OVA Scheme vs. Single Machine Approach in Feature Selection for Microarray Datasets

Chia Huey Ooi, Madhu Chetty, and Shyh Wei Teng

Gippsland School of Information Technology
Monash University, Churchill, VIC 3842, Australia
{chia.huey.ooi, madhu.chetty,
shyh.wei.teng}@infotech.monash.edu.au

Abstract. The large number of genes in microarray data makes feature selection techniques more crucial than ever. From rank-based filter techniques to classifier-based wrapper techniques, many studies have devised their own feature selection techniques for microarray datasets. By combining the OVA (one-vs.-all) approach and differential prioritization in our feature selection technique, we ensure that class-specific relevant features are selected while guarding against redundancy in predictor set at the same time. In this paper we present the OVA version of our differential prioritization-based feature selection technique and demonstrate how it works better than the original SMA (single machine approach) version.

Keywords: molecular classification, microarray data analysis, feature selection.

1 Feature Selection in Tumor Classification

Classification of tumor samples from patients is vital for diagnosis and effective treatment of cancer. Traditionally, such classification relies on observations regarding the location [1] and microscopic appearance of the cancerous cells [2]. These methods have proven to be slow and ineffective; there is no way of predicting with reliable accuracy the progress of the disease, since tumors of similar appearance have been known to take different paths in the course of time. Some tumors may grow aggressively after the point of the abovementioned observations, and hence require equally aggressive treatment regimes; other tumors may stay inactive and thus require no treatment at all [1]. With the advent of the microarray technology, data regarding the gene expression levels in each tumor samples now may prove a useful tool in aiding tumor classification. This is because the microarray technology has made it possible to simultaneously measure the expression levels for thousands or tens of thousands of genes in a single experiment [3, 4].

However, the microarray technology is a two-edged sword. Although with it we stand to gain more information regarding the gene expression states in tumors, the amount of information might simply be too much to be of use. The large number of features (genes) in a typical gene expression dataset (1000 to 10000) intensifies the need for feature selection techniques prior to tumor classification. From various filter-based procedures [5] to classifier-based wrapper techniques [6] to filter-wrapper

P. Perner (Ed.): ICDM 2006, LNAI 4065, pp. 10 – 23, 2006.

hybrid techniques [7], many studies have devised their own flavor of feature selection techniques for gene expression data. However, in the context of highly multiclass microarray data, only a handful of them have delved into the effect of redundancy in the predictor set on classification accuracy.

Moreover, the element of the balance between relative weights given to relevance vs. redundancy also assumes an equal, if not greater importance in feature selection. This element has not been given the attention it deserves in the field of feature selection, especially in the case of applications to gene expression data with its large number of features, continuous values, and multiclass nature. Therefore, to solve this problem, we introduced the element of the DDP (**d**egree of **d**ifferential **p**rioritization) as a third criterion to be used in feature selection along with the two existing criteria of relevance and redundancy [8].

2 Classifier Aggregation for Tumor Classification

In the field of classification and machine learning, multiclass problems are often decomposed into multiple two-class sub-problems, resulting in classifier aggregation. The rationale behind this is that two-class problems are easier to solve than multiclass problems. However, classifier aggregation may increase the order of complexity by up to a factor of B, B being the number of the decomposed two-class sub-problems. This argument for the single machine approach (SMA) is often countered by the theoretical foundation and empirical strengths of the classifier aggregation approach. The term single machine refers to the fact that a predictor set is used to train only one classifier. Here, we differentiate between internal and external classifier aggregation.

Internal classifier aggregation transpires when feature selection is conducted once based on the original multiclass target class concept. The single predictor set obtained is then fed as input into a single multiclassifier. The single multiclassifier trains its component binary classifiers accordingly, but using the same predictor set for all component binary classifiers. *External classifier aggregation* occurs when feature selection is conducted separately for each two-class sub-problem resulting from the decomposition of the original multiclass problem. The predictor set obtained for each two-class sub-problem is different from the predictor sets obtained for the other two-class sub-problems. Then, in each two-class sub-problem, the aforementioned predictor set is used to train a binary classifier.

Our study is geared towards comparing external classifier aggregation in the form of the one-vs.-all (OVA) scheme against the SMA. From this point onwards, the term *classifier aggregation* will refer to external classifier aggregation. Methods in which feature selection is conducted based on the multiclass target class concept are defined as SMA methods, regardless of whether a multiclassifier with internal classifier aggregation or a direct multiclassifier (which employs no aggregation) is used. Examples of multiclassifier with internal classifier aggregation are multiclass SVMs based on binary SVMs such as DAGSVM [9], "one-vs.-all" and "one-vs.-one" SVMs. Direct multiclassifiers include nearest neighbors, Naïve Bayes [10], other maximum likelihood discriminants and true multiclass SVMs such as BSVM [11].

