an observation. With an assumption of the underlying distribution of the observed data, a *parametric test* can be conducted.

T-test [10] is a typical parametric test. The t-statistic ranking criterion is derived from the t-test, which is used to test whether two Gaussian populations have different distributions in statistics. For gene selection, a high variance of distribution in different phenotypes indicates that the gene has a strong capability of discriminating different phenotypes.

Now, the *key problem* becomes, "*Does the gene expression data often follow the Gaussian distribution?*"

We use three well-known real data sets, Colon cancer data set [2], Breast cancer data set [19] and Leukemia data set [8], to test if they follow the Gaussian distribution. We use the Skewness and Kurtosis statistics to conduct the normality test. The null hypothesis is that a gene satisfies the normality condition. We choose a significance level of 0.05. If the null hypothesis of a gene is rejected, then the gene cannot be considered to follow the normal distribution. The error rate of mis-rejecting a gene that actually satisfies the normality condition is smaller than the significance level of 0.05.

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.

From this test, it is clear that the three data sets do not satisfy the normality condition. The three data sets are typical. *In general, the gene expression data may not satisfy the normality condition.*

Can we still use the t-statistic if the gene expression does not follow a Gaussian distribution? Generally, the t-statistic still can loosely measure the difference of the distributions between the phenotypes. However, it does not work as well as when the normality condition holds. As discussed before, there are two problems when the normality condition is violated. First, the order of genes in t-statistic may not reflect their capabilities 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, the t-statistic value of gene A can be larger than that of gene B. Consequently, gene A ranks higher than gene B. This may lead to an error in informative gene selection. The key is that the p-value of gene B should be calculated under the assumption of *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 a wrong order of genes.

Furthermore, if the normality condition is violated, the t-statistic will not follow the t-distribution any more. So we should not get the p-value of a gene from the t-distribution table, that is, using the significance level to select the informative genes does not make sense in statistics any more.

3.2 The Pitman Efficiency Theory

When the underlying distribution of data is unknown, a non-parametric test should be conducted. The *Pitman efficiency theory* [16], which is also called *asymptotic relative efficiency* (ARE) theory, gives a good explanation on the advantage of the non-parametric method when the normality condition is violated. Here, we review the basic ideas of the Pitman efficiency.

In statistic test, two types of errors may happen.

- A *type I error* is the mis-rejection of a true null hypothesis;
- A *type II error* is a failure to reject a false null hypothesis.

The probability of the type I errors is just the significance level, while the probability of the type II errors is equal to one subtract the power—the probability of rejecting the null hypothesis correctly.

Generally, a statistic test is good if it has relatively small probabilities of both type I and type II errors. Given a certain number of samples, there is usually a tradeoff between the probability of type I errors and that of type II errors. When the sample size is growing to infinity, both probabilities can approach to 0.

Suppose α and β are the probabilities of type I errors and type II errors, respectively. For any statistical test T, there exists a large enough sample size N such that T satisfies the given α and β. The relative efficiency $RE(A, B)$ is defined as $\frac{N_B}{N_A}$, where N_A and N_B are the sample sizes required for test A and B to give the same accuracy in terms of α and β, respectively.

$RE(A, B) > 1$ means that test A is more accurate than test B. However, $RE(A, B)$ is influenced by many factors and is difficult to compute. So the *asymptotic relative efficiency* $ARE(A, B)$ is usually used, which is the limitation of $RE(A, B)$ when α is fixed and β is approaching 0, as an objective criterion to compare the efficiencies of different tests in statistics. It is proved that $ARE(A, B)$ provides a good approximation of $RE(A, B)$ [16] even with a small data set size.

The *Wilcoxon rank sum test* [14, 20] is a non-parametric alternative to the two-sample t-test. So, we are concerned about $ARE(W, t)$, which is the asymptotic relative efficiency of Wilcoxon rank sum test to t-test. If we use $f(x)$ to represent the underlying distribution function of the data, and use σ^2 to represent its variance, then it can be shown [16] that, generally,

$$ARE(W, t) = 12\sigma^2(\int f^2(x)dx)^2.$$

Table 2 shows some concrete $ARE(W, t)$ values for some common distributions. When the data follows the Gaussian distribution perfectly, the t-test is a bit superior to Wilcoxon

	Colon cancer		Breast cancer		Leukemia	
	normal samples	tumor samples	normal samples	tumor samples	ALL samples	AML samples
Total # of genes	2000	2000	5776	5776	7129	7129
Rejected genes	730	1483	2250	2474	4542	2558

Table 1: The results on normality test on three real data sets. ("AML" for acute myeloid leukemia and "ALL" for acute lymphoblastic leukemia.)

rank sum test. However, when the data follows some other distributions, the Wilcoxon rank sum test is much better than the t-test.

