



A semiparametric approach for marker gene selection based on gene expression data

Zhong Guan¹ and Hongyu Zhao^{2,*}

¹Department of Mathematical Sciences, Indiana University South Bend, South Bend, IN 46634, USA and ²Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA

Received on April 21, 2004; revised on August 5, 2004; accepted on September 9, 2004
Advance Access publication September 16, 2004

ABSTRACT

Motivation: Identification of differentially expressed genes is a major issue in gene expression data analysis and selection of marker genes is critical in tumor classification using gene expression data. In this paper, we propose a semiparametric two-sample test to identify both differentially expressed genes and select marker genes for sample classification.

Results: A simulation study shows that the proposed method is more robust and powerful than the methods, generally used such as *t*-tests and non-parametric rank-sum tests, when the sample size is small. Cross-validation shows that the sample classification based on genes selected using this semiparametric method has lower misclassification rates.

Contact: hongyu.zhao@yale.edu

INTRODUCTION

Identifying differentially expressed genes is one major goal of microarray data analysis. Selecting marker genes for sample classification is also an important issue for disease classification based on gene expression data. Many methods have been proposed to select differentially expressed genes, including the two-sample *t*-tests (Dudoit *et al.*, 2002b), ANOVA (Kerr *et al.*, 2000), SAM (Tusher *et al.*, 2001), Wilcoxon non-parametric two-sample tests and others.

Since microarray technology is still expensive and requires biological materials that may be difficult to collect, most studies perform only a few replicated microarray experiments. However, the appropriateness of many statistical methods, especially parametric methods such as *t*-test and ANOVA, is questionable when the sample size is small. Based on our experience of analyzing both cDNA and Affymetrix microarray data, the difference between expression levels of genes in different samples is reflected both in means and variances, and the normality assumption for the underlying distribution may not hold. In order to take the effect of the treatments on the variances into account and still use two-sample *t*-test-like methods, some authors have proposed

different variance stabilization methods in microarray data analysis (e.g. see, Tusher *et al.*, 2001; Tibshirani, 1988a,b; Huber *et al.*, 2002). Variance shrinkage is another strategy to improve the estimation of variance (Long *et al.*, 2001). O'Brien (1988) considered this issue in the general setting of two-sample comparison by proposing and comparing several extensions of the *t*-, rank-sum and log-rank tests with the corresponding conventional tests. O'Brien (1988) observed that the conventional *t*-, rank-sum and log-rank tests are insensitive for a large class of alternatives that may be expected to occur commonly in practice and pointed out that the proposed methods should be useful for both identifying and interpreting group differences. Mantel and Brown (1974) also studied logistic-regression-based alternative tests for comparing normal distribution parameters. They showed that these tests are valid under various types of non-random sampling schemes and can be used for any distribution within the exponential family. The efficiency comparison between logistic regression and normal discriminant analysis was given by Efron (1975) (see also Halperin *et al.*, 1971).

In gene expression data analysis, especially in selecting differentially expressed genes, which may be used as gene markers to classify human diseases, we hope to find genes that reflect as many different aspects as possible between different samples. To avoid too many parameters and making the calculation too complicated, it may be necessary to consider the differences in both the means and variances into account. Because most gene expression datasets contain a small number of replicates, it is usually difficult to check the normality assumption of the underlying population distributions. One of the robust two-sample tests is the logistic regression method that is called an extension of the classic two-sample *t*-test. Our simulation study showed that, in the classification of tumor samples based on gene expression data, if the sample sizes of the two classes in the learning (training) data are not too small, the logistic regression method performs similarly to the non-parametric Wilcoxon test. Otherwise, the logistic method is more powerful than the non-parametric Wilcoxon test. In all the situations, logistic and Wilcoxon

*To whom correspondence should be addressed.

methods are more powerful than *t*-tests. This advantage of the semiparametric method is especially important in microarray data analysis because the number of replicates is usually small in this context.

METHODOLOGY

Our gene selection procedure is based on the multiple tests performed separately on each gene. For a given gene, let x_{ij} denote the gene expression level of the i -th replicate in the j -th group. For each fixed $j = 0, 1, \dots, m$, the $x_{ij}, i = 1, \dots, n_j$ are an iid sample from distribution F_j . Suppose that

$$f_j(x) = \exp[\alpha_j + \beta_j^\tau T(x)]f_0(x), \quad j = 1, \dots, m, \quad (1)$$