Various classification and feature selection studies have been conducted for multiclass microarray datasets. Most involved SMA with either one of or both direct and

internally aggregated classifiers [8, 12, 13, 14, 15]. Two studies [16, 17] did implement external classifier aggregation in the form of the OVA scheme, but only on a single split of a single dataset, the GCM dataset. Although in [17], various multiclass decomposition techniques were compared to each other and the direct multiclassifier, classifier methods, and not feature selection techniques, were the main theme of that study.

This brief survey of existent studies indicates that both the SMA and OVA scheme are employed in feature selection for multiclass microarray datasets. However, none of these studies have conducted a detailed analysis which applies the two paradigms in parallel on the same set of feature selection techniques, with the aim of judging the effectiveness of the SMA against the OVA scheme (or vice versa) on feature selection techniques for multiclass microarray datasets. To address this deficiency, we devise the OVA version of the DDP-based feature selection technique introduced earlier [8].

The main contribution of this paper is to study the effectiveness of the OVA scheme against the SMA, particularly for the DDP-based feature selection technique. A secondary contribution is an insightful finding on the role played by aggregation schemes such as the OVA in influencing the optimal value of the DDP.

We begin with a brief description of the SMA version of the DDP-based feature selection technique, followed by the OVA scheme for the same feature selection technique. Then, after comparing the results from both SMA and OVA versions of the DDP-based feature selection technique, we discuss the advantages of the OVA scheme over the SMA, and present our conclusions.

3 SMA Version of the DDP-Based Feature Selection Technique

For microarray datasets, the term *gene* and *feature* may be used interchangeably. The training set upon which feature selection is to be implemented, T, consists of N genes and M_t training samples. Sample j is represented by a vector, $\mathbf{x}_j$, containing the expression of the N genes $[x_{1,j},\ldots, x_{N,j}]^T$ and a scalar, y_j, representing the class the sample belongs to. The SMA multiclass target class concept $\mathbf{y}$ is defined as $[y_1, \ldots, y_{Mt}]$, $y_j \in [1,K]$ in a K-class dataset. From the total of N genes, the objective is to form the subset of genes, called the predictor set S, which would give the optimal classification accuracy. For the purpose of defining the DDP-based predictor set score, we define the following parameters.

- V_S is the measure of relevance for the candidate predictor set S. It is taken as the average of the score of relevance, $F(i)$ of all members of the predictor set [14]:

$$V_S = \frac{1}{|S|} \sum_{i \in S} F(i) \tag{1}$$

$F(i)$ indicates the correlation of gene i to the SMA target class concept $\mathbf{y}$, i.e., ability of gene i to distinguish among samples from K different classes at once. A popular parameter for computing $F(i)$ is the BSS/WSS ratios (the F-test statistics) used in [14, 15].

- U_S is the measure of antiredundancy for the candidate predictor set S. U_S quantifies the *lack of redundancy* in S.

$$U_S = \frac{1}{|S|^2} \sum_{i,j \in S} 1 - |R(i,j)| \tag{2}$$

$|R(i,j)|$ measures the similarity between genes i and j. $R(i,j)$ is the Pearson product moment correlation coefficient between genes i and j. Larger U_S indicates lower average pairwise similarity in S, and hence, smaller amount of redundancy in S.

The measure of goodness for predictor set S, $W_{A,S}$, incorporates both V_S and U_S.

$$W_{A,S} = (V_S)^{\alpha} \cdot (U_S)^{1-\alpha} \tag{3}$$

where the power factor $\alpha \in (0, 1]$ denotes the degree of differential prioritization between maximizing relevance and maximizing antiredundancy.

Decreasing the value of α forces the search method to put more priority on maximizing antiredundancy at the cost of maximizing relevance. Raising the value of α increases the emphasis on maximizing relevance (at the same time decreases the emphasis on maximizing antiredundancy) during the search for the optimal predictor set. A predictor set found using larger value of α has more features with strong relevance to the target class concept, but also more redundancy among these features. Conversely, a predictor set obtained using smaller value of α contains less redundancy among its member features, but at the same time also has fewer features with strong relevance to the target class concept.

The SMA version of the DDP-based feature selection technique has been shown to be capable of selecting the optimal predictor set for various multiclass microarray datasets by virtue of the variable differential prioritization factor [8]. Results from the application of this feature selection technique on multiple datasets [8] indicate two important correlations to the number of classes, K, of the dataset: As K increases,

1. the estimate of accuracy deteriorates, especially for K greater than 6; and
2. placing more emphasis on maximizing antiredundancy (using smaller α) produces better accuracy than placing more emphasis on relevance (using larger α).

