A Study on the Importance of Differential Prioritization in Feature Selection Using Toy Datasets

Chia Huey Ooi, Shyh Wei Teng, and Madhu Chetty

Faculty of Information Technology
Monash University, Australia
chiahuey.ooi@gms.edu.sg,
{shyh.wei.teng,madhu.chetty}@infotech.monash.edu.au

Abstract. Previous empirical works have shown the effectiveness of differential prioritization in feature selection prior to molecular classification. We now propose to determine the theoretical basis for the concept of differential prioritization through mathematical analyses of the characteristics of predictor sets found using different values of the DDP (degree of differential prioritization) from realistic toy datasets. Mathematical analyses based on analytical measures such as distance between classes are implemented on these predictor sets. We demonstrate that the optimal value of the DDP is capable of forming a predictor set which consists of classes of features which are well separated and are highly correlated to the target classes – a characteristic of a truly optimal predictor set. From these analyses, the necessity of adjusting the DDP based on the dataset of interest is confirmed in a mathematical manner, indicating that the DDP-based feature selection technique is superior to both simplistic rank-based selection and state-of-the-art equal-priorities scoring methods. Applying similar analyses to real-life multiclass microarray datasets, we obtain further proof of the theoretical significance of the DDP for practical applications.

1 Introduction

The aim of feature selection is to form, from all available features in a dataset, a relatively small subset of features capable of producing the optimal classification accuracy. This subset is called the predictor set. A feature selection technique is made up of two components: the predictor set scoring method (which evaluates the goodness of a candidate predictor set); and the search method (which searches the gene subset space for the predictor set based on the scoring method). This study focuses on filter-based technique which its classifiers are not invoked in the predictor set scoring method.

An important principle behind most filter-based feature selection studies can be summarized by the following statement: A good predictor set should contain features highly correlated to the target class concept, and yet uncorrelated with each other [1]. The predictor set attribute referred to in the first part of this statement, 'relevance', is the backbone of simple rank-based feature selection techniques. The aspect alluded to in the second part, 'redundancy', refers to pairwise relationships between all pairs of features in the predictor set.

M. Chetty, A. Ngom, and S. Ahmad (Eds.): PRIB 2008, LNBI 5265, pp. 311–322, 2008.

Previous studies [1, 2] have based their feature selection techniques on the concept of relevance and redundancy having equal importance in the formation of a good predictor set. We call the predictor set scoring methods used in such correlation-based feature selection techniques *equal-priorities scoring methods*. On the other hand, it is demonstrated in another study [3] using a two-class problem that seemingly redundant features may improve the discriminant power of the predictor set instead, although it remains to be seen how this scales up to multiclass domains with thousands of features. A previous study was implemented on the effect of varying the importance of redundancy in predictor set evaluation [4]. However, due to its use of a relevance score that is inapplicable to multiclass problems, the study was limited to only two-class classification.

Currently, when it comes to the use of filter-based feature selection for multiclass tumor classification, three popular recommendations are: 1) no selection [5, 6]; 2) select based on relevance alone [5, 7]; and finally, 3) select based on relevance and redundancy [2, 8]. Thus, so far, relevance and redundancy are the two existing criteria which have ever been used in predictor set scoring methods for multiclass tumor classification.

To these two criteria we introduce a third criterion: the relative importance placed between relevance and redundancy [9]. We call this criterion the degree of differential prioritization (DDP). DDP compels the search method to prioritize the optimization of one of the two criteria (of relevance or redundancy) at the cost of the optimization of the other. Unlike other existing correlation-based techniques, the DDP-based feature selection technique does not take for granted that the optimizations of both elements of relevance and redundancy are to have equal priorities in the search for the predictor set [10].

Although a large body of work has provided empirical support regarding the efficacy of the DDP concept in feature selection [9-11] , we have yet to establish the theoretical strengths and merits of the DDP-based feature selection technique. This is precisely the aim of this paper, which is to be realized through vigorous mathematical analyses of predictor sets found using the DDP-based feature selection technique and simple but illustrative examples using toy datasets.