where $\beta_j \in \mathbb{R}^d$ is a d -dimensional parameter ($d \geq 1$) and $T(x)$ is a known vector of functions of x such as $T(x) = x$ or $T(x) = (x, x^2)^\tau$. Qin and Zhang (1997) and Zhang (1999, 2001, 2002a) studied the goodness-of-fit tests for this model. Let Y be a multcategory response variable with $m + 1$ categories, $\pi_j = \Pr(Y = j)$ and F_j be the conditional distribution of X given $Y = j$ for $j = 0, 1, \dots, m$. It is easy to see that model (1) is equivalent to the following polychotomous logistic model (e.g. see Lesaffre and Albert, 1989a; Zhang, 2002a)

$$\log \left[\frac{P(Y = j|X = x)}{P(Y = 0|X = x)} \right] = \alpha_j^* + \beta_j^\tau T(x), \quad j = 1, \dots, m, \quad (2)$$

where $\alpha_j^* = \alpha_j - \log(\pi_0/\pi_j)$. Consider the following hypothesis:

$$H_0: \mathbf{B} = 0 \quad \text{versus} \quad H_1: \mathbf{B} \neq 0,$$

where $\alpha = (\alpha_1, \dots, \alpha_m)^\tau$ and $\mathbf{B} = (\beta_1^\tau, \dots, \beta_m^\tau)^\tau$. Let $n = n_0 + \dots + n_m$, $\alpha_0 = 0$, $\beta_0 = 0$, and denote the combined sample $\{x_{01}, \dots, x_{0n_0}, \dots, x_{m1}, \dots, x_{mn_m}\}$ by $\{u_1, \dots, u_n\}$. Based on the semiparametric model (1), the likelihood of the expression values x_{ij} is

$$L = \left(\prod_{k=0}^m p_k \right) \left[\prod_{j=1}^m \prod_{i=1}^{n_j} \exp[\alpha_j + \beta_j^\tau T(x_{ij})] \right],$$

where $p_k = dF_0(u_k) \geq 0$ and

$$\sum_{k=1}^n p_k \exp[\alpha_j + \beta_j^\tau T(u_k)] = 1, \quad \text{for } j = 0, \dots, m.$$

Using the Lagrangian multiplier method, we get (Zhang, 2002b)

$$\hat{p}_k = \frac{1}{n \sum_{j=0}^m \rho_j \exp[\alpha_j + \beta_j^\tau T(u_k)]}$$

and the profile semiparametric log-likelihood of (α, \mathbf{B})

$$l(\alpha, \mathbf{B}) = - \sum_{k=1}^n \log \left[\sum_{j=0}^m \rho_j \exp[\alpha_j + \beta_j^\tau T(u_k)] \right] + \sum_{j=1}^m \sum_{i=1}^{n_j} [\alpha_j + \beta_j^\tau T(x_{ij})] - n \log n,$$

where $\rho_j = n_j/n$ for $j = 0, \dots, m$. Let $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_m)^\tau$ and $\hat{\mathbf{B}} = (\hat{\beta}_1, \dots, \hat{\beta}_m)$ be the solution to the score equations:

$$\begin{aligned} \frac{\partial l(\alpha, \mathbf{B})}{\partial \alpha_r} &= n_r - \sum_{k=1}^n \frac{\rho_r \exp[\alpha_r + \beta_r^\tau T(u_k)]}{\sum_{j=0}^m \rho_j \exp[\alpha_j + \beta_j^\tau T(u_k)]} = 0, \\ \frac{\partial l(\alpha, \mathbf{B})}{\partial \beta_r} &= \sum_{i=1}^{n_r} T(x_{ir}) - \sum_{k=1}^n \frac{\rho_r T(u_k) \exp[\alpha_r + \beta_r^\tau T(u_k)]}{\sum_{j=0}^m \rho_j \exp[\alpha_j + \beta_j^\tau T(u_k)]} = 0, \quad \text{for } r = 1, \dots, m. \end{aligned}$$

The minus twice the logarithm of likelihood ratio test statistic for $H_0: \mathbf{B} = 0$ versus $H_1: \mathbf{B} \neq 0$ is

$$\begin{aligned} LR &= 2[l(\hat{\alpha}, \hat{\mathbf{B}}) + n \log n] \\ &= 2 \sum_{j=1}^m \sum_{i=1}^{n_j} [\hat{\alpha}_j + \hat{\beta}_j^\tau T(x_{ij})] \\ &\quad + 2 \sum_{k=1}^n \log \left[\sum_{j=0}^m \rho_j \exp[\hat{\alpha}_j + \hat{\beta}_j^\tau T(u_k)] \right], \end{aligned}$$

where $\hat{\alpha}_0 = 0$, $\hat{\beta}_0 = 0$. For large sample sizes, LR has an asymptotic χ^2 distribution with md degrees of freedom.

