Investigating the Class-Specific Relevance of Predictor Sets Obtained from DDP-Based Feature Selection Technique

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. Feature selection is crucial to tumor classification due to the high dimensionality of microarray datasets. With the aid of the degree of differential prioritization (DDP) between relevance and antiredundancy, our proposed DDP-based feature selection technique is capable of achieving better accuracies than those reported in previous studies, while using fewer genes in the predictor set. Additionally, we discovered a strong correlation between the DDP parameter in our feature selection technique and the number of classes in the dataset. This leads us to question if the measure of relevance in our feature selection technique becomes less efficient at capturing the class-specific relevance for each individual class of the dataset as the number of classes increases. In this study, we analyze the class-specific relevance of the predictor sets found using our feature selection technique. The analysis ultimately lays down the theoretical foundation for a beneficial improvement to our feature selection technique.

1 Introduction

The objective of feature selection is to form a subset of features, which would yield the optimal estimate of classification accuracy. This subset of features is termed the predictor set. The three levels of filter-based feature selection for tumor classification can be summarized as follows: 1) no selection, 2) based on relevance alone, and finally, 3) based on relevance and redundancy. Relevance measures the ability of the predictor set to distinguish among samples of different classes. Redundancy indicates the amount of similarity among the members of the predictor set.

Previous studies [1, 2] have based their filter-based feature selection techniques on the concept of relevance and redundancy having equal role in the formation of a good predictor set. On the other hand, using a simple 2-class problem, it has been demonstrated that seemingly redundant features may improve the discriminant power of the predictor set instead [3], although it remains to be seen if this also applies in cases of multiclass domains.

In [4], we introduced a third element: the relative importance placed between relevance vs. redundancy. We call this element the degree of differential prior-

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 60 – 70, 2006.

itization (DDP). DDP compels the search method to prioritize the optimization of one of the two elements (relevance or redundancy) at the cost of the optimization of the other.

The effectiveness of our DDP-based feature selection technique on the tumor classification of five multiclass microarray datasets has been reported in [4]. However, after adding two datasets into the collection of benchmark datasets, we observe indubitable correlations between several aspects of classification performance against the number of classes of the dataset – which bring on two hypotheses regarding the limitations of the DDP-based feature selection technique. To prove these hypotheses, an in-depth analysis of the predictor sets produced by the DDP-based feature selection technique is conducted.

The contributions of this paper are twofold. Firstly, we identify the source of the limitations of the DDP-based feature selection technique. Secondly, the weakness of the widely-used multiclass target class concept is depicted with the support of results from experiments on the benchmark datasets and cemented with the aid of the optimal value of the DDP for each of these datasets. At the same time, we present the case for using class-specific relevance scores in place of the all-classes-at-once relevance score in order to improve the feature selection technique.

We will begin with a brief description of the DDP-based feature selection technique, followed by a summary of the results, which leads to the hypotheses regarding the limitations of the DDP-based feature selection technique. From there, we present the analysis on the class-specific relevance of the predictor sets, discuss the results of the analysis, and state our conclusion.

2 DDP-Based Feature Selection Technique

The training set, T, consists of N genes and M_t training samples. Sample j is represented by vector $\mathbf{x}_j$ containing the expression of the N genes $[x_{1,j},\ldots, x_{N,j}]^{\mathrm{T}}$ and scalar y_j representing the class the sample belongs to. For microarray datasets, the term *gene* and *feature* may be used interchangeably. The 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 P genes, called the predictor set S, which would give the optimal classification accuracy.

The DDP-based predictor set scoring method measures the goodness of predictor set S as follows.

