Relevance, Redundancy and Differential Prioritization in Feature Selection for Multiclass Gene Expression Data

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 various ranking-based filter procedures to classifier-based wrapper techniques, many studies have devised their own flavor of feature selection techniques. Only a handful of the studies delved into the effect of redundancy in the predictor set on classification accuracy, and even fewer on the effect of varying the importance between relevance and redundancy. We present a filter-based feature selection technique which incorporates the three elements of relevance, redundancy and differential prioritization. With the aid of differential prioritization, our feature selection technique is capable of achieving better accuracies than those of previous studies, while using fewer genes in the predictor set. At the same time, the pitfalls of over-optimistic estimates of accuracy are avoided through the use of a more realistic evaluation procedure than the internal leave-one-out-cross-validation.

Keywords: Molecular classification, microarray data analysis, feature selection

1 Introduction

When it comes to multiclass microarray datasets, most of the previous classification studies have taken one of the following stances:

1. Feature selection does not aid in improving classification accuracy [1, 2], at least not as much as the type of classifier used.
2. Feature selection is often rank-based, and is implemented mainly with the intention of merely reducing cost/complexity of subsequent computations (since the transformed dataset is smaller), rather than also finding the feature subset which best explains the dataset [1, 3].
3. Studies proposing feature selection techniques with sophistication above that of rank-based techniques resort to an evaluation procedure which often gives overly-optimistic estimate of accuracy, but has the advantage of costing less computationally than procedures which yield a more realistic estimate of accuracy [4, 5].

From these stances, we see the three levels with which feature selection has been, and still is regarded for multiclass microarray datasets: 1) should not be considered at all, 2) simple rank-based methods for dataset truncation, and finally, 3) more complicated methods with sound theoretical foundation, but with dubious empirical results.

An important axiom governing the principles behind most feature selection works of the third level can be summarized by the following statement: A good predictor set

J.L. Oliveira et al. (Eds.): ISBMDA 2005, LNBI 3745, pp. 367–378, 2005.

should contain features highly correlated with the target class distinction, and yet uncorrelated with each other [6]. The attribute referred to in the first part of this statement is encapsulated in the term 'relevance', and has been the backbone for simple rank-based feature selection techniques, where genes are selected into the predictor set based on the score of their correlation to the target class distinction. The measurement of the aspect alluded to in the second part, 'redundancy' however, is not as straightforward, since the pairwise relationship between each pair of genes in the predictor set needs to be taken into account.

Previous studies [4, 6] 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, Guyon and Elisseeff demonstrated using a 2-class problem that seemingly redundant features may improve the discriminant power of the predictor set instead [7], although it remains to be seen how this scales up to multiclass domains with thousands of features. A study was implemented on the effect of varying the importance of redundancy in predictor set evaluation in [8]. However, due to its use of a relevance score that was inapplicable to multiclass problems, the study was limited to binary classification.

From here, we can rephrase the three levels of feature selection for tumor classification as follows: 1) no selection, 2) pick based on relevance alone, and finally, 3) pick based on relevance and redundancy. Thus, currently, relevance and redundancy are the two existing components used in predictor set scoring methods to evaluate the goodness of a predictor set.

We propose going one step further, by introducing the third element, this element being the relative importance placed between relevance vs. redundancy. This third element compels the search method to prioritize the optimization of one of the elements (of relevance and redundancy) at the cost of the optimization of the other. The degree of differential prioritization is determined by this third element. That is, unlike other existing redundancy-based feature selection studies, with our proposed feature selection technique, it is not taken for granted that the optimizations of both elements of relevance and redundancy are to have equal priorities in the search for the optimal predictor set.

The effectiveness of our proposed feature selection technique on the tumor classification of a multiclass microarray dataset has been reported in [9]. However, this paper aims to do more than illustrate the efficacy of the technique on various other multiclass microarray datasets. More importantly, applying our technique to *multiple* such datasets makes it possible for us to discern the relationship between dataset characteristics and the optimal degree of differential prioritization for a particular dataset.