Let $z_{ij} = j$, for $j = 0, \dots, m; i = 1, \dots, n_j$. By fitting the data $\{(x_{ij}, z_{ij}) : j = 0, \dots, m; i = 1, \dots, n_j\}$ with the polychotomous logistic model (2) and using the Newton–Raphson iteration method, we can also get the maximum-likelihood estimates $\hat{\alpha}$ and $\hat{\mathbf{B}}$. This can be performed by the built-in logistic regression functions of the statistical packages, such as R, S-plus and SAS. The simulation results of this paper ($m = 1$) are calculated using R function `glm` with `family = binomial`. The CATMOD procedure of SAS can be used to perform the analysis of generalized logits for polychotomous outcomes.

When $T(x) = (x, x^2)^\tau$ and $m = 1$, O’Brien (1988) called the above test ‘a natural generalization of the *t*-test’. It is actually a simultaneous test about the population means and variances. In fact, from (1), it is easy to see that the symmetrized Kullback–Leibler information distance between the two distributions F_0 and F_j in the exponential change point model measures the difference between $E_{F_0} T(X)$ and $E_{F_j} T(X)$, i.e.

$$\bar{I}(f_0, f_j) = \frac{1}{2} \beta^\tau [E_{F_0} T(X) - E_{F_j} T(X)].$$

$2\bar{I}(f, g)$ is also called *J*-divergence (see Jeffreys, 1946). Compared to the parametric models, such as Gaussian models, for simultaneous tests of means and the variances, this

semiparametric test is more robust in the sense that it assumes no specific forms of the underlying population distributions and only focuses on the relationship between them. Indeed, the two underlying distributions are assumed to be non-parametric except that the tilt has a parametric exponential form. From (1) and (2), we know that in this logistic model we regress the posterior log odds ratio against x . We also 'regress' the log ratio of the two unknown density (frequency) functions.

The above semiparametric method has been applied to changepoint problem in Guan (2004) and was shown to be more sensitive and robust than some non-parametric methods. Polychotomous discrimination was applied to multiclass cancer classification in Nguyen and Rocke (2002).

An important issue in logistic regression is the existence of the maximum-likelihood estimate of β . As pointed out by Albert and Anderson (1984) and Lesaffre and Albert (1989b), if for some gene, the data x_{ij} are completely or quasicompletely or partially separated, the maximum-likelihood estimate of B in the above logistic regression does not exist. In this case, the gene is clearly a marker gene and may not be detected using this semiparametric method. However, in this case, some other methods, such as t -test and Wilcoxon test, should be able to detect the marker gene. Moreover, Albert and Anderson (1984), Lesaffre and Albert (1989b) and Santner and Duffy (1986) provided methods to determine whether data are separated or overlapped. Therefore, if in the iteration of finding maximum-likelihood estimate of β exceed a given limit, one would like to apply these methods to check the separation status. If the data are separated into the correct groups, then this particular gene is obviously a marker gene. A better strategy to find marker genes is to first apply Wilcoxon test, and if the result of this test is significant, then select the gene, otherwise perform logistic regression test.

APPLICATION TO LEUKEMIA STUDY

The leukemia dataset contains gene expression levels in two types of acute leukemia: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) (Golub *et al.*, 1999). Gene expression levels were measured by using Affymetrix high-density oligonucleotide arrays containing 6817 human genes. The data consist of 47 cases of ALL (38 B-cell ALL and 9 T-cell ALL) and 25 cases of AML, and is available at <http://www.genome.wi.mit.edu/MPR>. This dataset has been analyzed by many authors. For example, Dudoit *et al.* (2002a) used this dataset as an example to compare several discrimination methods for the classification of tumors.

We applied the semiparametric logistic regression test to this dataset and compared the results with the classical two-sample t -test. Among the 40 genes with the smallest P -values for the semiparametric test with $d = 1$, 20 genes are not in the top 40 list of the two-sample t -tests as they have

very large P -values for t -test. These 20 genes are summarized in Table 1 and many of them are in fact cancer-related (<http://www.ncbi.nlm.nih.gov/>).

Based on Golub *et al.* (1999) training set (27 ALL and 11 AML) and the whole data, we compared the sets of top 40 significant genes of the five methods: Wilcoxon test (W), logistic regression with $d = 1$ (Lgt1) and $d = 2$ (Lgt2), BSS/WSS criterion (B/WSS) and t -test (T). The degree of overlap among these methods is summarized in Figures 1 and 2. These results suggest that different methods may lead to quite different sets of marker genes. By increasing the sample sizes, we can get more consistent sets of marker genes. Although the logistic test with $d = 2$ consider the changes in both mean and variance, the sets of marker genes are quite consistent. On the other hand, conventional t -test (assuming equal variances) and Welch adjusted t -test (assuming unequal variances) generated quite different sets of marker genes even for large sample sizes.