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

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

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 S [2]:

$$V_S = \frac{1}{|S|} \sum_{i \in S} F(i) \ . \tag{2}$$

The individual relevance score, $F(i)$, indicates the correlation of gene i to the target class concept **y**, i.e., the 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 (**b**etween-groups **s**um of **s**quares/**w**ithin-groups **s**um of **s**quares) ratio used in [2, 5]. For gene i,

$$F(i) = \frac{\sum_{j=1}^{M_t}\sum_{k=1}^{K} I(y_j = k)(\bar{x}_{ik} - \bar{x}_{i\bullet})^2}{\sum_{j=1}^{M_t}\sum_{k=1}^{K} I(y_j = k)(x_{ij} - \bar{x}_{ik})^2} \tag{3}$$

where $I(.)$ 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, while $\bar{x}_{ik}$ is the average of the expression of gene i across training samples belonging to class k. The BSS/WSS ratio, first used in [5] for multiclass tumor classification, is a modification of the F-ratio statistics for one-way ANOVA (Analysis of Variance).

U_S is the measure of antiredundancy for the candidate predictor set S.

$$U_S = \frac{1}{|S|^2}\sum_{i,j\in S, i\neq j} 1 - |R(i,j)| \ . \tag{4}$$

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

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.

For predictor set search, the linear incremental search method [2] is used, where the first member of S is chosen by selecting the gene with the highest $F(i)$ score. To find the second and the subsequent members of the predictor set, the remaining genes are screened one by one for the gene that would give the maximum $W_{A,S}$. The procedure is terminated when P has reached $P_{\max}$ (arbitrarily user-determined). This search method has a computational complexity of $O(NP_{\max})$ [2].

2.1 Experiment Settings

Different values of α from 0.1 to 1 were tested with equal intervals of 0.1. The F-splits procedure is employed in evaluating the performance of a predictor set of a certain size P and produced using a certain value of α. Using F different splits of training and test sets, the accuracies from all splits are averaged to give an indication of the performance of the predictor set. Optimal predictor sets ranging from sizes P=2,3,…,$P_{\max}$ were formed in each split. $P_{\max}$ and F are set to 100 and 10 respectively in this study. The microarray datasets used as benchmark datasets: GCM [6], NCI60 [7], lung [8], MLL [9], AML/ALL [10], PDL [11] and SRBC [12]

datasets, are described in Table 1. Datasets are preprocessed and normalized based on the recommended procedures in [5] for Affymetrix and cDNA microarray data.

Except for the GCM dataset, where the original ratio of training to test set size [6] is maintained to enable comparison with previous studies, for all datasets we use the standard 2:1 split ratio. DAGSVM [13], an all-pairs SVM-based multiclassifier, is used for performance evaluation.

Table 1. Descriptions of benchmark datasets

Dataset	Type	*N*	*K*	Training:test set
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

2.2 Size-Averaged Accuracy and α^*

For all predictor sets found using a particular value of α, we plot the estimate of accuracy obtained from the 10-splits procedure against the value of P of the corresponding predictor set (Figure 1a). The size-averaged accuracy for that value of α is the area under the curve in Figure 1a divided by the number of predictor sets, $(P_{max}-1)$.

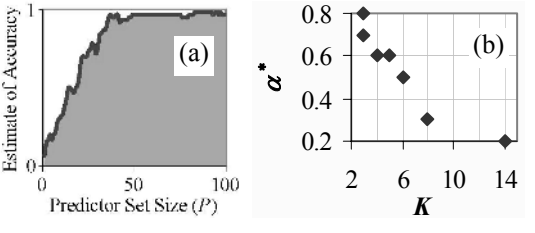

Fig. 1. a) Area under the accuracy-predictor set size curve; b) α^* vs. K

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.

Figure 1b shows the value of α^* from each dataset plotted against the value of K from corresponding datasets. We observe that as K increases, placing **more emphasis on maximizing antiredundancy** produces better accuracy than placing more emphasis on relevance. Maximizing antiredundancy becomes less important as K decreases – therefore supporting the assertion in [3] that redundancy does *not* hinder the discriminant power of the predictor set when K is 2. Conversely, in order to get the best size-averaged accuracy as K increases, the predictor set scoring method has to **decrease** the emphasis on relevance.