From these observations, we conclude that as K increases, for majority of the classes, features highly relevant with regard to a specific class are more likely to be 'missed' by a multiclass score of relevance (i.e., given a low multiclass relevance score) than by a class-specific score of relevance. In other words, the measure of relevance computed based on the SMA multiclass target class concept is not efficient enough to capture the relevance of a feature when K is larger than 6.

Moreover, there is an imbalance among the classes in the following aspect: For class k (k = 1, 2, …, K), let h_k be the number of features which have high class-specific (class k vs. all other classes) relevance and are also deemed highly relevant by the SMA multiclass relevance score. For all benchmark datasets, h_k varies greatly from class to class. Hence, we need a classifier aggregation scheme which uses class-specific target class concept catering to a particular class in each sub-problem and is thus better able to capture features with high correlation to a specific class. This is where the proposed OVA scheme is expected to play its role.

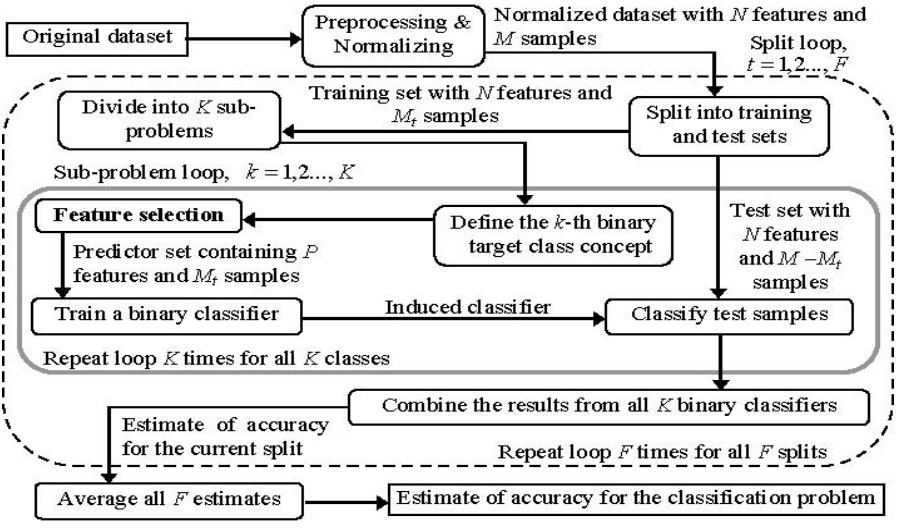

Fig. 1. Feature selection using the OVA scheme

4 OVA Scheme for the DDP-Based Feature Selection Technique

In the OVA scheme, a *K*-class feature selection problem is divided into *K* separate 2-class feature selection sub-problems (Figure 1). Each of the *K* sub-problems has a target class concept different from the target class concept of the other sub-problems and that of the SMA. Without loss of generality, in the *k*-th sub-problem (k = 1, 2, …, K), we define class 1 as encompassing all samples belonging to class *k*, and class 2 as comprising of all samples *not* belonging to class *k*. In the *k*-th sub-problem, the target class concept, $\mathbf{y_k}$, is a 2-class target class concept.

$$\mathbf{y_k} = \begin{bmatrix} y_{k,1} & y_{k,2} & \cdots & y_{k,M_t} \end{bmatrix} \tag{4}$$

where

$$y_{k,j} = \begin{cases} 1 & \text{if} \quad y_j = k \\ 2 & \text{if} \quad y_j \neq k \end{cases} \tag{5}$$

In solving the *k*-th sub-problem, feature selection finds the predictor set S_k, the size of which, *P*, is generally much smaller than *N*. Therefore, for each tested value of P = 2, 3, …, P_{max}, *K* predictor sets are obtained from all *K* sub-problems. For each value of *P*, the *k*-th predictor set is used to train a component binary classifier which then attempts to predict whether a sample belongs or does *not* belong to class *k*. The predictions from *K* component binary classifiers are combined to produce the overall prediction. In cases where more than one of the *K* component binary classifiers pro-claims a sample as belonging to their respective classes, the sample is assigned to the class corresponding to the component binary classifier with the largest decision value.

Equal predictor set size is used for all K sub-problems, i.e., the value of P is the same for all of the K predictor sets.

In the k-th sub-problem, the predictor set score for S_k, W_{A,S_k}, is given as follows.

$$W_{A,S_k} = \left(V_{S_k}\right)^{\alpha} \cdot \left(U_{S_k}\right)^{1-\alpha} \tag{6}$$

The significance of α in the OVA scheme remains unchanged in the general meaning of the SMA context. However, it must be noted that the power factor $\alpha \in (0, 1]$ now represents the degree of differential prioritization between maximizing relevance *based on the 2-class target class concept*, $\mathbf{y_k}$, (instead of relevance based on the K-class target class concept $\mathbf{y}$ of the SMA) and maximizing antiredundancy.