It is not surprising to see such differences because different methods work under different model assumptions. Some methods are not very robust and depend heavily upon the goodness of fit of the model to the data. Some are robust and insensitive to the data distribution. The parametric methods, e.g. t -tests and ANOVA, generally assume normal distributions or large sample sizes. In microarray data analyses, we often have small sample sizes and the distribution of most datasets does not follow, or even differs much from the normal distribution. It is also well known that such parametric methods are not robust. Although non-parametric methods, such as the Wilcoxon test, make almost no assumption on the underlying distributions and may lose useful information, even they are distribution-free and robust. The proposed semiparametric method treats the underlying distributions almost non-parametrically and only assumes that the log ratios of density (frequency) functions are a known parametric functions of observations. This procedure can be viewed as to regress the log ratio of density (frequency) function. Therefore, as a local fit of the data, this semiparametric approach provides a flexible, robust and powerful alternative to the existing methods.

CLASSIFICATION

As mentioned in Dudoit *et al.* (2002a), the identification of 'marker' genes for the classification of tumors is an important issue and t -tests are generally unable to identify genes that discriminate between all the classes. Using logistic regression with $d = 2$, we take the change in the variance of training data into account. We do not mean, and it is also impossible, to use the instability of the expression level of the selected gene to discriminate different classes. However, this feature is actually present in gene expression data. From this point of view, we see that bagging (Breiman, 1998, 1996) or boosting (Freund and Schapire, 1997) would be necessary even for

Table 1. Description of the 20 genes

GenBank Id	Gene Description (NCBI annotation)
M31166	PTX3 pentaxin-related gene, rapidly induced by IL-1 β [Human tumor necrosis factor-inducible (TSG-14) mRNA, complete cds]
M27783	ELA2 elastase 2, neutrophil (Human neutrophil elastase mRNA, 3' end)
M91432	ACADM acyl-coenzyme A dehydrogenase, C-4 to C-12 straight chain [Human medium-chain acyl-CoA dehydrogenase (MCAD) gene, exon 12]
D88422	Cystatin A (<i>Homo sapiens</i> gene for cystatin A, exon 3 and complete cds)
M92287	CCND3 cyclin D3 [<i>H.sapiens</i> cyclin D3 (CCND3) mRNA, complete cds]
M31523	TCF3 transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47) [Human transcription factor (E2A) mRNA, complete cds]
X74262	Retinoblastoma binding protein P48 (<i>H.sapiens</i> RbAp48 mRNA encoding retinoblastoma binding protein)
U85767	Myeloid progenitor inhibitory factor-1 MPIF-1 mRNA (Human myeloid progenitor inhibitory factor-1 MPIF-1 mRNA, complete cds)
U46499	Glutathione S-transferase, microsomal [<i>H.sapiens</i> microsomal glutathione transferase (<i>MGST1</i>) gene, 3' sequence]
X74801	T-complex protein 1, gamma subunit (<i>H.sapiens</i> Cctg mRNA for chaperonin)
D26308	NADPH-flavin reductase (Human mRNA for NADPH-flavin reductase, complete cds)
U32944	Cytoplasmic dynein light chain 1 (hdlc1) mRNA (Human cytoplasmic dynein light chain 1 (hdlc1) mRNA, complete cds)
X62320	GRN granulin (<i>H.sapiens</i> mRNA for epithelin 1 and 2)
L42572	Motor protein [<i>H.sapiens</i> transmembrane protein (p87/89) mRNA, complete cds]
L47738	Inducible protein mRNA (<i>H.sapiens</i> inducible protein mRNA, complete cds)
X15949	IRF2 interferon regulatory factor 2 [Human mRNA for interferon regulatory factor-2 (IRF-2)]
J05243	SPTAN1 spectrin, alpha, non-erythrocytic 1 (alpha-fodrin) [Human non-erythroid alpha-spectrin (SPTAN1) mRNA, complete cds]
U62136	Putative enterocyte differentiation promoting factor mRNA, partial cds (<i>H.sapiens</i> enterocyte differentiation associated factor EDAF-1 mRNA, complete cds)
J04990	Cathepsin G precursor (Human cathepsin G gene, complete cds)
M12959	TCRA T-cell receptor alpha-chain (Human T-cell receptor active alpha-chain mRNA from JM cell line, complete cds. Human T-cell leukemic cell line JM, cDNA to mRNA, clone pJM3E11.)

some 'stable' classifiers such as the nearest neighbors (Fix and Hodges, 1951) procedure.

After the selection of the genes, we can use quadratic discriminant analysis to classify a new sample. In the following analysis, we only consider the simplest case $K = 2$. Let $x_i = (x_{i1}, \dots, x_{ip})^T$ denote the expression profile of p selected genes of the i -th sample in the training dataset, and y_i is the class label of this sample.