Table 3 shows an example of the $ARE(W,t)$ values, when normal population is partly contaminated. The standard normal distribution $\Phi(x)$ is contaminated with a different scale normal distribution $\Phi(x/3)$, the contaminated portion is ξ. The contaminated distribution function is

$$F_\xi(x) = (1-\xi)\Phi(x) + \xi\Phi(x/3).$$

In Table 3, we can see only contaminated portion 0.15 makes the t-test much worse than Wilcoxon rank sum test. Generally, under different situations, the $ARE(W,t)$ value varies in the interval $[0.864, +\infty]$. Thus, in most cases, the Wilcoxon rank sum test is superior to the t-test. For gene selection problem, the superiority means that the selected genes are more reliable in terms of having different distributions in different phenotypes. Therefore, it is more reasonable and reliable to use the rank sum test method for gene selection rather than the t-statistic method.

ξ	0	0.05	0.1	0.15
$ARE(W,t)$	0.955	1.196	1.373	1.497

Table 3: The $ARE(W,t)$ values for contaminated normal distribution $F_\xi(x) = (1-\xi)\Phi(x) + \xi\Phi(x/3)$.

3.3 Rank Sum Test

When the data does not follow a normal distribution, a distribution free *non-parametric test* should be conducted. The rank sum test is a big category of non-parametric tests. The general idea 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 test much less sensitive to outliers and noises than the classical (parametric) tests. 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, it is more suitable to apply the rank sum test on informative gene selection.

Depending on the number of classes in the data sets, we have different kinds of rank sum tests. The *Wilcoxon rank sum test* [20, 14] is a non-parametric alternative to the two-samples t-test. The *Kruskal-Wallis rank sum test* is used for multi-class testing.

In this paper, we focus on the Wilcoxon rank sum test. Given a data set of two classes. The statistic W is the sum of ranks of the samples in the smaller class. The major steps of the Wilcoxon rank sum test are as follows.

1. 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;
2. Add all the ranks associated with the observations from the smaller group. This gives the Wilcoxon statistic;
3. Finally, the p-value associated with the Wilcoxon statistic is found from the Wilcoxon rank sum distribution table, or a statistics toolkit, such as Matlab or SAS.

The method is demonstrated as follows.

EXAMPLE 1 (WILCOXON RANK SUM TEST). Suppose X and Y are the expression levels of a certain gene in normal and tumor samples, respectively. The normal set X contains 12 observed values and the tumor set Y contains 7 observed values, as shown in Table 4. We want to test whether null hypothesis $P(X > Y) = P(X < Y)$ holds, which means the distribution in normal samples is identical to the distribution in tumor samples.

We combine all the samples in X and Y, and sort them in the value ascending order. The ranks are assigned to samples based on the order. If k samples have the same value of rank i, then each of them has an average rank

$$(i + \frac{k-1}{2}).$$

The results are shown in Table 5.

Let n_1 and n_2 be the numbers of samples in the smaller and larger groups, respectively. The tumor set has a smaller sample size, 7, so $n_1 = 7$ and $n_2 = 12$. Then, we compute the sums of the tumor ranks and have the statistic $W = 1+3+4+6+8.5+11+16 = 49.5$.