Aside from these differences, the role of α is the same in the OVA scheme as in the SMA. For instance, at $\alpha = 0.5$, we still get an equal-priorities scoring method, and at $\alpha = 1$, the feature selection technique becomes rank-based.

The measure of relevance for S_k, V_{S_k}, is computed by averaging the score of relevance, $F(i,k)$ of all members of the predictor set.

$$V_{S_k} = \frac{1}{|S_k|} \sum_{i \in S_k} F(i,k) \tag{7}$$

The score of relevance of gene i in the k-th sub-problem, $F(i,k)$, is given as follows.

$$F(i,k) = \frac{\sum_{j=1}^{M_t} \sum_{q=1}^{2} I\left(y_{k,j} = q\right)\left(\bar{x}_{iq} - \bar{x}_{i\bullet}\right)^2}{\sum_{j=1}^{M_t} \sum_{q=1}^{2} I\left(y_{k,j} = q\right)\left(x_{ij} - \bar{x}_{iq}\right)^2} \tag{8}$$

$I(.)$ is an indicator function returning 1 if the condition inside the parentheses is true, otherwise it returns 0. $\bar{x}_{i\bullet}$ is the average of the expression of gene i across all training samples. $\bar{x}_{iq}$ is the average of the expression of gene i across training samples belonging to class k when q is 1. When q is 2, $\bar{x}_{iq}$ is the average of the expression of gene i across training samples *not* belonging to class k.

The measure of antiredundancy for S_k, U_{S_k}, is computed the same way as in the SMA.

$$U_{S_k} = \frac{1}{|S_k|^2} \sum_{i,j \in S_k} 1 - |R(i,j)| \tag{9}$$

For search method, in the k-th sub-problem, we use the linear incremental search [14] given below. The order of computation is $O(NKP_{\max})$.

1. For $k = 1, 2, \ldots, K$, do
   1.1. Choose the gene with the largest $F(i,k)$ as the first member of S_k.

1.2. For $P = 2, 3, \ldots, P_{\max}$

1.2.1. Screen the remaining $(N - P + 1)$ genes one by one to find the gene that would enable S_k to achieve the maximum W_{A,S_k} for the size P.

1.2.2. Insert such gene as found in 1.2.1 into S_k.

5 Results

Feature selection experiments were conducted on seven benchmark datasets using both the SMA and the OVA scheme. In both approaches, different values of α from 0.1 to 1 were tested with equal intervals of 0.1. The characteristics of microarray datasets used as benchmark datasets: the GCM [16], NCI60 [18], lung [19], MLL [20], AML/ALL [21], PDL [22] and SRBC [23] datasets, are listed in Table 1. For NCI60, only 8 tumor classes are analyzed; the 2 samples of the prostate class are excluded due to the small class size. Datasets are preprocessed and normalized based on the recommended procedures in [15] for Affymetrix and cDNA microarray data.

Table 1. Descriptions of benchmark datasets. N is the number of features after preprocessing.

Dataset	Type	N	K	Training:Test set size
GCM	Affymetrix	10820	14	144:54
NCI60	cDNA	7386	8	40:20
PDL	Affymetrix	12011	6	166:82
Lung	Affymetrix	1741	5	135:68
SRBC	cDNA	2308	4	55:28
MLL	Affymetrix	8681	3	48:24
AML/ALL	Affymetrix	3571	3	48:24

With the exception of the GCM dataset, where the original ratio of training to test set size used in [16] is maintained to enable comparison with previous studies, for all other datasets we employ the standard 2:1 split ratio. The DAGSVM classifier is used throughout the performance evaluation. The DAGSVM is an all-pairs SVM-based multiclassifier which uses less training time compared to either the standard algorithm or Max Wins while producing accuracy comparable to both [9].

5.1 Evaluation Techniques

For the OVA scheme, the exact evaluation procedure for a predictor set of size P found using a certain value of the DDP, α, is shown in Figure 1. In case of the SMA, the sub-problem loop in Figure 1 is conducted only once, and that single sub-problem represents the (overall) K-class problem. Three measures are used to evaluate the overall classification performance of our feature selection techniques. The first is the *best averaged accuracy*. This is simply taken as the largest among the accuracy obtained from Figure 1 for all values of P and α. The number of splits, F, is set to 10.