Several discrimination methods can be used to classify tumors based on gene expression data. Dudoit *et al.* (2002a) compared many methods and concluded that the k -nearest neighbors (k NNs) classifier perform remarkably

well compared to more sophisticated methods such as aggregated classification trees (CART). In our analysis, we aggregate the nearest neighbors classifier using bagging and boosting perturbations. A total of 30 genes are selected using the 30 most significant genes based on five tests: semiparametric tests (logistic regression) with $d = 1, 2$, Wilcoxon test, t -tests with unequal variances and BSS/WSS criterion (Dudoit *et al.*, 2002a), which is equivalent to the t -test with equal variance. The value of k is chosen to be 3. Using Breiman's (1998) adapted boosting algorithm (see Freund and Schapire, 1998), we correctly classified all the 34 observations in the test leukemia dataset with two classes. The prediction votes (PVs) are given

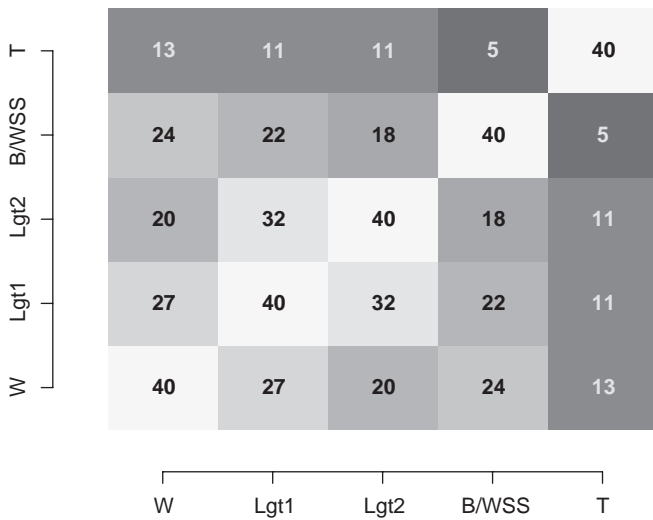


Fig. 1. Overlap of five methods with genes selected based on Golub *et al.*'s learning set.

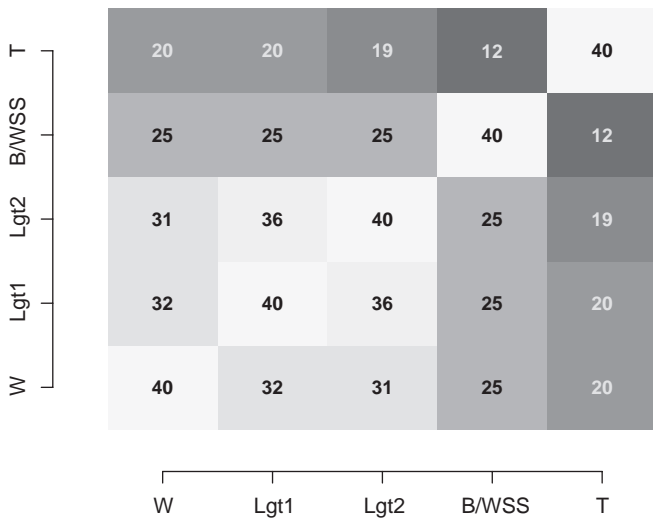


Fig. 2. Overlap of five methods with genes selected based on all data.

in Table 2. Two ALL cases [indices 71 (B-cell) and 67 (T-cell)] and one AML case (index 66) are the most difficult observations to classify. Observations 66 and 67 were in the list of three observations tended to be difficult with the classification of Dudoit *et al.* (2002a) and were misclassified and have low prediction strength of 0.27 and 0.15, respectively, when compared with Golub *et al.* (1999). Using standard bagging (non-parametric bootstrap) procedure, only 2 (67 and 71) out of 34 test observations are misclassified. To compare different marker gene selection methods in terms of the misclassification rates, we use learning set/test set (LS/TS) resampling procedure of 2:1, 1:2, 1:3 and 1:6 schemes (see Dudoit *et al.*, 2002a). That is, with $(n_L, n_T) = (48, 24), (24, 48), (18, 54)$ and $(10, 62)$, the misclassification rates are estimated based

Table 2. Prediction votes of aggregated *k*NN classifier using boosting

Index	39	40	42	47	48	49	41
PV	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Index	43	44	45	46	70	71	72
PV	1.00	1.00	1.00	1.00	1.00	0.59	1.00
Index	68	69	67	55	56	59	52
PV	1.00	1.00	0.63	1.00	1.00	1.00	1.00
Index	53	51	50	54	57	58	60
PV	0.97	1.00	1.00	1.00	0.97	0.96	0.95
Index	61	65	66	63	64	62	
PV	1.00	0.98	0.79	1.00	1.00	1.00	