If null hypothesis $H : P(X > Y) = P(X < Y)$ holds, then the statistic W should be around the expectation value

$$\frac{(n_1+n_2+1)\cdot n_1}{2} = 70.$$

If the value of W is either too large or too small, the null hypothesis is likely to be false. Using the Matlab, the p-value is computed as 0.0873. Thus, if we set the significance level α is $0.05 < 0.0873$, the null hypothesis H_0 cannot be rejected; but if we set the significance level α is $0.1 > 0.0873$, then the null hypothesis H_0 is rejected. □

Generally, for a multi-class informative gene discovery application (more than two phenotypes), the Kruskal-Wallis

Distribution	Distribution function	$ARE(W,t)$
$U(-1,1)$	$\frac{1}{2}I(-1,1)$	1
$N(0,1)$	$\frac{1}{\sqrt{2p}}e^{\frac{x^2}{2}}$	$\frac{3}{\pi} \simeq 0.955$
Logistic	$e^{-x}(1+e^{-x})^{-2}$	$\pi^{\frac{2}{9}} \simeq 1.097$
Double Exp.	$\frac{1}{2}e^{-\lvert x\rvert}$	1.5
Exponential	$\lambda e^{-\lambda x}$	3

Table 2: The $ARE(W,t)$ values for some common distributions.

Types	Num	Expression level											
X (nor)	12	134	146	104	119	124	161	107	83	113	129	97	123
Y (tum)	7	70	118	101	85	107	132	94					

Table 4: Gene expression levels in two phenotypes.

rank sum test, which is a generalization of the Wilcoxon rank sum test, can be used. Limited by space, we omit the details here.

3.4 Selecting Informative Genes by Significance Level Threshold

Many previous methods select the top k genes as the informative ones, where k is a parameter specified by the user. As discussed before and will be illustrated using real data later, the value k is very different from one data set to the other. Thus, it is very hard for a user to guess the right value.

Instead of guessing a magic number, is it possible that the user specify the quality requirement on the informative gene selection? Here, we propose a *significance level threshold* approach.

In statistics, the significance level measures the probability of type I errors. Thus, a user can use the significance level to specify the quality requirement. By this idea, our informative gene selection method takes a *significance level threshold* α_{max}. Only the genes whose p-values that are less than the threshold are selected.

The major steps of the informative gene selection are as follows.

1. Specify a significance level threshold α_{max} (e.g. 0.01), to indicate the quality requirement of informative gene selection;
2. Compute the Wilcoxon-statistic for every gene;
3. Use the statistics to compute the corresponding p-values;
4. Select the genes whose p-values are smaller than the significance level threshold α_{max}, which means the distributions between phenotypes are not identical.

3.5 Verification

To verify the effect of our rank sum test method for informative gene discovery, we build classifiers using support vector machines (SVM). As indicated by the previous studies, SVM is capable of classification on high dimensional data sets with a small number of samples.

4. EXPERIMENTS

We test the effectiveness of the rank sum test method for informative gene discovery using two real data sets as follows.

- **The colon cancer data set [2]**. It contains the expression profiles of $2,000$ genes in 22 normal tissues and 40 colon tumor tissues.
- **The leukemia data set [8]**. It consists of $7,129$ genes in 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML) samples.

The experiments are implemented using Matlab 6.5 and SVM-Light [13]. SVM-Light is a free SVM software downloadable at http://svmlight.joachims.org/.

As reported in Table 1, the two data sets do not follow the Gaussian distribution.

4.1 Evaluation of the Rank Sum Test

To evaluate the effectiveness of the rank sum method, we compare the accuracy of the SVM classifiers with and without the informative gene discovery.

For the rank sum method, we try four different significance levels: 0.1, 0.05, 0.01 and 0.001, respectively. The informative gene selection returns 210, 109, 34 and 8 informative genes on the colon data set, and 1837, 1425, 844 and 398 genes on the leukemia data set, respectively.

We also try the following 3 kinds of support vector machines.

- Linear SVM (no kernel);
- 3-poly SVM (cubic polynomial kernel); and
- Radial basis function SVM (RBF kernel).

Before plugging the data into the SVM-Light tool, we normalize the original expression data sets so that the mean is 0 and the standard deviation is 1.

To make the test more robust, we conduct the 4-fold cross-validation experiments [5]. 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

Values	70	83	85	94	97	101	104	107	107	113
Types	t	n	t	t	n	t	n	n	t	n
Ranks	1	2	3	4	5	6	7	8.5	8.5	10
Values	118	119	123	124	129	132	134	146	161	
Types	t	n	n	n	n	t	n	n	n	
Ranks	11	12	13	14	15	16	17	18	19	

Table 5: Assign ranks to each sample.