Figure 2a indicates that overall accuracy (represented by the highest size-averaged accuracy) deteriorates as *K* increases.

2.3 Class Accuracy

This is computed in the same way as the size-averaged accuracy, except that we compute the class-specific accuracy for each class of the dataset. Hence there are a total of *K* class accuracies for a *K*-class dataset. For the sake of brevity, the plot for individual class accuracy from each class is not shown, but the trend is the same as shown by size-averaged accuracy in Figure 2a. The class accuracy of majority of classes in the dataset decreases as *K* increases.

To measure the balance among classes in terms of class accuracy, the range of class accuracy is used. First, we compute the class accuracy for each class using the optimal DDP value of the dataset, α^*. Then, the range of class accuracy is computed by subtracting the worst class accuracy from the best class accuracy among the classes in the dataset. The plot of the range of class accuracy against the value of *K* of the corresponding datasets (Figure 2b) shows that as *K* increases, the *imbalance* among classes in terms of class accuracy also increases.

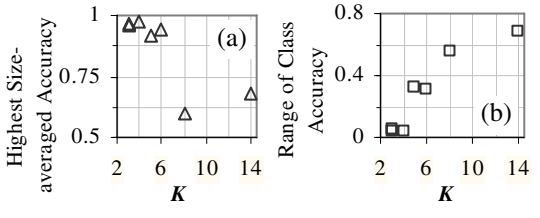

Fig. 2. a) Highest size-averaged accuracy and b) range of class accuracy vs. *K*

Therefore, the greater difference in class accuracy among classes and generally smaller class accuracy for majority of classes as *K* increases causes the deterioration of overall accuracy as *K* increases.

3 Analyzing the Predictor Sets

Two hypotheses are formed based on the results presented in the previous section:

1. Based on the results in Section 2.2: The measure of relevance computed based on **y** is not efficient enough to capture the class-specific relevance of a feature when *K* is larger than 6.
2. Based on the results in Sections 2.2 and 2.3: For each class of the dataset, class-specific relevance is essential to achieving good class accuracy, which, along with the class accuracies from other classes, contributes to the overall accuracy.

In this section we present the analyses implemented in order to prove both hypotheses.

3.1 OVA (One-vs.-All)-Based Relevance of the Predictor Sets

The target class concept currently used in the DDP-based feature selection technique, **y**, is multiclass; there is only one target class concept for all K classes. The measure of relevance computed based on **y** is also termed *all-classes-at-once* relevance. OVA (**o**ne-**v**s.-**a**ll)-based target class concept, on the other hand, is binary; there is one target class concept for each of the K classes in the dataset, making for a total of K binary target class concepts. Therefore, the relevance computed based on one of these binary target class concepts is called *class-specific relevance*.

For class k (k = 1, 2, …, K), the OVA-based target class concept is represented by

$$\mathbf{y_k} = \begin{bmatrix} y_{k,1} & y_{k,2} & \cdots & y_{k,M_t} \end{bmatrix}, \tag{5}$$

where

$$y_{k,j} = \begin{cases} 1 & \text{if} \quad y_j = k \\ 2 & \text{if} \quad y_j \neq k \end{cases}. \tag{6}$$

The score of relevance of gene i for class k, $F(i,k)$, is given as follows.

$$F(i,k) = \frac{\sum_{j=1}^{M_t} \sum_{q=1}^{2} I(y_{k,j} = q)(\bar{x}_{iq} - \bar{x}_{i\bullet})^2}{\sum_{j=1}^{M_t} \sum_{q=1}^{2} I(y_{k,j} = q)(x_{ij} - \bar{x}_{iq})^2}. \tag{7}$$

$I(.)$ returns 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, the value $\bar{x}_{iq}$ is the average of the expression of gene i across training samples *not* belonging to class k.