The second measure is obtained by averaging the estimates of accuracy from different sizes of predictor sets ($P = 2, 3, \ldots, P_{max}$) obtained using a certain value of α to get the *size-averaged accuracy* for that value of α. This parameter is useful in predicting the value of α likely to produce the optimal estimate of accuracy since our feature selection technique does not explicitly predict the best P from the tested range of $[2, P_{max}]$. The size-averaged accuracy is computed as follows. First, for all predictor sets found using a particular value of α, we plot the estimate of accuracy obtained from the procedure outlined in Figure 1 against the value of P of the corresponding predictor set (Figure 2). The size-averaged accuracy for that value of α is the area under the curve in Figure 2 divided by the number of predictor sets, $(P_{max}-1)$.

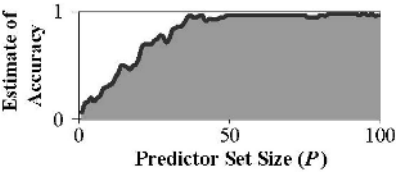

Fig. 2. Area under the accuracy-predictor set size curve

The value of α associated with the highest size-averaged accuracy is deemed the empirical optimal value of the DDP or the empirical estimate of α^*. Where there is a tie in terms of the highest size-averaged accuracy between different values of α, the empirical estimate of α^* is taken as the average of those values of α.

The third measure is *class accuracy*. This is computed in the same way as the size-averaged accuracy, the only difference being that instead of overall accuracy, we compute the class-specific accuracy for each class of the dataset. Therefore there are a total of K class accuracies for a K-class dataset.

In this study, P_{max} is deliberately set to 100 for the SMA and 30 for the OVA scheme. The rationale for this difference is that more features will be needed to differentiate among K classes at once in the SMA, whereas in the OVA scheme, each predictor set from the k-th sub-problem is used to differentiate between only two classes, hence the smaller upper limit to the number of features in the predictor set.

5.2 Best Averaged Accuracy

Based on the best averaged accuracy, the most remarkable improvement brought by the OVA scheme over the SMA is seen in the dataset with the largest number of classes ($K = 14$), GCM (Table 2). The accuracy of 80.6% obtained from the SMA is increased by nearly 2% to 82.4% using the OVA scheme. For the NCI60, lung and SRBC datasets there is a slight improvement of 1% at most in the best averaged accuracy when the OVA scheme is compared to the SMA. The performance of the SMA version of the DDP-based feature selection technique for the two most challenging benchmark datasets (GCM and NCI60) has been compared favorably to results from

previous studies in [8]. Therefore it follows that the accuracies from the OVA scheme compare even more favorably to accuracies obtained in previous studies on these datasets [12, 14, 15, 16, 17].

Naturally, the combined predictor set size obtained from the OVA scheme is greater than that obtained from the SMA. However, we must note that the predictor set size *per component binary classifier* (i.e., the number of genes per component binary classifier) associated with the best averaged accuracy is smaller in case of the OVA scheme than the SMA (Table 2). Furthermore, we consider two facts: 1) There are K component binary classifiers involved in the OVA scheme where the component DAGSVM reverts to a plain binary SVM in each of the K sub-problems. 2) On the other hand, there are ${}^{K}C_{2}$ component binary classifiers involved in the multiclassifier used in the SMA, the all-pairs DAGSVM. Therefore, 1) the smaller number of component binary classifiers and 2) the smaller number of genes used per component binary classifier in the OVA scheme serve to emphasize the superiority of the OVA scheme over the SMA in producing better accuracies for datasets with larger K such as the GCM and NCI60 datasets.

For the PDL dataset, the best averaged accuracy deteriorates by 2.8% when the OVA scheme replaces the SMA. For the datasets with the least number of classes ($K = 3$), the best averaged accuracy is the same whether obtained from predictor set produced from feature selection using the SMA or the OVA scheme.

Table 2. Best averaged accuracy (± standard deviation across F splits) estimated from feature selection using the SMA and OVA scheme, followed by the corresponding differential prioritization factor and predictor set size ('gpc' stands for 'genes per component binary classifier')

Dataset	SMA	OVA
GCM	80.6 ± 4.3%, α=0.2, 85 gpc	82.4 ± 3.3%, α=0.3, 24 gpc
NCI60	74.0 ± 3.9%, α=0.3, 61 gpc	75.0 ± 6.2%, α=0.3, 19 gpc
PDL	99.0 ± 1.0%, α=0.5, 60 gpc	96.2 ± 1.1%, α=0.6, 16 gpc
Lung	95.6 ± 1.6%, α=0.5, 31 gpc	96.0 ± 1.7%, α=0.5, 14 gpc
SRBC	99.6 ± 1.1%, α=0.7, 13 gpc	100 ± 0%, α=0.8, 2 gpc
MLL	99.2 ± 1.8%, α=0.6, 12 gpc	99.2 ± 1.8%, α=0.7, 4 gpc
AML/ALL	97.9 ± 2.2%, α=0.8, 11 gpc	97.9 ± 2.2%, α=0.6, 6 gpc

