Cancer Informatics

METHODOLOGY

Comparison of Two Output-Coding Strategies for Multi-Class Tumor Classification Using Gene Expression Data and Latent Variable Model as Binary Classifier

Sandeep J. Joseph[1], Kelly R. Robbins[1], Wensheng Zhang[1] and Romdhane Rekaya[1,2,3]

[1]Rhodes Centre for Animal and Dairy Science, [2]Institute of Bioinformatics, [3]Department of Statistics, University of Georgia, Athens, GA-30605, USA.

Abstract: Multi-class cancer classification based on microarray data is described. A generalized output-coding scheme based on One Versus One (OVO) combined with Latent Variable Model (LVM) is used. Results from the proposed One Versus One (OVO) output-coding strategy is compared with the results obtained from the generalized One Versus All (OVA) method and their efficiencies of using them for multi-class tumor classification have been studied. This comparative study was done using two microarray gene expression data: Global Cancer Map (GCM) dataset and brain cancer (BC) dataset. Primary feature selection was based on fold change and penalized t-statistics. Evaluation was conducted with varying feature numbers. The OVO coding strategy worked quite well with the BC data, while both OVO and OVA results seemed to be similar for the GCM data. The selection of output coding methods for combining binary classifiers for multi-class tumor classification depends on the number of tumor types considered, the discrepancies between the tumor samples used for training as well as the heterogeneity of expression within the cancer subtypes used as training data.

Keywords: binary classifier, multi-class tumor classification, supervised classification, latent variable model, gibbs sampling, gene expression

Cancer Informatics 2010:9 39–48

This article is available from http://www.la-press.com.

Introduction

Improvements in cancer classification have been of great importance in cancer treatment. Conventional diagnostic methods are based on subjective evaluation of the morphological appearance of the tissue sample, which requires a visible phenotype and a trained pathologist to interpret the view.[1] It is difficult for those approaches to distinguish tumours with similar histo-pathological appearance (phenotype) but different clinical course and response to therapy. Since the advent of microarray technology, researchers have begun to use expression array analysis as a quantitative phenotyping tool. The potential advantage of using arrays for phenotyping is that they provide a simultaneous quantitative measure of thousands of parameters (gene expression levels) some of which are likely to have disease relevance. Due to the ability to quantify a large number of parameters, the use of expression array in diagnosing, promises both more accurate class prediction and the identification of subclasses that could not be defined by traditional methods. Even though this technology is promising in disease diagnosis, there are many huddles that the researcher should over come in order to achieve such goals. Since extraction of mRNA from a single cell is extremely difficult task, researchers are forced to pool tissues that seem to share the same fate or the same functions to obtain the adequate quantity of mRNA. This may imply that the expression levels calculated are the means of all the cells in the pool. Another issue is with the genetic variability of two individuals that will affect the expression of genes. The accumulation of noise to affect the outcome, at various points of experiment is yet another issue. Finally, the samples collected are small in numbers that results in high dimensional data with very large (thousands to tens of thousands) number of genes. The most fundamental problem that needs to solve the above mentioned caveats of the technology is to identify genes whose expression patterns either characterize a particular cell state or predict a certain forthcoming cell state. The first step in solving this problem is the development of tools for classifying samples according to their gene expression. Various clustering, classification and predicting techniques have been used to analyze and understand the gene expression data resulted from DNA microarray. Some recent applications include: molecular classification of acute leukaemia,[1] classification of human cancer cell lines,[2] Support Vector Machine (SVM) classification of cancer samples.[3]

A challenge in predicting diagnostic categories using microarray data is that the number of genes is usually significantly greater than the number of tissue samples available, and only a subset of the genes is relevant in distinguishing different classes. Selection of relevant genes for classification is known as feature selection. This has three main applications: first, the classification accuracy is often improved using a subset instead of the entire set of genes: second, a small set of relevant genes is convenient for developing diagnostic tests; and third, these genes may lead to biologically interesting insights that are characteristics of the classes of interest. There have been many reports that address the classification and feature-selection problems.[4] However, many of these methods are tailored towards binary classification in which there are only two classes. While majority of the cancer phenotypes have more than two subtypes, which leads us to a multi-class prediction scenario.