Figure 3 shows the procedure involved in analyzing the OVA-based relevance of predictor sets of size 100 obtained from different values of α. $TC_{k,S(100,\alpha)}$ is the split-averaged OVA-based relevance of the predictor set of size 100 found using the differential prioritization factor α for class k. We arbitrarily use the size of 100 because this is the value of $P_{\max}$ used in the evaluating our feature selection technique on the datasets. Fixing the predictor set size and the number of top features on the $F(i,k)$ ranking to be counted in the predictor set at any other number does not significantly change the conclusions of the analysis.

For each class in a dataset, we average the values of $TC_{k,S(100,\alpha)}$ across the tested range of α. Let us denote this average as $TC_{k,S(100,\text{avg})}$. To succinctly analyze the behavior of $TC_{k,S(100,\alpha)}$ across classes and different values of α, we plot three parameters against corresponding value of α (Figure 4).

1. The $TC_{k,S(100,\alpha)}$ values from the class which produces the maximum $TC_{k,S(100,\text{avg})}$
2. The $TC_{k,S(100,\alpha)}$ values from the class which produces the minimum $TC_{k,S(100,\text{avg})}$
3. The $TC_{k,S(100,\alpha)}$ values from the class which produces the median $TC_{k,S(100,\text{avg})}$

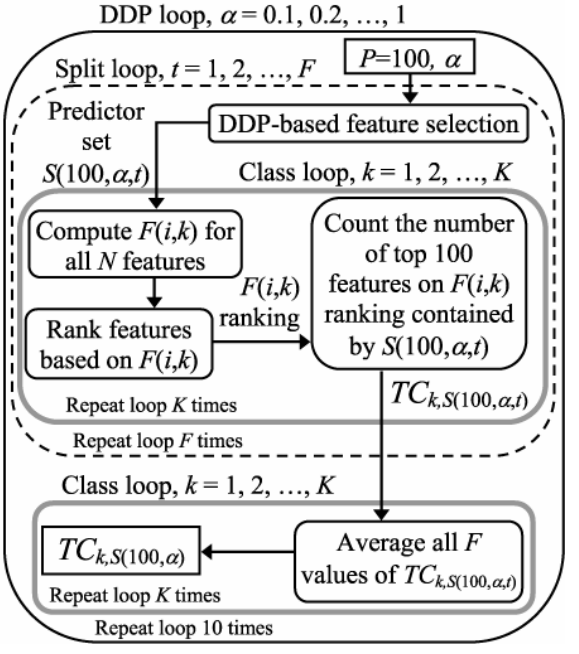

Fig. 3. Analyzing OVA-based relevance

among all the classes in the dataset. In case of even K, we take the average of the $TC_{k,S(100,\alpha)}$ values from the two classes which produce the median $TC_{k,S(100,\text{avg})}$.

There are two important observations to be made from Figure 4. Firstly, in an ideal situation, the difference between the largest and smallest $TC_{k,S(100,\alpha)}$ across classes should be 0. However, even at the smallest tested value of α, difference between largest and smallest $TC_{k,S(100,\alpha)}$ is greater than 8 for all datasets. For five of the datasets, the median-$TC_{k,S(100,\text{avg})}$ plot lies closer to the minimum-$TC_{k,S(100,\text{avg})}$ plot than the maximum-$TC_{k,S(100,\text{avg})}$ plot. For the remaining two datasets, the median-$TC_{k,S(100,\text{avg})}$ plot is halfway between the two other plots. Secondly, $TC_{k,S(100,\alpha)}$ is often larger for datasets with smaller K than datasets with larger K.

The first observation indicates that the OVA-based relevance of predictor sets found using the DDP-based feature selection technique is imbalanced among different classes for each benchmark dataset, especially in predictor sets where the emphasis on the all-classes-at-once relevance are large (α greater than 0.7). It also bears out the claim that the all-classes-at-once relevance score does not capture the OVA-based relevance well for **majority** of the classes in each benchmark dataset. The second observation tells us that the all-classes-at-once relevance score becomes less efficient at capturing the OVA-based relevance for all classes as the number of classes in the dataset increases. These two observations support the first hypothesis stated earlier in this section.