5.3 Size-Averaged Accuracy

The best size-averaged accuracy for the OVA scheme is better for all benchmark datasets except the PDL and AML/ALL datasets (Table 3). The peak of the size-averaged accuracy plot against α for the OVA scheme appears to the right of the peak of the SMA plot for all datasets except the PDL and lung datasets, where they stay the same for both approaches (Figure 3). This means that the value of the optimal DDP (α^*) when the OVA scheme is used in feature selection is greater than the optimal DDP (α^*) obtained from feature selection using the SMA, except for the PDL and lung datasets. In Section 6, we will look into the reasons for the difference in the empirical estimates of α^* between the two approaches of the SMA and the OVA scheme.

Table 3. Best size-averaged accuracy estimated from feature selection using the SMA and OVA scheme, followed by the corresponding DDP, α^*. A is the number of times OVA outperforms SMA, and B is the number of times SMA outperforms OVA, out of the total of tested values of $P = 2, 3, \ldots, 30$.

Dataset	SMA	B	OVA	A
GCM	68.2%, α^*=0.2	0	76.0%, α^*=0.5	29
NCI60	60.1%, α^*=0.3	0	64.4%, α^*=0.6	29
PDL	94.0%, α^*=0.5	0	92.3%, α^*=0.5	19
Lung	91.8%, α^*=0.6	1	92.3%, α^*=0.6	12
SRBC	97.3%, α^*=0.6	0	99.9%, α^*=0.9	26
MLL	96.8%, α^*=0.7	0	97.4%, α^*=0.8	12
AML/ALL	95.9%, α^*=0.8	0	95.6%, α^*=0.9	9

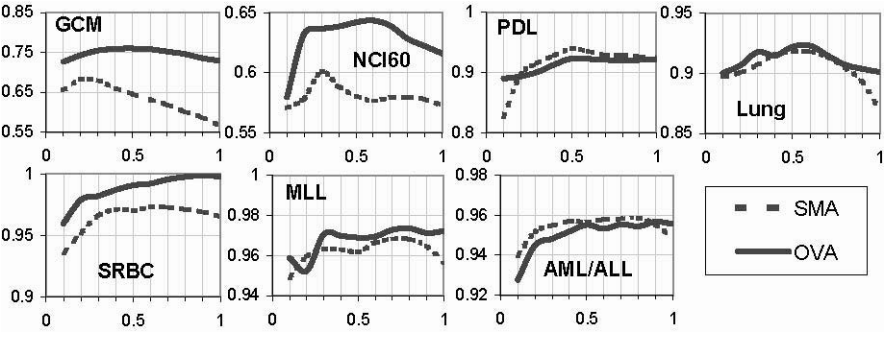

Fig. 3. Size-averaged accuracy plotted against α

We have also conducted statistical tests on the significance of the performance of each of the approaches (SMA or OVA) over the other for each value of P (number of genes per component binary classifier) from $P = 2$ up to $P = 30$. Using Cochran's Q statistic, the number of times the OVA approach outperforms the SMA, A, and the number of times the SMA outperforms the OVA approach, B, at 5% significance level, are shown in Table 3. It is observed that $A > B$ for all seven datasets, and that A is especially large (in fact, maximum) for the two datasets with largest number of classes, the GCM and NCI60 datasets. Moreover, A tends to increase as K increases, showing that the OVA approach increasingly outperforms the SMA (at 5% significance level) as the number of classes in the dataset increases.

5.4 Class Accuracy

To explain the improvement of the OVA scheme over the SMA, we look towards the components that contribute to the overall estimate of accuracy: the estimates of the class accuracy. Does the improvement in size-averaged accuracy in the OVA scheme translate to similar increase in the class accuracy of each of the classes in the dataset?

To answer the question, for each class in a dataset, we compute the difference between class accuracy obtained from the OVA scheme and that from the SMA using corresponding values of α^* from Table 3. Then, we obtain the average of this difference from all classes in the same dataset. **Positive** difference indicates **improvement** brought by the OVA scheme against the SMA. For each dataset, we also count the number of classes whose class accuracy is better under the OVA scheme than in the SMA and divide this number by K to obtain a percentage. These two parameters are then plotted for all datasets (Figure 4).