To generate toy datasets for this purpose, we employ a model which is well-known and recognized not only in the domains of molecular classification and microarray analysis but also conventional data minings. Later in this paper, we also show how close conditions in real-life multiclass microarray datasets resemble those of our toy datasets. Additional advantages of toy datasets include the unlimited number of datasets we can generate [vs. the limited number of available real-life microarray datasets]; the control we are able to exercise over dataset characteristics such as the number of classes and features [11]; and prior knowledge of the members of the ideal predictor set, which provides the ultimate means for measuring the efficacy of the feature selection technique without involving the inductions of actual classifiers.

The organization of the paper is as follows: Beginning with a description of the DDP-based feature selection technique, we proceed to present a model for producing the toy datasets. Then, we analyze the class separation property of the predictor sets obtained from each of the toy datasets. We then apply the same analysis to eight real-life multiclass microarray datasets. Finally, we present the conclusions of the study.

2 Differential Prioritization

For gene expression datasets, the terms *gene* and *feature* may be used interchangeably. From the total of N genes, the objective is to form the subset of genes, called the predictor set S, which gives the optimal classification accuracy.

The score of goodness for predictor set S is given as follows.

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

V_S represents the relevance of S. V_S measures the average of the correlation of the members of S to the target class concept.

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

The target class concept is represented by the target class vector **y**, which is defined as $[y_1, y_2, \ldots, y_{|T|}]$, $y_j \in [1, K]$ in a K-class dataset. y_j is the class label of sample j. The training set, T, consists of all training samples of K classes. Based on **y**, the relevance of gene i is computed as follows.

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

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 in T. $\bar{x}_{i,k}$ is the average of the expression of gene i across training samples belonging to class k. $x_{i,j}$ is the expression of gene i in sample j. $F(i)$ is the BSS/WSS (between-groups sum of squares/within-groups sum of squares) ratio for gene i [12]. It indicates the ability of the gene in discriminating among samples belonging to K different classes.

U_S represents the antiredundancy of S. Antiredundancy is a measure opposite to redundancy in quality [9].

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

The absolute value of the Pearson product moment correlation coefficient between genes i and j, $|R(i,j)|$, is used to measure the correlation between genes i and j.

The power factor $\alpha \in (0, 1]$ in Eq. 1 denotes the DDP 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 (and at the same time decreases the emphasis on maximizing antiredundancy) during the search for the predictor set [10].

A predictor set found using a 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 a 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. At $\alpha = 0.5$, we get an equal-priorities scoring method. At $\alpha = 1$, the feature selection technique becomes rank-based.

We posit that different datasets will require different values of the DDP between maximizing relevance and maximizing antiredundancy in order to come up with the most efficacious predictor set. Therefore the optimal range of α (leading to the predictor set giving the best accuracy) is dataset-specific.

The linear incremental search is conducted as follows: 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 $W_{A,S}$. Since the combination of our predictor set scoring method and this search method does not specify an output as to the final size of the predictor set to be used, the maximum size of the predictor set, P, will have to be predetermined by the user.

3 Toy Datasets Based on One-vs.-All (OVA) Model

In using toy datasets, we aim to provide simple but clear and demonstrative examples which highlight the importance of choosing the correct value of the DDP in forming the best predictor set. Furthermore, another advantage of toy datasets is the fact that we know exactly just how large a predictor set should be found for each case, facilitating the task of choosing the value of P.

It is widely accepted that over-expression or under-expression (suppression) of genes causes the difference in phenotype among samples of different classes. The categorization of gene expression is given as follows.

1. A gene is over-expressed: if its expression value is above baseline.
2. A gene is under-expressed: if its expression value is below baseline.
3. Baseline interval: the normal range of expression value.

Usually the mean of the expression across genes is taken as the middle of the baseline interval. In analyses of microarray data, the conventional data normalization procedures often set the mean across genes for each sample to zero. Hence, over-expression is represented by positive values and under-expression by negative values. With this categorization we next employ a well-known paradigm leading to the one-vs.-all (OVA) model, which is then used to generate toy datasets.