Multiple class prediction is more difficult than binary prediction because the classification algorithm has to consider a greater number of separation boundaries or relations.[5,6] In binary classification, an algorithm can make a decision boundary for only one of the classes; the other class is simply the complement. In multi-class classification problems, each class has to be defined explicitly. A multi-class problem can be decomposed into a set of binary problems and then combined to make a final multi-class prediction. The general term used for such procedures are called Ensemble methods.

The basic idea behind combining binary classifiers is to decompose the multiclass problem into a set of easier and more accessible binary problems. The main advantage in this divide-and-conquer strategy is that any binary classification algorithm can be used. Besides choosing a decomposition scheme and a classifier for the binary decompositions, one also needs to devise a strategy for combining the binary classifiers and providing a final prediction. The problems of combining binary classifiers have been studied in the computer science literature[7,8] from a theoretical and empirical perspective. However, the literature is inconclusive, and the best method for combining binary classifiers for any particular problem is open.

Standard modern approaches for combining binary classifiers can be stated in terms of what is called output coding.[9] The basic idea behind output coding can be illustrated by the following example: Given three classes, the first classifier may be trained to discriminate classes 1 and 2 from 3, the second classifier is trained to discriminate classes 2 and 3 from 1, and the third classifier is trained to discriminate classes 1 and 3 from 2. Two common examples of output coding are the one-versus-all (OVA) and one-versus-one (OVO), also called all-pairs (AP) approaches. OVA and OVO combinations of binary classifiers have been applied to the analysis of DNA microarray data.[10]

In this paper, a generalized OVO combination of binary classifiers has been proposed and applied to multiclass tumour classification. Latent Variable Model (LVM) was chosen as the binary classifier, which has been successfully applied to microarray classification. Results from the proposed OVO method was compared to that obtained using OVA strategy by applying them to two publicly available microarray gene expression datasets Two major categories of feature selection methods have been tested: fold change as well as penalized t-test.

Materials and Methods

Datasets

Two data sets have been used for this comparative study. The first dataset on which the proposed method applied was to the well-known GCM dataset.[4] It consisted of 144 and 54 training and test samples, respectively, representing 14 tumor types. These tumor types included BR (breast adenocarcinoma), PR (prostate adenocarcinoma), LU (lung adenocarcinoma), CO (colorectal adenocarcinoma), LY (lymphoma), BL (bladder transitional cell carcinoma), ML (melanoma), UT (uterine adenocarcinoma), LE (leukemia), RE (renal cell carcinoma), PA (pancreatic adenocarcinoma), OV (ovarian adenocarcinoma), ME (pleural mesothelioma) and CNS (central nervous system). All samples were primary tumors with the exception of eight metastatic tumors in the test set. Expression data was generated using Affymetrix high-density oligonucleotide microarrays containing 16,043 known human genes or expressed sequence tags (EST). The distribution of training and testing samples among the 14 classes is listed in Table 1. The expression intensities for each gene were calculated using Affymetrix GENECHIP analysis software.

The second dataset (BC)[11] used in this study contained 92 brain cancer expression profiles consisting of 7129 genes using an Affymetrix oligonucleotide array. These samples are grouped into 6 classes: 46 samples of classic medulloblastoma (CMD); 14 of desmoplastic medulloblastoma (DMD); 10 of malignant gliomas (MG); 10 of atypical teratoid/rhabdoid tumors (AR); 4 of normal cerebellum (NC) and 8 of supratentorial primitive neuroectodermal tumors (PN). Ideally we should include all cancer and non-cancer subtypes in the dataset for classification. After performing some preliminary studies, we found that the expression level of the NC group varies significantly from the other cancer related sub-groups. With such large variability and the extreme small size (4 samples) we decided to remove the NC group. Additionally, due the large number of samples in the CMD group (half of all samples) and our concern that it will dominate the classification progress, we decided to remove this group and to focus only on 4 groups with relative equal number of samples. Consequently we used 4 classes of cancer were two of them, when classified solely by morphological characteristics, were in controversy of whether they belong to the same group or not. The distribution of training and testing samples among the 4 subtypes of the BC dataset is listed in Table 2.