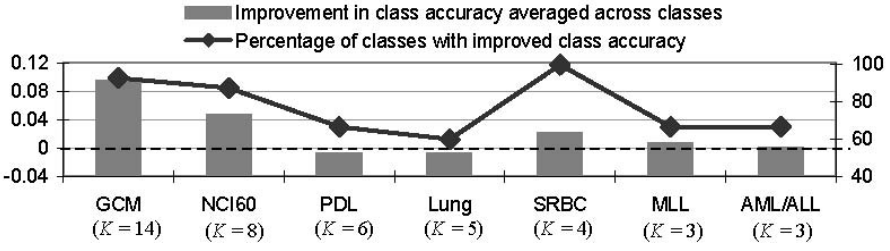

Fig. 4. Improvement in class accuracy averaged across classes (left axis) and percentage of classes with improved class accuracy (right axis) for the benchmark datasets

Figure 4 provides two observations. Firstly, for **all** datasets, the minimum percentage of classes whose class accuracy has been improved by the OVA scheme is 60%. This indicates that the OVA scheme feature selection is capable of increasing the class accuracy of the *majority* of the classes in a multiclass dataset. Secondly, the average improvement in class accuracy is highest in datasets with largest K, the GCM and the NCI60 (above 4%). Furthermore, only one class out of 14 and 8 classes for the GCM and NCI60 datasets respectively does not show improved class accuracy under the OVA scheme (compared to the SMA). Therefore, the OVA scheme brings the largest amount of improvement over the SMA for datasets with large K.

In several cases, improvement in class accuracy occurs only for classes with small class sizes, which is not sufficient to compensate for the deterioration in class accuracy for classes with larger class sizes. Therefore, even if majority of the classes show improved class accuracy under the OVA scheme, this does not get translated into improved overall accuracy (PDL and AML/ALL datasets) or improved averaged class accuracy (PDL and lung datasets) when a few of the larger classes have worse class accuracy.

6 Discussion

For both approaches, maximizing antiredundancy is less important for datasets with smaller K (less than 6) – therefore supporting the assertion in [24] that redundancy does not hinder the performance of the predictor set when K is 2. In the SMA feature selection, the value of α^* is more strongly influenced by K compared to the case in the OVA scheme feature selection. The correlation between α^* and K in the SMA is

found to be –0.93, whereas in the OVA scheme the correlation is –0.72. In both cases, the general picture is that of α^* decreasing as K increases.

However, on a closer examination, there is a marked difference in the way α^* changes with regard to K between the SMA and the OVA versions of the DDP-based feature selection technique (Figure 5). In the SMA, α^* decreases in accordance with every step of increase in K. In the OVA scheme, α^* stays near the range of equal-priorities predictor set scoring method (0.5 and 0.6) for the four datasets with larger K (the GCM, NCI60, PDL and lung datasets). Then, in the region of datasets with smaller K, α^* in the OVA scheme increases so that it is nearer the range of rank-based feature selection technique (0.8 and 0.9 for the SRBC, MLL and AML/ALL datasets).

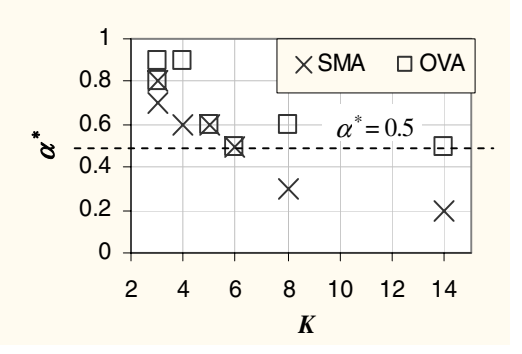

Fig. 5. Optimal value of DDP, α^*, plotted against K for all benchmark datasets

The steeper decrease of α^* as K increases in the SMA implies that the measure of relevance used in the SMA fails to capture the relevance of a feature when K is large. In the OVA scheme, the decrease of α^* as K increases is more gradual, implying better effectiveness than the SMA in capturing relevance for datasets with larger K.

Furthermore, for **all** datasets, the value of α^* in the OVA scheme is greater than or equal to the value of α^* in the SMA. Unlike in the SMA, the values of α^* in the OVA scheme never fall below 0.5 for all benchmark datasets (Figure 5). This means that the measure of relevance implemented in the OVA scheme is more effective at identifying relevant features, regardless of the value of K. In other words, K different groups of features, each considered highly relevant based on a different binary target class concept, $\mathbf{y_k}$ ($k = 1,2,...,K$), are more capable of distinguishing among samples of K different classes than a single group of features deemed highly relevant based on the K-class target class concept, $\mathbf{y}$.

Since in none of the datasets has α^* reached exactly 1, antiredundancy is still a factor that should be considered in the predictor set scoring method. This is true for both the OVA scheme and the SMA. Redundancy leads to unnecessary increase in classifier complexity and noise. However, for a given dataset, when the optimal DDP leans closer towards maximizing relevance in one case (Case 1) than in another case (Case 2), it is usually an indication that the approach used in measuring relevance in Case 1

is *more effective* than the approach used in Case 2 at identifying truly relevant features. In this particular study, Case 1 represents the OVA version of the DDP-based feature selection technique, and Case 2, the SMA version.