The crux of the OVA concept has gained wide, albeit tacit, acceptance among microarray and tumor gene expression researchers. The fact that particular genes are only over-expressed in tissues of certain type of cancer, and not any other types of cancer or normal tissues [13], is part of the entrenched domain knowledge. Hence the

term 'marker' – for genes that mark the particular cancer they are associated with. In the OVA model, certain groups of genes, also called the 'marker genes', are only over-expressed (or under-expressed) in samples belonging to a particular class and never in all samples of other classes. This model emphasizes that a group of marker genes is specific to one class. Therefore for a K-class dataset, there are K different groups of marker genes.

Let us denote as G the number of genes in each group of marker genes, $X_{\max}$ and $X_{\min}$ the maximum and minimum limits respectively to the absolute value of the class means for the whole dataset. Thus, for the g-th gene in a group of marker genes, the maximum limit to the absolute value of the class means is defined as:

$$x_{\max,g} = X_{\max} - (\Delta X)(g-1) \tag{5}$$

where $g = 1, 2, \ldots, G$, and

$$\Delta X = \frac{X_{\max} - X_{\min}}{G-1} \tag{6}$$

Among the K classes, the class means are made to vary such that there is an imbalance in terms of class means. The reasons are firstly to mimic a condition prevalent in multiclass microarray datasets (imbalance among classes in terms of class means even after normalization), especially in datasets with very large number of classes; and secondly, to present a challenge to the feature selection technique in choosing relevant but non-redundant genes. We will provide further elucidation on the second reason later in this section. For the g-th gene in a group of marker genes, the difference between the class means of subsequent classes is defined in the following manner.

$$\Delta x_g = \frac{2x_{\max,g}}{K-1} \tag{7}$$

Next, initialize a matrix $M:=(\mu_{i,k})_{N\times K}$ of zeros where N is the total number of genes in the dataset, and, in this case, is the product of G and K. This is the matrix of class means, whose element, $\mu_{i,k}$, represents the mean of gene i across samples belonging to class k ($k = 1, 2, \ldots, K$).

$$\mu_{(g-1)K+k,k} = \left(-1^g\right)\left[x_{\max,g} - (\Delta x_g)(k-1)\right] \tag{8}$$

The $[(g-1)K+k]$-th gene is the g-th member of the k-th group of marker genes and therefore has non-zero class mean for class k and zero class mean for all other classes – the archetypal OVA trait. The term $\left(-1^g\right)$ serves to change the sign of the class mean at different values of g so as to produce both over- and under-expressed marker genes. The strongest marker genes are composed of the first genes ($g = 1$) of each group of marker genes while the weakest marker genes consist of the last genes ($g = G$) of each group of marker genes.

Standard deviation among samples of the same class, or class standard deviation, is set to 1 for all instances, $\sigma_{i,k} = 1$ for all k and i. To produce a K-class toy dataset, a

Table 1. A 4-class example from the OVA model (μi,k represents the mean of gene i across samples belonging to class k)

g	k	$\mu_{i,k}$	$\mu_{i,1}$	$\mu_{i,2}$	$\mu_{i,3}$	$\mu_{i,4}$
1	1	$\mu_{1,k}$	$-X_{\max}$	0	0	0
1	2	$\mu_{2,k}$	0	$-0.5X_{\max}$	0	0
1	3	$\mu_{3,k}$	0	0	$0.5X_{\max}$	0
1	4	$\mu_{4,k}$	0	0	0	$X_{\max}$
2	1	$\mu_{5,k}$	$X_{\max}-\Delta X$	0	0	0
2	2	$\mu_{6,k}$	0	$0.5(X_{\max}-\Delta X)$	0	0
2	3	$\mu_{7,k}$	0	0	$0.5(\Delta X-X_{\max})$	0
2	4	$\mu_{8,k}$	0	0	0	$\Delta X-X_{\max}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
G	1	$\mu_{(G-1)K+1,k}$	$(-1^G)X_{\min}$	0	0	0
G	2	$\mu_{(G-1)K+2,k}$	0	$0.5(-1^G)\ X_{\min}$	0	0
G	3	$\mu_{(G-1)K+3,k}$	0	0	$-0.5(-1^G)\ X_{\min}$	0
G	4	$\mu_{(G-1)K+4,k}$	0	0	0	$-(-1^G)X_{\min}$

total of m samples are generated for class k (k = 1, 2, …, K) using Gaussian distribution of mean $\mu_{i,k}$ and standard deviation $\sigma_{i,k}$ for gene i.