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 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. We let the SVM-Light toolkit accept all testing patterns. That is, no testing patterns are rejected without labels. The prediction results are shown in Table 6 and Table 7, respectively.

On the colon data set, the best accuracy achieved by using the informative genes is 98.3%, where only one prediction error happens in the 4-fold cross-validation experiment. On the leukemia data set, the accuracy of the informative gene based method is even 100%.

The effectiveness of the informative genes is significant: using the informative genes improves the accuracy dramatically. This is consistent with the results in previous studies.

Moreover, the accuracy is not sensitive to the significance levels. The accuracy will decrease slightly, when either too many or too few informative genes are selected. That is because, on the one hand, too many genes (with a large significance level value) may include some unrelated genes that bring noise to the classifiers; on the other hand, too few genes (with a small significance level value) may filter out some useful information for the classifiers.

To analyze whether the number of genes determined by significance level is proper, we use the Wilcoxon statistic as the ranking criterion, and select the top k genes as common feature ranking methods do. The number k is determined by

$$k_n = 10 \times 2^{n-1},\ n = 1, 2, \ldots$$

That is, we select 10, 20, 40, ..., 1280 genes for the colon data set, and 10, 20, 40, ..., 5120 genes for the leukemia data set, respectively. Then, we construct SVM classifiers to test the accuracy. The results are shown in Figure 3(a) (for colon data set) and Figure 3(b) (for leukemia data set).

From the figures, we can see that the optimal numbers of informative genes are quite different between colon data and leukemia data. On the colon data set, the classifier using the top 40 informative genes gives approximately the best accuracy, while on the leukemia data set, the classifier using the top k genes has a perfect 100% accuracy when k is in the range of 160 to 1280. This experiment clearly shows why it is hard for a user to choose the value for parameter k in informative gene selection: the value of k can be very different in various data sets. Thus, it is hard for a user to choose a good value manually.

In the same figures, we also plot the numbers of informative genes selected by significance levels and the corresponding prediction accuracy using the informative genes. It is interesting to notice that the significance level of 0.01 leads to almost the best performance on both data sets. This strongly suggests that the significance level indicates the quality of the informative genes properly. By rank sum test, we do not have to guess the number of informative genes. Instead, we can use some proper significance level, such as 0.01, to specify the quality requirement of the informative genes. We will come back to this point in Section 5.

4.2 Comparison with Previous Methods

We compare the rank sum test method with some typical existing feature ranking methods. We choose two popular methods in the comparison, the method developed by Golub et al. [9] (called Golub's method) and the t-statistic method developed by Alon et al. [1]. Their ranking criteria are discussed in Section 2. For comparison, we select the same number of informative genes as the rank sum test method does at different significance levels for Golub's method and the t-statistic method, and then construct the SVM classifiers using the selected informative genes. The comparison results of the prediction accuracy (average among the 3 SVM variants) on colon and leukemia data sets are shown in Table 8 and Table 9, respectively.

The results clearly show that the rank sum method is consistently better than the Golub's method and the t-statistic method, while the latter two methods are comparable in accuracy. The improvement in accuracy of the rank sum method against the other two methods is about $2\% - 5\%$. This improvement is non-trivial, considering the baseline methods also achieve accuracy above 90%.

Moreover, we use the significance level to guide the selection of number of informative genes. That is, the Golub's method and the t-statistic method are also benefited from the proper numbers of informative genes in these experiments.

The experimental results are consistent with the Pitman efficiency theory. According to our normality test, both the colon data set and the leukemia data set have many genes violating the normality condition. So the statistics theory guarantees that the rank sum test method will outperform the t-statistic like methods in these two data sets. The experimental results concur with the statistics theory very well. In general, the rank sum test method is better than the t-statistic like methods for informative gene discovery.

	Without informative gene discovery	Significance level (# informative genes)			
		0.1 (210)	0.05 (109)	0.01 (34)	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%
RBM SVM	45.3%	87.2%	93.6%	98.3%	88.8%
Average	44.3%	79.6%	91.4%	96.2%	88.8%

Table 6: The result on the colon data set.

	Without informative gene discovery	Significance level (# informative genes)			
		0.1 (1837)	0.05 (1425)	0.01 (844)	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%
RBM SVM	52.8%	94.4%	100%	100%	100%
Average	60.7%	93.5%	95.4%	100%	98.6%

Table 7: The result on the leukemia data set.