7 Conclusions

Based on one or more of the following criteria: class accuracy, best averaged accuracy and size-averaged accuracy, the OVA version of the DDP-based feature selection technique outperforms the SMA version. Despite the increase in computational cost and predictor set size by a factor of K, the improvement brought by the OVA scheme in terms of overall accuracy and class accuracy is especially significant for the datasets with the largest number of classes and highest level of complexity and difficulty, such as the GCM and NCI60 datasets. Furthermore, the OVA scheme brings the degree of differential prioritization closer to relevance for most of the benchmark datasets, implying better efficiency in the OVA approach at measuring relevance than the SMA.

References

1. Slonim, D.K., Tamayo, P., Mesirov, J.P., Golub, T.R., Lander, E.S.: Class prediction and discovery using gene expression data. In: RECOMB 2000 (2000) 263–272
2. Garber, M.E., Troyanskaya, O.G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G.D., Perou, C.M. , Whyte, R.I., Altman, R.B., Brown, P.O., Botstein, D., Petersen, I.: Diversity of gene expression in adenocarcinoma of the lung. Proc. Natl. Acad. Sci. 98(24) (2001) 13784–13789
3. Schena, M., Shalon, D., Davis, R.W., Brown, P.O.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270 (1995) 467–470
4. Shalon, D., Smith, S.J., Brown, P.O.: A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Research 6(7) (1996) 639–645
5. Yu, L., Liu, H.: Redundancy Based Feature Selection for Microarray Data. In: Proc. 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2004) 737–742
6. Li, L., Weinberg, C.R., Darden, T.A., Pedersen, L.G.: Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 17 (2001) 1131–1142
7. Xing, E., Jordan, M., Karp, R.: Feature selection for high-dimensional genomic microarray data. In: Proc. 18th International Conference on Machine Learning (2001) 601–608
8. Ooi, C.H., Chetty, M., Teng, S.W.: Relevance, redundancy and differential prioritization in feature selection for multiclass gene expression data. In: Oliveira, J.L., Maojo, V., Martín-Sánchez, F., and Pereira, A.S. (Eds.): Proc. 6th International Symposium on Biological and Medical Data Analysis (ISBMDA-05) (2005) 367–378
9. Platt, J.C., Cristianini, N., Shawe-Taylor, J.: Large margin DAGs for multiclass classification. Advances in Neural Information Processing Systems 12 (2000) 547–553
10. Mitchell, T.: Machine Learning, McGraw-Hill, 1997
11. Hsu, C.W., Lin, C.J.: A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks 13(2) (2002) 415–425

12. Li, T., Zhang, C., Ogihara, M.: A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 20 (2004) 2429–2437
13. Chai, H., Domeniconi, C.: An evaluation of gene selection methods for multi-class microarray data classification. In: Proc. 2nd European Workshop on Data Mining and Text Mining in Bioinformatics (2004) 3–10
14. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. In: Proc. 2nd IEEE Computational Systems Bioinformatics Conference. IEEE Computer Society (2003) 523–529
15. Dudoit, S., Fridlyand, J., Speed, T.: Comparison of discrimination methods for the classification of tumors using gene expression data. JASA 97 (2002) 77–87
16. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., Golub, T.R.: Multi-class cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. 98 (2001) 15149–15154
17. Linder, R., Dew, D., Sudhoff, H., Theegarten D., Remberger, K., Poppl, S.J., Wagner, M.: The 'subsequent artificial neural network' (SANN) approach might bring more classificatory power to ANN-based DNA microarray analyses. Bioinformatics 20 (2004) 3544–3552
18. Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C.F., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D., Brown, P.O.: Systematic variation in gene expression patterns in human cancer cell lines, Nature Genetics 24(3) (2000) 227–234
19. Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J., Meyerson, M.: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. 98 (2001) 13790–13795
20. Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., Korsmeyer, S.J.: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics 30 (2002) 41–47
21. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286 (1999) 531–537
22. Yeoh, E.-J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Relling, M.V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.-H., Evans, W.E., Naeve, C., Wong, L., Downing, J. R.: Classification, subtype discovery, and prediction of outcome in pediatric lymphoblastic leukemia by gene expression profiling. Cancer Cell 1 (2002) 133–143
23. Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S.: Classification and diagnostic prediction of cancers using expression profiling and artificial neural networks. Nature Medicine 7 (2001) 673–679
24. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3 (2003) 1157–1182