In Table 1, an entry on the i-th row and k-th column represents the class mean of class k for gene i, where $i = \lfloor (g-1)K + k \rfloor$, and therefore gene i is the g-th member of the k-th group of marker genes. We can see that using relevance alone as a criterion, and with uniform class size, marker genes associated with class 1 and 4 will always be favored more than marker genes specific to any other classes, regardless of the value of g. Including antiredundancy as the second criterion will obviate this imbalanced predilection – therein lies the reason for us to use unequal values for class means among different classes. But how much weight is to be assigned to relevance, and how much to antiredundancy?

The apparent answer would be equal weights, which is the foundation of existing equal-priorities scoring methods. But as mentioned previously in Section 1, it has been shown that antiredundancy is not as important as relevance for the two-class case – this is obvious in the case of our OVA toy dataset; any subset of sufficiently relevant genes will be capable of differentiating between the two classes. Hence as the number of classes increases (an important theme in multiclass classification studies), will not the importance of antiredundancy (w.r.t. relevance) increase as well? This question is to be answered from the analyses in this study.

4 Experiment Settings

Ten values of α are tested from 0.1 to 1 with equal intervals of 0.1. G is set to 3, 5, 10, 20, and 30. $X_{\max}$ and $X_{\min}$ is set to 100 and 1 respectively, while the number of samples per class, or class size, m, is set to 100 uniformly for all classes. We test for $K = 2$ to $K = 30$. Since no inductions of classifiers are to be implemented in this study, whole datasets are used as training sets during feature selection.

The minimum predictor set size necessary to differentiate among the K classes is $K-1$. The optimal predictor set is actually any subset of $K-1$ genes from the first K of the marker genes (that is, at $g=1$) generated using the class means defined in Eq. 8. Thus, P is set to $K-1$.

5 Analyses of Predictor Sets

In this section we critically analyze the properties of the predictor sets produced from different values of α. The property of separation of classes is studied. Finally, we apply the analyses to real-life datasets for comparison to results from toy datasets.

5.1 Separation of Classes

The goodness of a predictor set can be measured by how well separated the classes of features in the predictor set are. If there are two predictor sets with the same relevance score, the set with better separated classes of features is deemed to be better as there is less redundancy among its features. A natural way to measure separation of classes is the distance between pairs of class means. In this study, we use the Euclidean distance metric. For the q-th pair of classes, $C_q = \{c_{1,q}, c_{2,q}\}$, the separation between classes given by the predictor set found through a DDP value of α, S_α, measured using the Euclidean metric is given below.

$$d_{E,\alpha}(c_{1,q}, c_{2,q}) = \sqrt{\sum_{i \in S_\alpha} \left| \bar{x}_{i,c_{1,q}} - \bar{x}_{i,c_{2,q}} \right|^2} \tag{9}$$

$\bar{x}_{i,k}$ is the average of the expression of gene i across samples belonging to class k. Averaging across all ${}^{K}C_2$ pairs of classes, we obtain the mean Euclidean distance between all pairs of classes as measured by S_α.

$$\bar{d}_{E,\alpha} = \frac{1}{{}^{K}C_2} \sum_{q=1}^{{}^{K}C_2} d_{E,\alpha}(c_{1,q}, c_{2,q}) \tag{10}$$

The value of the DDP leading to the best separation of classes in terms of the Euclidean metric is the one which gives the largest $\bar{d}_{E,\alpha}$.

$$\alpha_E^* = \arg\max_{\alpha} \left(\bar{d}_{E,\alpha} \right) \tag{11}$$

If there is more than one value of α satisfying Eq. 11, the mean among these values is taken as α_E^*. Since these values are generally adjacent to each other, taking the mean will still provide a good picture of how the DDP affects separation of classes.