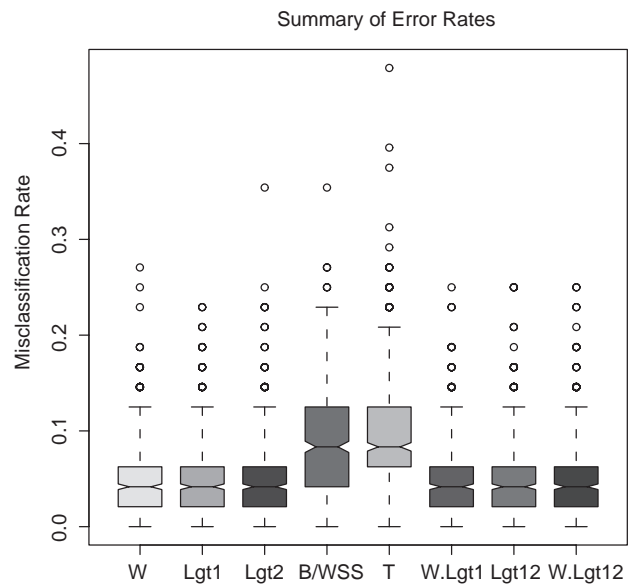


Fig. 3. Summary of error rates (500 replicates, 1:2 scheme, 40 different genes).

on random partitions of the combined dataset of $n = 72$ observations into a learning set of n_L observations and a test set of n_T observations, respectively.

Error rate estimation using simulation

We first select the top $p = 30$, and 40 significant genes based on eight tests: Wilcoxon (W), logistic regression with $d = 1$ (Lgt1), $d = 2$ (Lgt2), t -test (T), BSS/WSS criterion (B/WSS), the combination of Wilcoxon and logistic regression with $d = 1$ (W.Lgt1), the combination of logistic regression with $d = 1$ and $d = 2$ (Lgt12), and the combination of Wilcoxon and logistic regression with $d = 1$ and $d = 2$ (W.Lgt12). For the combined method, the P -value of a gene is determined by the smaller or smallest one of the tests to be combined. We then simulate 500 LS/TS (2:1, 1:2, 1:3 and 1:6 schemes) samples and use the *k*NN classifier with $k = 3$ to classify the test set, the misclassification rates are summarized by box-and-whisker plots in Figures 3–6.

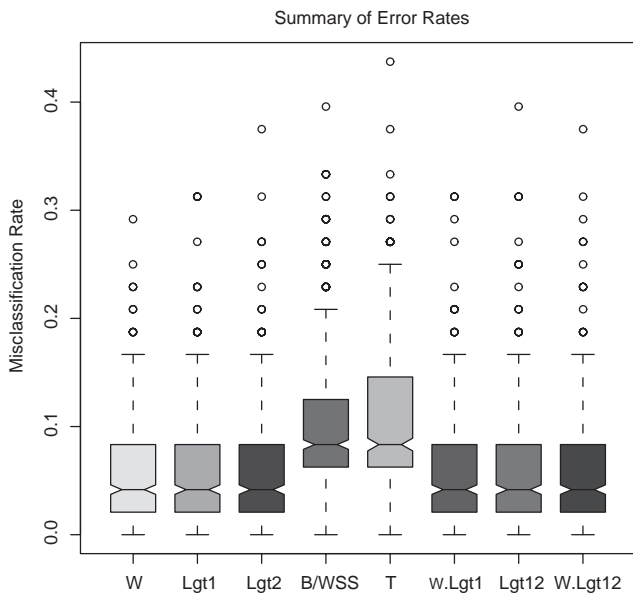


Fig. 4. Summary of error rates (500 replicates, 1:2 scheme, 30 different genes).

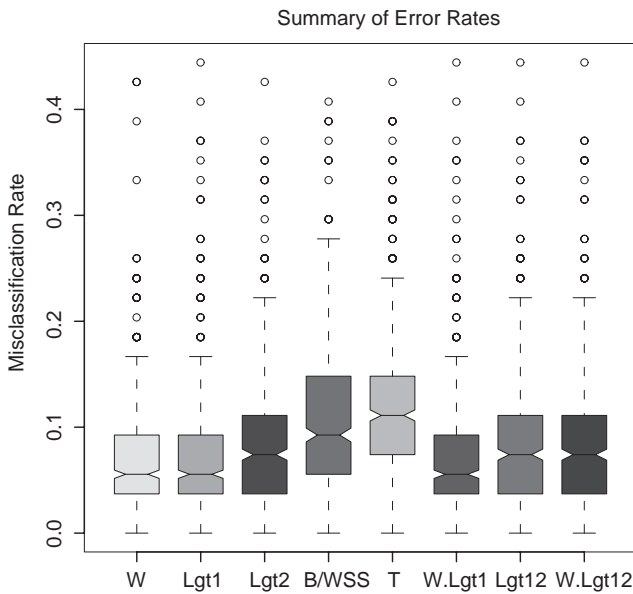


Fig. 5. Summary of error rates (500 replicates, 1:3 scheme, 40 different genes).