Data preprocessing

Data preprocessing of both the datasets consisted of threshold treatment to the expression intensities with

Table 1. GCM dataset: Number of samples per tumor class.[a]

Cancer class	BR	PR	LU	CO	LY	BL	ME	UT	LE	RE	PA	OV	ML	CNS
Training	8	8	8	8	16	8	8	8	24	8	8	8	8	16
Testing	4	6	4	4	6	3	2	2	6	3	3	4	3	4

Abbreviations: [a]BR, Breast; PR, Prostate; LU, Lung; CO, Colorectal; LY, Lymphoma; BL, Bladder; ME, Melanoma; UT, Uterus; LE, Leukemia; RE, Renal; PA, Pancreas; OV, Ovary; ML, Mesothelioma; CNS, Brain.

Table 2. BC datasets: Number of samples per subtype.[b]

Subtype	DMD	MG	AR	PN
Training	14	10	10	8
Testing	4	3	3	2

Abbreviations: [b]DMD, Desmoplastic medulloblastoma; MG, Malignant gliomas; AR, Atypical rhabdoid; PN, Primitive neuroectodermal.

20 and 1600 as the lower and upper limit, respectively, after which the $\log_2$ transformation was applied. Further, genes with the highest transformed intensity being smaller than two times the minimum expression were deleted.

Feature selection (Initial gene pool)

The initial gene pool was build to reduce number of genes or features in a data. Many researches have revealed that when the number of features is large, the performance of the learning methods degrades. Ideally, one would like to select an optimal subset of features that would yield maximum predictive power for a given classification algorithm. In the case of high-dimensional data sets, this can be very computationally demanding; consequently, many statistical and rank based methods are used. In this study, two criteria were used for feature selection. Each of the two criteria was computed for every gene separately.

Fold Change (FC)

At the $\log_2$ scale, the fold change is the absolute value of the difference of expression intensity means between two groups. For OVA partitioning of data, this can be expressed as FC = $|M_o - M_r|$, where M_o represents the mean of the training samples in a single tumor type to be separated from the others, and M_r represents the mean of the training samples in all other cancer types. For OVO partitioning, M_1 and M_2 represents the mean of the training samples of the two cancer types under consideration. Intrinsically. FC assigns equal variance to every gene.

Penalized t-statistics

T-statistics is defined as:

$$t = \frac{M_o - M_r}{s_p \sqrt{1/N_o + 1/N_r}}$$

where S_p is the pooled standard deviation, N_o and N_r are the numbers of the training samples in the two groups, respectively, and M_o and M_r for OVA and M_1 and M_2 for OVO are the same as defined above.

For genes with very small S_p, their t-statistics could be large even when the fold change is quite small. The penalized statistics helps to overcome this shortcoming. It consists of adding a positive quantity, a (90th percentile of the distribution of the pooled standard deviation of all the genes), to the denominator of the t-statistics leading to:

$$t = \frac{M_o - M_r}{a + s_p \sqrt{1/N_o + 1/N_r}}$$

In this study, the size of the feature (genes) set used was 50, 75, 100 and 200 genes for the BC dataset and 50, 100 and 200 genes for the GCM dataset.