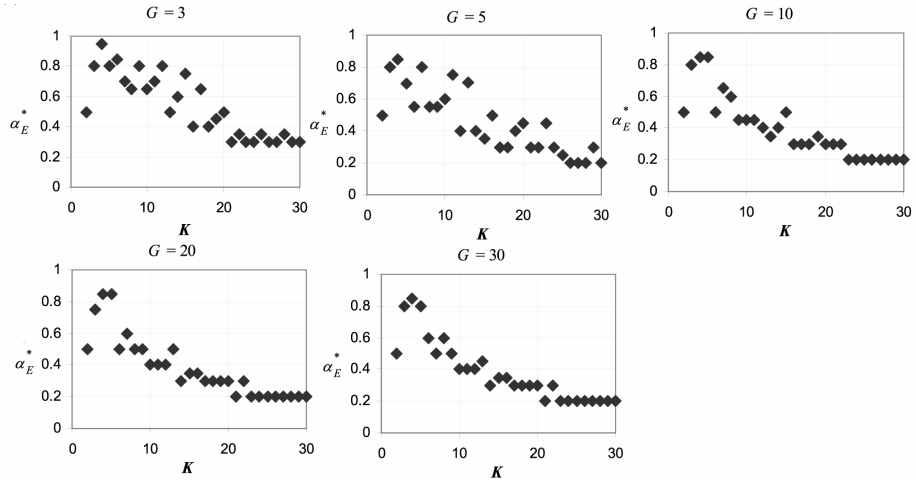

Fig. 1. The DDP producing the optimal separation of classes as measured by the Euclidean distance, α_E^*, as a function of K for toy datasets generated from (a) the OVA model and (b) the PW model

Fig. 1 shows that the number of classes, K, influences the value of α_E^*, regardless of the value set to G. Larger G tends to produce a more distinct α_E^*-K plot. As K increases beyond 20, α_E^* settles to a smaller value (around 0.2). Regardless of the parameter values of G, the number of classes in the dataset undoubtedly affects the value of the DDP which brings about the best separation of classes in terms of the Euclidean distance.

6 Real-Life Datasets

We now apply the previous analyses to real-life datasets. Descriptions of eight real-life microarray datasets are shown in Table 2. The Brown (BRN) dataset [14] includes 15 broad cancer types. Following a previous study [15], the skin tissue samples due to small class size (3 samples) are excluded from analysis. The GCM dataset [13] contains 14 tumor classes. For the NCI60 dataset [16], only 8 tumor classes are analyzed; the 2 samples of the prostate class are excluded due to the small class size.

The PDL dataset [17] consists of 6 classes, each class representing a diagnostic group of childhood leukemia. The SRBC dataset [18] consists of 4 subtypes of small, round, blue cell tumors (SRBCTs). In the 5-class lung dataset [19], 4 classes are subtypes of lung cancer; the fifth class consists of normal samples. The MLL dataset [20] contains 3 subtypes of leukemia: ALL, MLL, and AML. The AML/ALL dataset [21] also contains 3 subtypes of leukemia: AML, B-cell ALL, and T-cell ALL.

Table 2. Descriptions of real-life datasets. *N* is the number of genes after preprocessing. *K* is the number of classes in the dataset.

Dataset	Type	*N*	*K*	Training:Test set size
BRN	cDNA	7452	14	174:83
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

Except for the BRN and SRBC datasets (which are only available as preprocessed in their originating studies), datasets are preprocessed and normalized based on the recommended procedures [12] for Affymetrix and cDNA microarray data. Except for the GCM dataset, for which the ratio of training to test set sizes used in the originating study [13] is maintained to enable comparison with previous studies, for all datasets we employ the standard 2:1 split ratio.

But before applying the analyses to real-life datasets, we investigate how close conditions in real-life datasets match those of toy datasets.