5. DISCUSSION

As shown by the experiments, the rank sum method has a clear advantage. In this section, we present a further discussion on the rationale of informative gene discovery by rank sum test.

To quantify the *quality* of informative genes, it is important to measure the capability of a gene to discriminate phenotypes. Such a measure should be independent to the distribution of the gene expression levels. That is, such a measure should be assumption free for data distribution. Therefore, the Wilcoxon statistic (generally, the rank-sum statistics) and the p-value in rank sum test are ideal candidates for the quality measurement of informative genes.

The ranking criterion is critical for feature ranking methods. A good ranking criterion should provide the following two guarantees. First, it should be able to quantify the significance of the difference between phenotypes. That is, a user can specify the quality requirement of the informative genes by specifying a threshold parameter using the ranking criterion. Second, a ranking criterion should give an order of the genes in quality. It should guarantee that gene g_1 is better than g_2 if g_1 is before g_2 in the order.

Most of the ranking criteria in the previous studies cannot fully provide the above two guarantees. t-statistic can provide the guarantees if the data follows the normal distribution, which may not be true in microarray data.

Rank sum test is independent of data distribution. Thus, the significance level reflects the quality of the informative genes. This is guaranteed by the statistic properties.

6. CONCLUSIONS

Discovery of informative genes is crucial and indispensable for gene expression data analysis. It has extensive applications in bioinformatics. Most of the previous methods take a feature ranking criterion and select the top k genes with the highest ranking values as the informative genes, where k is a user specified parameter. Usually, it is hard for the users to guess the values. The t-statistic method and its variants are the state-of-the-art feature ranking methods. However, this kind of methods inherently assume that the data follows the Gaussian distribution, which is often violated in real microarray data sets according to our investigation.

In this paper, to overcome the defections mentioned above, we present a non-parametric rank sum test method for informative gene discovery. In statistics theory, the Pitman efficiency indicates it is more reasonable and reliable to use a rank sum test method rather than a t-statistic like method. By experiments on real data sets, we also show that the rank sum test approach is more accurate and robust.

Moreover, we propose the method to select informative genes using a significance level threshold. We show that the significance level threshold carries the quality requirement in statistics, is easy to specify, and performs consistent well on real data sets.

As future work, it is interesting to measure the quality of the informative genes more accurately. Moreover, we would like to explore more extensive applications of non-parametric testing methods in other bioinformatics problems, such as analyzing time-series gene expression data, three-dimensional gene expression data sets [12] and the interaction between genes.

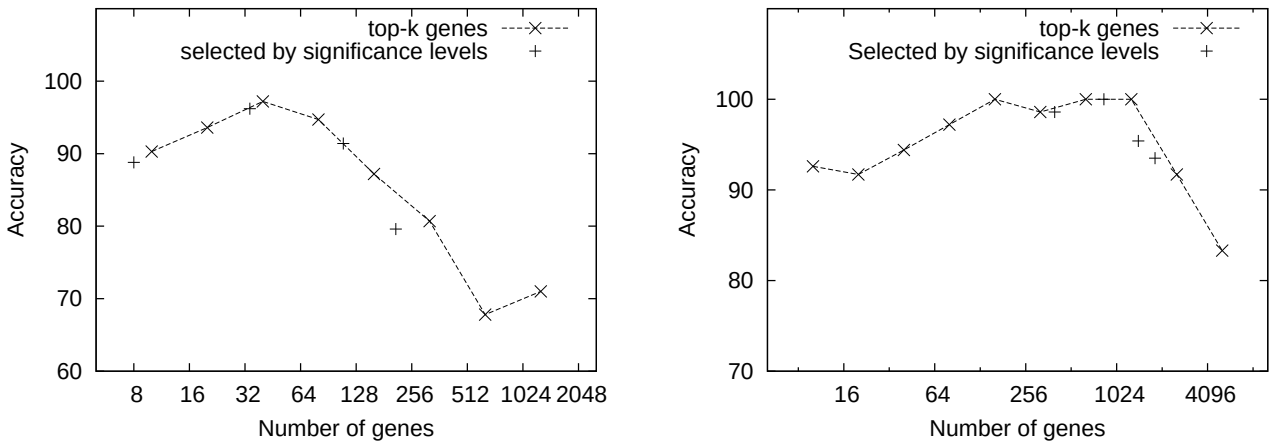
Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments, which help to improve the quality of this paper.