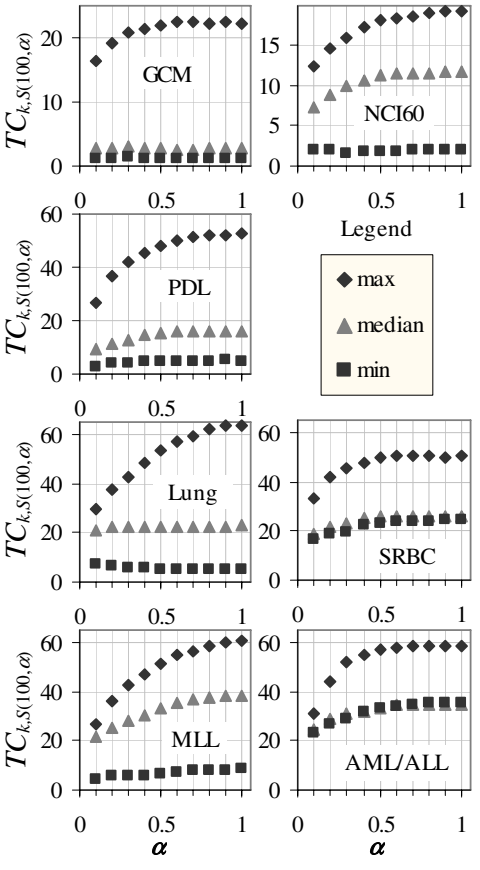

Fig. 4. Statistics from $TC_{k,S(100,\alpha)}$ plotted against corresponding value of α

3.2 OVA-Based Relevance and Class Accuracy

For each value of α and each class of a dataset, we plot the class accuracy obtained from that value of α against corresponding value of $TC_{k,S(100,\alpha)}$ (Figure 5). At first, class accuracy climbs rapidly as $TC_{k,S(100,\alpha)}$ increases. However, when $TC_{k,S(100,\alpha)}$ reaches a certain 'limiting' value, class accuracy begins to stagnate or even deteriorate in some datasets (e.g., NCI60, MLL and AML/ALL).

The stagnation or deterioration in class accuracy with respect to $TC_{k,S(100,\alpha)}$ can be explained by the fact that as the number of features with high $F(i,k)$ scores increases, the likelihood of these features being strongly correlated to each other also increases. This leads to lower level of antiredundancy in the predictor set, and hence, the resulting plateau/degeneration in class accuracy.

The correlation between OVA-based relevance (as represented by $TC_{k,S(100,\alpha)}$) and corresponding class accuracy indicates that to a certain extent (as limited by redundancy in the predictor set), large $TC_{k,S(100,\alpha)}$ is beneficial for the class accuracy of the corresponding class. This confirms the second hypothesis stated earlier in this section.