6.1 Investigating the Imbalance of Class Means in Real-Life Datasets

We have mentioned in the section on the generation of toy datasets that imbalance in terms of class means among classes is prevalent in highly multiclass microarray datasets. Investigation is conducted on whole datasets (no splitting) in order to determine the extent of the aforementioned imbalance. For class k ($k = 1,2,\ldots,K$), we choose the class mean with the greatest absolute value (equivalent to the absolute value of $\mu_{(g-1)K+k,k}$ from Eq. 8 or $\mu_{(g-1)\binom{K}{2}+q,c_{b,q}}$ from Eq. 9 at $g = 1$) among all *N* class means.

$$\bar{x}_{0,k} = \max_{i=1,2,\ldots,N} \left(\left| \bar{x}_{i,k} \right| \right) \tag{12}$$

Next, to illustrate the imbalance among classes in terms of class means, we compute the range of class means.

$$R\left(\bar{x}_{0,k}\right) = \max_{k}\left(\bar{x}_{0,k}\right) - \min_{k}\left(\bar{x}_{0,k}\right) \tag{13}$$

The result is shown in Table 3. We observe that the range $R\left(\bar{x}_{0,k}\right)$ for datasets with large *K* (such as BRN, GCM, and NCI60) is greater than the range $R\left(\bar{x}_{0,k}\right)$ for datasets with smaller *K*. Looking at the maximum and minimum values of $\bar{x}_{0,k}$ across *k* in Table 3, we can say with certainty that there is an imbalance among classes in terms of class means, especially in datasets containing more than 6 classes.

Table 3. Range of class means in real-life datasets

Dataset	$\max_k(\bar{x}_{0,k})$	$\min_k(\bar{x}_{0,k})$	$R(\bar{x}_{0,k})$
BRN	11.57	3.93	7.65
GCM	73.00	7.77	65.23
NCI60	7.64	4.20	3.44
PDL	2.77	2.67	0.10
Lung	9.62	8.39	1.23
SRBC	2.27	1.59	0.68
MLL	4.35	4.08	0.27
AML/ALL	7.14	6.76	0.38

We expect this imbalance to either increase or at least stay constant as K increases beyond 14 (which is the largest number of classes to be found among real-life datasets). Therefore the implementation of unequal maximum limits to the absolute value of the class means for different classes in Eqs. 8 and 9 is justified, particularly in analyses involving K as high as 30 for toy datasets, as is the case in this study.

6.2 Applying the Analyses to Real-Life Datasets

For real-life datasets, the analyses are implemented separately upon the training set of each split, there being a total of 10 splits of training and test sets. The mean across all splits is taken for the $\overline{d}_{E,\alpha}$ measured in the analyses, and then used to find the corresponding value of the DDP which optimizes $\overline{d}_{E,\alpha}$.

We will assume that the optimal P for each real-life dataset is directly proportional to K (as is the case for toy datasets). However, allowing for remnant noise (left even after data preprocessing), we assign larger values to P for real-life datasets ($30K$) than for toy datasets with similar K.

Fig. 2 shows that for the majority of real-life datasets, the trend regarding $\overline{d}_{E,\alpha}$ is similar to the trend for toy datasets. However, in the α_E^*–K plot, one dataset

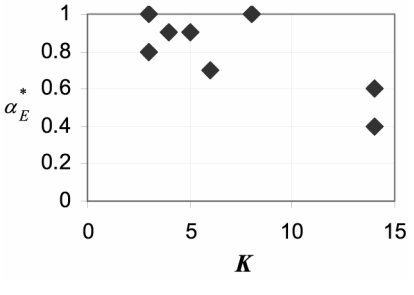

Fig. 2. α_E^* values of the DDP which optimize various predictor set characteristics as a function of K for real-life microarray datasets

(NCI60) produces a point ($\alpha_E^* = 1$ at $K = 8$) which diverges from the $\alpha_E^* - K$ plots observed in toy datasets. Despite this discrepancy (due to small class sizes and the heterogeneity of some of the classes), the overall picture provided by Fig. 2 indicates that the effect of *K* on the values of the DDP which optimize $\overline{d}_{E,\alpha}$ in real-life datasets is the same as the effect in toy datasets.