Binary classification method

Latent variable models are often used in binary classification. It consists of establishing a relationship between the observed binary response, $Y = (Y_1, Y_2, \ldots, Y_n)$, and a continuous and unobserved latent variable $l = (l_1, l_2, \ldots, l_n)$ such that:

$$Y_i = \begin{cases} 1 & \text{if} \quad l_i \geq 0 \\ 0 & \text{if} \quad l_i < 0 \end{cases}$$

Further, if a set of exploratory factors, X_i for ith sample are collected, the liability l_i could be modeled through a simple linear regression model as:

$$l_i = X'_i\beta + e_i \quad E(l_i) = X'_i\beta \quad e_i \sim N(0, 1) \qquad (1)$$

The link function of the systematic component $X'_i\beta$ with the binary response Y_i could be structured via a probit model.[12,13]

Thus,

$$p_i(Y_i = 1) = \phi(X'_i\beta) \quad \text{and} \quad p_i(Y_i = 0) = 1 - \phi(X'_i\beta) \qquad (2)$$

where $\phi(\cdot)$ is the standard normal cumulative distribution function. In the case of using expression profiling for disease classification, the matrix X will include the expression intensities and β will include the vectors of gene effects. For the classification of binary

response, the probability p_i can be used, leading to the following discrimination rule:

$$Yi = \begin{cases} 1 & if \ \phi(X'_i\beta) \geq 0.5 \\ 0 & if \ \phi(X'_i\beta) < 0.5 \end{cases} \quad (3)$$

Inferences on the model parameters β, and the point and interval estimates of the quantity of main interest, $p_i(Y_i = 1)$ was achieved by Bayesian approach implemented via a Markov Chain Monte Carlo (MCMC) algorithm (Gibbs sampling).[13] All Conditional distributions required for the implementation of the model via Gibbs sampler were in closed form being normal for the position parameters, scaled inverted Chi square for the residual variance, and truncated normal distribution for the latent variables and were sampled following the relationship:

$$l_i = X'_i\beta + \phi^{-1}\,[y_i\phi(-u_i) + u_i(y_i + (1 - 2y_i)\ \phi(-\mu_i))] \quad (4)$$

Convergence of the sampling process was accessed based on visual inspection of the samples trace plots, and was always reached in less than 200 rounds. Conservatively, a long chain strategy of 20,000 iterations was implemented with the first 5,000 iterations discarded as burn-in. In order to alleviate autocorrelations, one in every 50 samples of the remaining 15,000 draws was maintained for post Gibbs analysis.

Due to the fact that the number of genes is much greater than the number of samples; a dimension reduction technique called singular value decomposition (SVD) is performed before fitting the regression model. For the BC dataset, five replications were done by randomly selecting the training samples and testing samples from each subtype by maintaining the same number of samples in both training and testing samples as mentioned in Table 1. For the GCM dataset, exact split done by Ramaswamy et al[4] was done.

Output coding strategies implemented for multiple classifications

Once the discriminative genes (features) are selected based on fold change and penalized t-test, samples that contain only those selected (top ranked) genes were used for training as well as for testing. The two output coding strategies implemented in this study using LVM (binary classifier) are described below.

One-versus-all method

The one-versus-all approach[4] divides the classes into two groups each time, with one group consisting of a single class and the other group consisting of samples in all the other classes. In other words, given k classes, k independent classifiers are constructed were the ith classifier is trained to separate samples belonging to class i from all others. The codebook is a diagonal matrix, and the final prediction is based on the classifier that produces the strongest confidence:

$$\text{Class} = \arg\max f_i$$

where f_i is the signed confidence measure of the ith classifier. The maximum confidence rule with $p_i(Y_i = 1)$ is used as the confidence measure.

For example, in this study, if we take DMD as one group, the other group will contain the remaining samples from MG, AR and PN. The DMD samples will be given the binary status 1 and all other tumor samples except DMD will have 0 as binary status. The predictive ability of LVM was tested using a cross-validation procedure using the test samples (Tables 1 and 2), in which the binary response for one test sample was treated as unknown and the regression coefficients were calculated using the training samples of the 2 tumor types. The estimated coefficients were then used to predict the status of the unknown test sample. This process is repeated for each test sample of all tumor subtypes.