This simulation study shows that different marker gene selection methods affect the output of the classification dramatically.

Since we mainly focus on the comparison of different methods of marker gene selection, we can compare these methods by selecting genes based on all datasets available and then compare the classification results using different sets of marker genes. We note that the misclassification rates

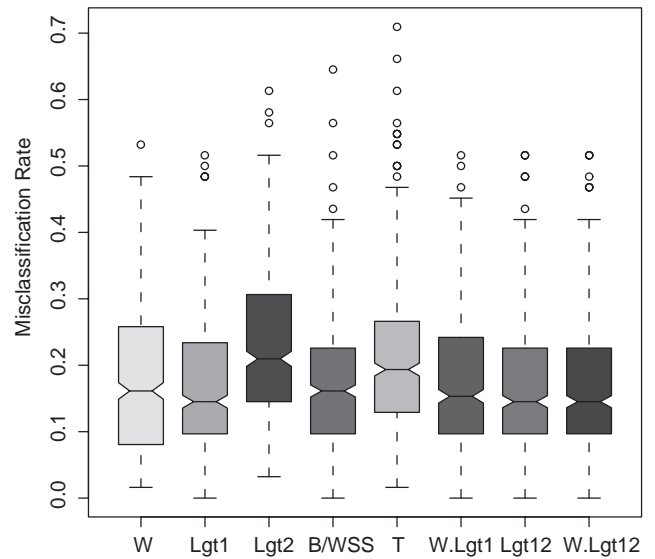


Fig. 6. Summary of error rates (500 replicates, 1:6 scheme, 40 different genes).

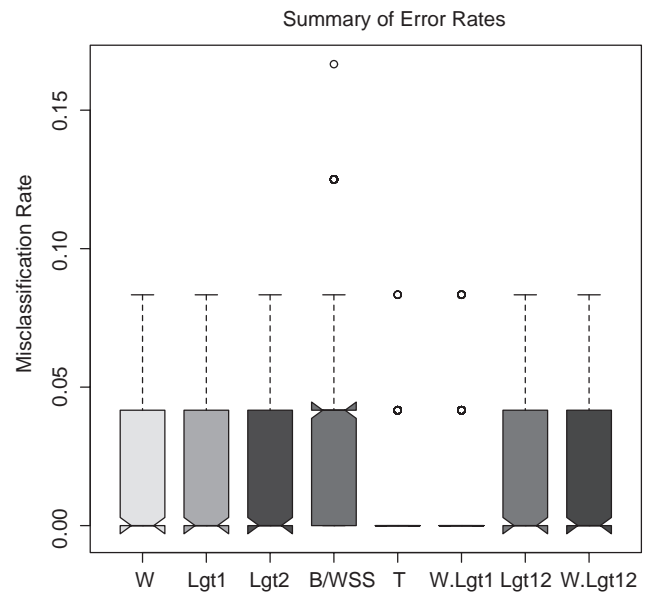


Fig. 7. Summary of error rates (500 replicates, 2:1 scheme, 40 same genes).

may be underestimated, although the set of marker genes may converge for each gene selection method when the learning dataset is large enough. Figure 7 compares the results based on the same set of marker genes that are selected using all the datasets available.

DISCUSSION

Owing to the expense of microarray experiments and difficulty in biological sample collection, the sample sizes of

most microarray data are quite small. In this case, parametric statistical methods such as t -test and ANOVA are less reliable than the non-parametric or semiparametric methods due to likely deviations from the parametric assumptions. For example, although both t -test and logistics regression test with $d = 1$ compare the means, t -test depends heavily on the correctness of the normality assumption of the data values. This is especially the case for microarray data since there are many unknown uncontrollable factors that affect the observed expression levels and the normal assumptions are also difficult to test due to small sample sizes. Although Wilcoxon non-parametric two-sample test (equivalent to Mann–Whitney test) and its k -sample version Kruskal–Wallis test are distribution-free methods, they make no use of any information about the distributions of the data. In contrast, the proposed semiparametric method makes no assumption on the underlying distribution except that there is a parametric link between two groups and this link can be interpreted as regression. Therefore, this semiparametric method is both more powerful and robust than the other methods. In selecting differentially expressed genes based on gene expression data, we recommend the combination of several methods, e.g. t -test or ANOVA, logistic regression with $d = 1, 2$ and non-parametric methods, and consider the union of the sets of significant genes as the set of candidate genes. In selecting marker genes for classification, the logistic regression method with $d = 1$ seems appropriate if linear classifiers are applied.