7 Conclusions

In this paper, we have proposed a systematic method to model toy datasets based on the OVA concept. We have used these toy datasets for analyzing the DDP concept in feature selection. The findings in this study have shown that DDP is necessary because it subsumes existing techniques such as equal-priorities scoring methods and rank-based selection. The optimal value of α is not always 0.5 (equal-priorities scoring methods) or 1 (rank-based selection). Instead, it is based on the number of classes in the datasets. By using this optimal value in the DDP-based feature selection technique, we can then find the predictor set with the best adjustment of maximum relevance and minimum redundancy pertinent to the number of classes in the dataset.

Finally, since all findings have been achieved based on analytical evaluations, not empirical evaluations involving classifiers, this study establishes the theoretical basis for the usefulness of the DDP.

References

1. Hall, M.A., Smith, L.A.: Practical feature subset selection for machine learning. In: Paper presented at the Proc. 21st Australasian Computer Science Conf. (1998)
2. Ding, C., Long, F., Peng, H.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
3. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Machine Learning Research 3, 1157–1182 (2003)
4. Knijnenburg, T.A., Reinders, M.J.T., Wessels, L.F.A.: The selection of relevant and non-redundant features to improve classification performance of microarray gene expression data. In: Proc. 11th Annual Conf. of the Advanced School for Computing and Imaging, Heijen, NL (2005)
5. 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, 2429–2437 (2004)
6. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., et al.: Multi-class cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA 98, 15149–15154 (2001)
7. Chai, H., Domeniconi, C.: An evaluation of gene selection methods for multi-class microarray data classification. In: Paper presented at the Proc. 2nd European Workshop on Data Mining and Text Mining in Bioinformatics (2004)
8. Yu, L., Liu, H.: Redundancy based feature selection for microarray data. In: Paper presented at the Proc. of ACM SIGKDD 2004 (2004)

9. Ooi, C.H., Chetty, M., Gondal, I.: The role of feature redundancy in tumor classification. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072. Springer, Heidelberg (2004)
10. 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., Pereira, A.S. (eds.) ISBMDA 2005. LNCS (LNBI), vol. 3745. Springer, Heidelberg (2005)
11. Ooi, C.H., Chetty, M., Teng, S.W.: Modeling microarray datasets for efficient feature selection. In: Paper presented at the Proc. 4th Australasian Conf. on Knowledge Discovery and Data Mining (AusDM 2005) (2005a)
12. Dudoit, S., Fridlyand, J., Speed, T.: Comparison of discrimination methods for the classification of tumors using gene expression dat. J. Am. Stat. Assoc. 97, 77–87 (2002)
13. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., et al.: Multi-class cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA 98, 15149–15154 (2001)
14. Munagala, K., Tibshirani, R., Brown, P.: Cancer characterization and feature set extraction by discriminative margin clustering. BMC Bioinformatics 5, 21 (2004)
15. Park, M., Hastie, T.: Hierarchical classification using shrunken centroids. Department of Statistics, Stanford University. Technical Report (2005), http://www-stat.stanford.edu/~hastie/Papers/hpam.pdf
16. Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., et al.: Systematic variation in gene expression patterns in human cancer cell lines. Nat. Genet. 24, 227–235 (2000)
17. Yeoh, E.-J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., Mahfouz, R., et al.: Classification, subtype discovery, and prediction of outcome in pediatric lymphoblastic leukemia by gene expression profiling. Cancer Cell 1(2), 133–143 (2002)
18. Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., et al.: Classification and diagnostic prediction of cancers using expression profiling and artificial neural networks. Nat. Med. 7, 673–679 (2001)
19. Bhattacharjee, A., Richards, W.G., Staunton, J.E., Li, C., Monti, S., Vasa, P., et al.: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. USA 98, 13790–13795 (2001)
20. Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., et al.: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat. Genet. 30, 41–47 (2002)
21. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., et al.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)