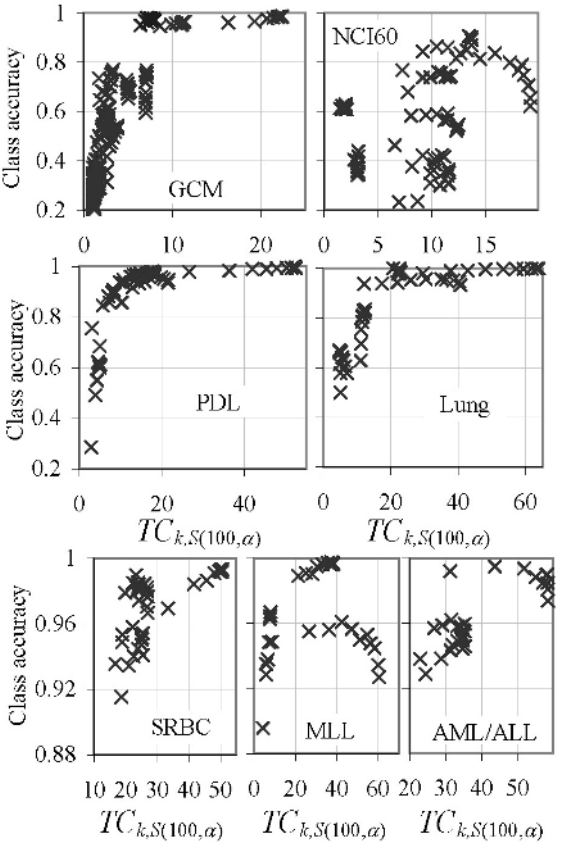

Fig. 5. Class accuracy plotted against corresponding value of $TC_{k,S(100,\alpha)}$

Proving the two hypotheses gives us indications on the areas we need to improve on in the DDP-based feature selection technique. In order to eliminate or at least alleviate the adverse effect of increasing K on overall accuracy, we need to fine-tune the DDP-based feature selection technique so that it becomes more capable of finding predictor sets which contain high level of $TC_{k,S(100,\alpha)}$ (OVA-based relevance) for all classes in a dataset, while at the same time having minimal difference between the largest and the smallest $TC_{k,S(100,\alpha)}$ across all classes.

Therefore, instead of using **y** (as in the original, all-classes-at-once version of the DDP-based feature selection technique), the score of relevance for gene i should be computed using the class-specific relevance, $F(i,k)$, which is defined in equation (7). The use of class-specific relevance leads to the conception of the OVA version of the DDP-based feature selection technique [14].

The OVA version of the DDP-based feature selection technique should produce some improvement in the estimates of overall accuracy and class accuracy for datasets with relatively large K, such as the GCM and NCI60 datasets. Improvement is also expected for datasets with smaller K but not as significant as that for the datasets with larger K (>6). With the current technique, the accuracy for the larger-K

datasets is much lower than the accuracy for other datasets with smaller K (Figure 2a). Hence, the improvement expected from the modified feature selection technique will certainly be highly beneficial.

Tentative results from the OVA version of the DDP-based feature selection technique compared against the original (all-classes-at-once) DDP-based feature selection technique are available in [14]. These results confirm the predictions in the previous paragraph.

4 Conclusions

With the help of the DDP parameter of α, we have identified the source of the limitations of the DDP-based feature selection technique. The flaw in the widely-used multiclass target class concept has been proven both empirically through the feature selection experiments on the benchmark datasets and conceptually through the analysis of the OVA-based relevance of predictor sets found using our feature selection technique. The case for using class-specific relevance scores in place of the all-classes-at-once relevance score in order to improve the technique has also been proven. By fine-tuning the existing technique, the accuracy will be significantly improved for multiclass microarray datasets, especially datasets with large number of classes.

References

1. Hall, M.A., Smith, L.A.: Practical feature subset selection for machine learning. In: McDonald, C. (ed.): Proc. 21st Australasian Computer Science Conference. Springer, Singapore (1998) 181–191
2. 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
3. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3 (2003) 1157–1182
4. 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
5. Dudoit, S., Fridlyand, J., Speed, T.: Comparison of discrimination methods for the classification of tumors using gene expression data. JASA 97 (2002) 77–87
6. 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
7. 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

8. 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
9. 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
10. 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
11. 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
12. 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
13. Platt, J.C., Cristianini, N., Shawe-Taylor, J.: Large margin DAGs for multiclass classification. Advances in Neural Information Processing Systems 12 (2000) 547–553
14. Ooi, C.H., Chetty, M., Teng, S.W.: OVA Scheme vs. Single Machine Approach in Feature Selection for Microarray Datasets. In: Perner, P. (Ed): Proc. 6th Industrial Conference on Data Mining (ICDM2006) (2006), in press