7. REFERENCES

[1] U. Alon and N. Barkai et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96:6745–6750, 1999.

[2] U. Alon and N. Barkai et al. The colon microarray data set. *Availabe at*



(a) The colon data set (b) The Leukemia data set

Figure 3: The test of the top k informative genes versus the classifier accuracy.

Method	Significance level (# informative genes)			
	0.1 (210)	0.05 (109)	0.01 (34)	0.001 (8)
Rank sum	79.6%	91.4%	96.2%	88.8%
Golub's	79.6%	90.3%	93.6%	88.8%
t-statistic	75.9%	90.3%	91.4%	88.8%

Table 8: The comparison of multiple methods on the colon data set.

http://microarray.princeton.edu/oncology/affydata/index.html, 1999.

[3] J. DeRisi and L. Penland et al. Use of a cdna microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14:457–460, 1996.

[4] Chris H. Q. Ding. Analysis of gene expression profiles: class discovery and leaf ordering. In *Proceedings of the sixth annual international conference on Computational biology*, pages 127–136. ACM Press, 2002.

[5] R. Duda, P. Hart, and D. Stork. *Pattern Classification (2nd edition)*. John Wiley & Sons, New York, 2001.

[6] S. Dudoit, J. Fridyand, and T.P. Speed. Comparison of discrimination methods for the classification of tumor using gene expression data. *Journal of American Statistical Association*, 97(457):77–87, 2002.

[7] T. S. Furey, N. Christianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Hauessler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[8] T.R. Golub and D.K. Slonim et al. The leukemia data set. *Availabe at http://www.genome.wi.mit.edu/MPR/data_set_ALL_AML.htm*, 1999.

[9] T.R. Golub and D.K. Slonim et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[10] C.H. Goulden. *Methods of Statistical Analysis (2nd edition)*. John Wiley & Sons, New York, 1956.

[11] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[12] D. Jiang, J. Pei, M. Ramanathan, C. Tang and A. Zhang. *Mining Coherent Gene Clusters from Three-Dimensional Microarray Data*. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*, August 22 - 25, Seattle, WA, USA.

[13] T. Joachims. Making large-scale svm learning practical. In B. Scholkoph et al., editor, *Advances in Kernel Methods – Support Vector Learning*. MIT Press, 1999.

[14] E.L. Lehmann. *Non-parametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco, 1975.

[15] W. Li and I. Grosse et al. Gene selection criterion for discriminant microarray data analysis based on extreme value distributions. In *Proceedings of the seventh annual international conference on Computational molecular biology*, pages 217–223. ACM Press, 2003.

[16] Ya. Nikitin. *Asymptotic efficiency of non-parametric tests*. Cambrige University Press, 1995.

[17] Paul Pavlidis, Jason Weston, Jinsong Cai, and William Noble Grundy. Gene functional classification from heterogeneous data. In *Proceedings of the fifth annual international conference on Computational biology*, pages 249–255. ACM Press, 2001.

Method	Significance level (# informative genes)			
	0.1 (1837)	0.05 (1425)	0.01 (844)	0.001 (398)
Rank sum	93.5%	95.4%	100%	98.6%
Golub's	87%	91.7%	94.4%	94.4%
t-statistic	92.6%	91.7%	95.4%	94.4%

Table 9: The comparison of multiple methods on the leukemia data set.

[18] Donna K. Slonim, Pablo Tamayo, Jill P. Mesirov, Todd R. Golub, and Eric S. Lander. Class prediction and discovery using gene expression data. In *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 263–272. ACM Press, 2000.

[19] Stanford Breast Cancer Microarray Project. The breast cancer microarray data set. *Availabe at http://genome-www.stanford.edu/breast_cancer/sbcmp/data.shtml*, 1999.

[20] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

[21] Hwanjo Yu, Jiong Yang, Wei Wang, and Jiawei Han. Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines. In *Proceedings of the 2nd IEEE computer society bioinformatics conference*, pages 220–228. IEEE Computer Society, 2003.