One-versus-One method

This approach involves implementing LVM for each pair of tumor classes. This ensures that each of these groups is compared with each of the remaining groups one by one, and the corresponding selected genes can represent the most significant differences. The usual way to perform this is, given k classes, $(k(k-1))/2$ classifiers are constructed, with each classifier trained to discriminate between a class pair say, i and j. But in this study, two sets of $(k(k-1))/2$ classifiers are generated, were one set is the complement of the other as described in Box 1. This can be thought of as a $(k-1)$ x k matrix, where the ijth entry corresponds to a classifier that discriminates between classes i and j. The codebook, in this study was done

Box 1. Proposed One Versus One (OVO) output-coding strategy.

Let A, B, C & D represents four tumor types

Step 1

The predictive ability of LVM is tested using cross validation procedure, in which the binary response for one test sample is treated as unknown and regression coefficients are calculated using the training dataset (only 2 classes at a time).

Step 2

For a particular test sample (for ex. A), cross validation procedure is implemented for each pair wise combination to estimate $P(Y_i = 1)$, resulting in the following $(k–1) \times k$ matrix for each test sample.

$P(Y_i = A)$ A (1)* vs. B (0)	$P(Y_i = B)$ B (1) vs. A (0)	$P(Y_i = C)$ C (1) vs. A (0)	$P(Y_i = D)$ D (1) vs. A (0)
$P(Y_i = A)$ A (1) vs. C (0)	$P(Y_i = B)$ B (1) vs. C (0)	$P(Y_i = C)$ C (1) vs. B (0)	$P(Y_i = D)$ D (1) vs. B (0)
$P(Y_i = A)$ A (1) vs. D (0)	$P(Y_i = B)$ B (1) vs. D (0)	$P(Y_i = C)$ C (1) vs. D (0)	$P(Y_i = D)$ D (1) vs. C (0)

*Represents the binary status of the training data while implementing LVM.

Step 3

Mean and median of each column (for ex. column 1 contains the probability that the test sample is predicted to be A compared to all other tumor classes (pair wise comparison)) is estimated and the test sample will be assigned to that particular tumor class whose estimated mean or median is highest. For example, if column 1 has the highest mean compared to other 3 columns then the test sample will be classified as tumor class A.

Steps 1, 2 and 3 are repeated for each test sample for all tumor classes.

by simply calculating the mean and median of each column and selects the column for which measures of central tendency were the highest.

Results

Prediction accuracy determination of the optimum number of genes (features) to be used by the classification algorithm is usually a difficult task that depends on several factors including the classifier rules and the complexity of the data set. Finding the optimal number of genes is generally very difficult. Many practical solutions are based on experience or some heuristics.[14] For binary regression algorithms, previous studies[15,16] showed a feature set of one to two hundred top genes was adequate for a simple two-category problem. In this study, the size of the feature set was 50, 75, 100 and 200 for the BC dataset and 50, 100 and 200 for the GCM dataset.

BC dataset

The prediction accuracies when OVA and OVO output-coding strategies were implemented for the 5 replications using fold change and penalized t-test for feature selection are summarized in Tables 3 and 4 respectively. The highest average prediction accuracy of the 5 replications obtained for fold change for OVO was 93.33% for 75 genes, where as, for OVA the highest average was 83.33% for 75 genes. For penalized t-test, the highest average for OVO was 91.66% for 50 genes while, for OVA, the highest

Table 3. Prediction accuracies for the BC dataset using fold change as the feature selection method.

No. of genes	50		75		100		200	
Coding method	OVA (%)	OVO (%)	OVA (%)	OVO (%)	OVA (%)	OVO (%)	OVA (%)	OVO (%)
Rep 1	66.67	91.66	75.00	91.66	91.66	91.66	91.66	91.66
Rep 2	75.00	66.67	75.00	83.33	75.00	83.33	66.67	75.00
Rep 3	83.33	91.66	83.33	91.66	83.33	91.66	83.33	100.00
Rep 4	83.33	100.00	91.66	100.00	91.66	100.00	91.66	100.00
Rep 5	83.33	100.00	91.67	100.00	66.67	100.00	66.67	91.67
Average	78.332	89.99	83.33	93.33	81.66	93.33	79.99	91.66

Table 4. Prediction accuracies for the BC dataset using penalized t-statistics as the feature selection method.