Selection method for the classification of tumors and cancers based on high-throughput data is also important. Different methods indeed yield quite different misclassification rates for various datasets. In the examples of the present paper, we used only k NN classification just because it has been shown by Dudoit *et al.* (2002a) that this method is better than other commonly used classification methods. For other kinds of datasets, other methods may be used. For example, Wu *et al.* (2003) compared several discriminant methods including linear and quadratic discriminant analysis, k NN and random forest (RF) for the classification of ovarian cancer using mass spectrometry data, and showed that RF performs better than all the other methods considered in the comparison.

ACKNOWLEDGEMENTS

We thank Dr Biao Zhang for introducing the paper of O'Brien (1988) and his advice on empirical likelihood techniques. This work was supported in part by NIH grant GM59507 and NSF grant DMS 0241160.

REFERENCES

- Albert, A. and Anderson, J.A. (1984) On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71**, 1–10.
- Breiman, L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.
- Breiman, L. (1998) Arcing classifier. *Ann. Stat.*, **26**, 801–824.
- Dudoit, S., Fridly, J. and Speed, T.P. (2002a) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.
- Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002b) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sinica*, **12**, 111–139.
- Efron, B. (1975) The efficiency of logistic regression compared to normal discriminant analysis. *J. Am. Stat. Assoc.*, **70**, 892–898.
- Fix, E. and Hodges, J. (1951) Discriminatory analysis, nonparametric discrimination: consistency properties. Technical Report. USAF School of Aviation Medicine, Randolph Field, TX.
- Freund, Y. and Schapire, R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Sys. Sci.*, **55**, 119–139.
- Freund, Y. and Schapire, R.E. (1998) Comment on ‘Arcing classifiers’. *Ann. Stat.*, **26**, 824–832.
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasen, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, 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.
- Guan, Z. (2004) A semiparametric changepoint model. *Biometrika*, **91**, 849–862.
- Halperin, M., Blackwelder, W.C. and Verter, J.I. (1971) Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches. *J. Chronic Dis.*, **24**, 125–158.
- Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96–S104.
- Jeffreys, H. (1946) An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond., Ser. A*, **186**, 453–461.
- Kerr, M.K., Martin, M. and Churchill, G.A. (2000) Analysis of variance for gene expression microarray data. *Technical Report*. The Jackson Laboratory, Bar Harbor, ME.
- Lesaffre, E. and Albert, A. (1989a) Multiple-group logistic regression diagnostics. *J. R. Stat. Soc. Ser. C*, **38**, 425–440.
- Lesaffre, E. and Albert, A. (1989b) Partial separation in logistic discrimination. *J. R. Statist. Soc. Ser. B*, **51**, 109–116.
- Long, A., Mangalam, H.J., Chan, B.Y., Toller, L., Hatfield, G.W. and Baldi, P. (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression profiling in *Escherichia coli* K12. *J. Biol. Chem.*, **276**, 19937–19944.
- Mantel, N. and Brown, C. (1974) Alternative tests for comparing normal distribution parameters based on logistic regression. *Biometrics*, **30**, 485–497.
- Nguyen, D.V. and Rocke, D.M. (2002) Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, **18**, 1216–1226.
- O'Brien, P.C. (1988) Comparing two samples: extensions of the t , rank-sum, and log-rank tests. *J. Am. Stat. Assoc.*, **83**, 52–61.
- Qin, J. and Zhang, B. (1997) A goodness-of-fit test for logistic regression models based on case–control data. *Biometrika*, **84**, 609–618.

- Santner,T.J. and Duffy,D.E. (1986) A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **73**, 755–758.
- Tibshirani,R. (1988a) Estimating transformations for regression via additivity and variance stabilization. *J. Am. Stat. Assoc.*, **83**, 394–405.
- Tibshirani,R. (1988b) Variance stabilization and the bootstrap. *Biometrika*, **75**, 433–444.
- Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, **98**, 5116–5121.
- Wu,B., Abbott,T., Fishman,D. McMurray,W., Mor,G., Stone,K., Ward,D., Williams,K. and Zhao,H. (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, **19**, 1636–1643.
- Zhang,B. (1999) A chi-squared goodness-of-fit test for logistic regression models based on case–control data. *Biometrika*, **86**, 531–539.
- Zhang,B. (2001) An information matrix test for logistic regression models based on case–control data. *Biometrika*, **88**, 921–932.
- Zhang,B. (2002a) Assessing goodness-of-fit of generalized logit models based on case–control data. *J. Multivariate Anal.*, **82**, 17–38.
- Zhang,B. (2002b) An EM algorithm for a semiparametric finite mixture model. *J. Stat. Comput. Simul.*, **72**, 791–802.