No. of genes	50		75		100		200	
Coding method	OVA (%)	OVO (%)	OVA (%)	OVO (%)	OVA (%)	OVO (%)	OVA (%)	OVO (%)
Rep 1	83.33	91.66	75.00	100.00	75.00	100.00	91.66	91.66
Rep 2	75.00	91.66	75.00	75.00	75.00	75.00	66.67	75.00
Rep 3	91.67	91.66	91.66	83.33	91.66	83.33	91.66	83.33
Rep 4	83.33	83.33	75.00	91.66	66.67	91.66	66.7	91.66
Rep 5	50.00	100.00	83.33	100.00	83.33	91.67	91.67	100.00
Average	76.66	91.66	80.00	89.99	78.33	88.32	81.67	88.32

prediction accuracy was 81.67% for 200 genes. This result shows that OVO strategy could give a higher prediction accuracy compared to OVA. Also features (genes) selected by fold change gained highest accuracies for both OVO and OVA. There was an improvement of around 12% by using OVO method compared to OVA method for fold change and 15% improvement for penalized t-test for the BC dataset. Leung et al[16] got 17% improvement by using OVO over OVA using t-test for feature selection and k-means clustering for classification.

GCM Dataset

The prediction accuracy of the 54 validation (test) samples, using fold change and penalized t-test, is summarized in Tables 5 and 6 respectively. The highest prediction accuracy for OVO with fold change feature selection was 75.92% for 50 genes where as for the same number of genes and fold change, OVA performed higher (79.6%). For 200 genes, the fold change feature selection method performed equally for both OVO and OVA (70.4%). The highest prediction accuracy for OVO with penalized t-test was 74% for 50 genes where as an accuracy of 79.6% was obtained for OVO method for 100 genes when the second highest mean was also considered for each sample (data not shown). For 200 genes selected based on penalized t-test, the OVO had a higher accuracy of 3% compared to OVA, which is not that significant. But for 50 genes and penalized t-test showed higher performance in OVA compared to OVO.

Several works have been done using this GCM dataset using various feature selection methods and classification methods on OVA set up. Using a recursive feature selection procedure and support vector machine (SVM) classification algorithm, Ramaswamy et al[4] obtained their best result with 42 tumors correctly predicted among the 54 test samples, corresponding to an accuracy of 78%. Using a feature selection algorithm based on overlaps of gene expression values between different classes in conjunction with the Covering Classification Algorithm (CCA), a modification of the k-NN method, Bagirov et al[17] achieved prediction accuracy of around 80%. Based on the concept of gene interaction, Antonov et al[18] proposed a Maximal Margin Linear Programming (MAMA) procedure that combines linear programming and SVM and they got around 85.2% classification accuracy on an OVA set up. Combining all these previous research and the results obtained from this study, it may be concluded that for GCM dataset, OVA performs almost equally as that of OVO.

Discussions

Classification problems aim at building an efficient, effective model for predicting class membership

Table 5. Prediction accuracies for the GCM dataset using fold change as the feature selection method.

No. of genes	OVO (Accuracy in %)	OVA (Accuracy in %)
200	70.3	70.4
100	72.2	74.5
50	75.92	79.6

Table 6. Prediction accuracies for the GCM dataset using penalized t-statistics as the feature selection method.

No. of genes	OVO (Accuracy in %)	OVA (Accuracy in %)
200	66.00	63.00
100	65.00	73.25
50	74.00	75.92

from the data. The learning process is done on the training data, which consists of data points chosen from the input data space and their class label. A model built from the training data is expected to produce the correct label on the training data and also to predict correctly the identity of an unknown class. In cases where there is only two classes, a classification problem is said to be a binary, while many real-world problems are multi-class problems. Majority of the solutions for multi-class problems is done by decomposing the multiple classes into binary ones. One-versus-the rest methods, pair wise comparison, error-correcting output-coding (ECOC)[20] are examples. The main criticism against OVR is because of the inconsistent class assignment as well as solving potentially asymmetric problems using a symmetric approach.[14,21] OVO also can be criticized for solving asymmetric problems symmetrically, but one advantage of using this method is that each classifier is easy to train since it is purely a binary problem even though the number of comparisons will be higher than OVR.

The usual way of implementing OVO is based on vote, which class label gets the largest number of votes, that test sample will be assigned that class label. While the OVO implemented in this study uses summary of the posterior distribution (mean and median) of the classification probability to determine the fate of the test sample.

Latent Variable models (LVMs) postulate that the observed discrete (binary data) is only an indicator of an underlying non-observed random variable that follows a certain distribution (often normal distribution). The discrete responses are observed when the latent variable exceeds a certain threshold(s). For example for the binary situation, the binary response is 1 (case) when the latent variable exceeds threshold, T, (i.e. $T > 0$) otherwise it is zero (control). The main advantage of using LVM along with the proposed OVO in the context of a Bayesian implementation via Gibbs sampler is because its ability of providing the full posterior distribution of the classification probability.[13,15,22] Other binary classifiers like SVM will output only the class label (+, or −) of the test sample. However, knowing the class label or the predictive value a lot of times is not good enough to evaluate a classification. This makes LVM a good classifier that suits well with the proposed OVO strategy.

It has been mentioned that there is probably no multiclass method using binary decomposition approach that outperforms every thing else and that for practical purposes the choice of the method has to be made depending on the constraints, such as the desired level of accuracy, the time available for development and training, and the nature of the classification problem.[6] The same conclusion has been made in the current study since different gene expression data sets showed differences in the prediction accuracies. In OVO decomposition strategy, we need to do $2x(K-1)/2$-times as many binary classifiers than in OVA method which results in more computational time. One of the rather power full approach of the OVO method is the high demands for the representatively of the training set. We noticed that the OVO method for GCM data did not make much improvement in prediction accuracies. The main reason based on above discussion might be because of the disparity among the number of training samples in each of the 14 classes. Moreover, high heterogeneity observed within subclasses of the GCM data especially the Breast cancer

Table 7. Number of misclassified test samples of Breast (BR) and their predicted tumour subtype according to OVO for both fold change and penalized t-statistics.

No. of test samples for BR tumour type	No. of genes (features)	No. of misclassified samples (using fold change)[c]	No. of misclassified samples (using penalized t-statistics)
4	200	3 (LE, UT, PA)	3 (LU, LU, PA)
4	100	3 (LE, UT, PA)	3 (LU, LU, PA)
4	50	2 (LE, PA)	3 (LE, PA, LU)

[c]In parenthesis is the assigned tumour types for the misclassified BR test samples.

(BR) sample might be one of the reasons for less prediction accuracies (Table 7). In fact, among the 4 BR test samples, previous studies by Ramaswamy et al[4] and Bagirov et al[17] could correctly predict only 2 and 1 test samples, respectively, indicating potential prediction uncertainty which in turn affects the classification accuracy. Potential chances of mislabeling of tumor samples as proved by Zhang et al[15] might be another reason for reduced classification accuracy. It is obvious that higher prediction accuracies obtained while using the BC dataset in OVO methods might be because of sufficient representation of training data in each cancer type as well as homogeneous gene expression data within each class. Therefore it is intuitive that datasets that have enough representation in the training data for each class and training samples that have more or less homogeneous expression patterns of within each subtype, OVO will definitely out-perform OVA as we observe in BC data set. One of the drawbacks of using the proposed OVO might be the occurrence of over fitting although the latent variable model handled it quite robustly.

Conclusions

The output-coding scheme from machine learning has been successfully applied to multi-class microarray classification in this paper. It has been shown that a good coding matrix can lead to high accuracy of multi-class microarray classification. Better coding strategies are required to further improve the performance of the output coding scheme. This study demonstrated that the choice of feature selection statistics, comparison method between groups and classification algorithms could all affect the interpretation of final results of microarray data. More emphasis is given on the comparison methods and it could be concluded that the selection of OVO or OVA depends upon the data structure and the type of microarray experiment. Based on the BC dataset, we can assume that when dealing with multi-class cancer type datasets, OVO comparison method can give better performance accuracy than the commonly used OVA. But we could not conclude the same inference with the GCM dataset. The main reason could be: large number of tumor types is considered at the same time or may be because of the heterogeneity within and among the cancer subtypes or chances of potential mislabeling of tumor subtypes in the training data.

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors report no conflicts of interest.

References

1. Golub TR, Slonim DK, Tamayo P, et al. C Molecular classification of cancer: Class discovery and class prediction by gene expression prediction. *Science*. 1999;286:531–7.
2. Ross DT, Seherf U, Eisen MB, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*. 2000;24:227–35.
3. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*. 2000;16:906–14.
4. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceeding of National Academy of Sciences*. 2001;98(26):15149–54.
5. Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for classification of tumors using gene expression data. *J Am Stat Ass*. 2002;97:77–87.
6. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*. 2005;21:631–43.
7. Hastie TJ, Tibshirani RJ. Classification by pairwise coupling. In: Jordan MI, Kearnsa MJ, Solla SA, eds. Advances in neural information processing systems. Volume 10, MIT Press, 1998.
8. Allwein EL, Schapire RE, Singer Y. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*. 2000;1:113–41.
9. Dietterich TG, Bakiri G. Error-correcting output codes: A general method for improving multiclass inductive learning programs. Proceedings of the Ninth National Conference on Artificial Intelligence. AAAI Press. 1991; 572–7.
10. Yeang CH, Ramaswamy S, Tamayo P, et al. Molecular classification of multiple tumor types. *Bioinformatics*. 2001;17(1):S316–22.
11. Pomeroy SL, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumor outcome based on gene expression. *Nature*. 2002;415(6870):436–42.
12. Johnson VE, Albert JH. Ordinary data model. Springer: New York, 1999.
13. West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of National Academy of Sciences*. 2001;98(20):11462–7.
14. Li T, Zhang C, Ogihara M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*. 2004;20:2429–37.
15. Zhang W, Rekaya R, Bertrand JK. Method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer. *Bioinformatics*. 2006;22:317–25.
16. Leung YY, Chang CQ, Hung YS, Fung PCW. Gene selection for brain cancer classification. *Proceedings of the 28th IEEE EMBS Annual Internationl Conference*. New York City, USA, Aug 30–Sept 3, 2006.
17. Bagirov AM, Ferguson B, Ivkovic S, Aaunders G, Yearwood J. Modified global k-means algorithm for clustering in gene expression data sets. *Bioinformatics*. 2003;19:1800–7.
18. Antonov AV, Tetko IV, Mader MT, Budczies J, Mewes HW. Optimization models for cancer classification: extracting gene interaction information from microarray expression data. *Bioinformatics*. 2004; 20:644–52.

19. Scholkopf B, Smola JA. Learning with Kernels. MIT Press, Cambridge, MA, 2002.
20. Dietterich TG, Bakiri G. Solving multiclass learning problem via error-correcting output codes. *J Artif Intell Res*. 1995;2:263–86.
21. Scholkopf B, Smola JA. Learning with Kernels. MIT Press, Cambridge, MA, 2002.
22. Robbins K, Joseph S, Zhang W, Rekaya R, Bertrand JK. Classification of incipient Alzheimer patients using gene expression data: Dealing with potential misdiagnosis. *Online J Bioinformatics*. 2006;7:22–31.