Having introduced the element of differential prioritization, we go on to demonstrate the importance of applying evaluation procedure which yields more realistic estimate of accuracy than the internal cross validation procedure used in recent tumor classification studies [3, 4, 5]. This is done by evaluating our feature selection techniques using two evaluation procedures: the first being the F-splits procedure, the second is the aforementioned internal cross validation procedure.

The contributions of this study are threefold: 1) to show that a degree of freedom in adjusting the priorities between maximizing relevance and minimizing redundancy is necessary to produce the best classification performance (i.e. equal-priorities techniques might not yield the optimal predictor set); 2) to demonstrate the relationship between dataset characteristics and the optimal degree of differential prioritization; and 3) to highlight the importance of using a realistic evaluation procedure.

2 Methods

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 target class vector $\mathbf{y}$ is defined as $[y_1, \ldots, y_{Mt}]$, $y_j \in [1,K]$ in a K-class dataset. Gene i, on the other hand, is represented by vector $\mathbf{g}_i$, containing expression of gene i across the M_t samples in the training set, $[x_{i,1},\ldots, x_{i,Mt}]$. From the total of N genes, the objective of feature selection is to form the subset of genes, called the predictor set S, which would give the optimal classification accuracy.

2.1 The Antiredundancy-Based Scoring Method

A score of goodness incorporating both the elements of maximum relevance and minimum redundancy ensures that the optimal predictor set should possess maximal power in discriminating between different classes (maximum relevance), while at the same time containing features with minimal correlation to each other (minimal redundancy).

For the purpose of defining our predictor set scoring method, without loss of generality, we define the following parameters.

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

Both V_S and U_S are to be maximized in the search for the optimal predictor set.

U_S quantifies the *lack of redundancy* in S. With U_S, we have an antiredundancy-based scoring method in which the measure of goodness for predictor set S is given as follows.

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

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

2.2 Significance of the Differential Prioritization Factor, α

In the previous section it has been stated that an optimal predictor set is to be found based on two criteria: maximum relevance and maximum antiredundancy. However, the quantification of the priority to be assigned to each of these two criteria remains an unexplored area.

In the antiredundancy-based scoring method, 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 vector, 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 vector. 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 degrees of prioritization between maximizing relevance and maximizing antiredundancy in order to come up with the most efficacious predictor set. Therefore the optimal range of α (optimal as in leading to the predictor set giving the best estimate of accuracy) is dataset-specific.

2.3 Definitions of Relevance and Antiredundancy

The measure of relevance for S is computed by averaging up the score of relevance, $F(i)$ of all members of the predictor set, as recommended in [4]:

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

$F(i)$ is the score of relevance for gene i. It indicates the correlation of gene i to the target class vector $\mathbf{y}$. For continuous-valued datasets, a popular parameter for computing $F(i)$ is the BSS/WSS ratios (the F-test statistics) used in [4, 10]. 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 [10] for multiclass tumor classification, is a modification of the F-ratio statistics for one-way ANOVA (Analysis of Variance). It indicates the gene's ability in discriminating among samples belonging to the K different classes.

The measure of antiredundancy for S is computed by summing up one minus absolute values of the measures of correlation between all possible pairwise combinations of the members of S, and normalizing by division with the square of the size of S. Since both correlation and anti-correlation contribute to redundancy in S, absolute values of correlation are used.

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

For continuous-valued datasets, a conventional measure of correlation between pairs of genes is the absolute value of the Pearson product moment correlation coefficient, which measures similarity between two genes. Between genes p and q, the measure of correlation $R(p,q)$ is the Pearson product moment correlation coefficient between genes p and q. Larger U_S indicates lower average pairwise correlation in S, and hence, smaller amount of redundancy among the members of S.

2.4 The Search Method

An exhaustive search for the optimal predictor set is computationally expensive. For instance, in searching for the best S of a certain size P, the order of complexity is $O(N^P)$. We employed the linear incremental search method, 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 $W_{A,S}$. This search method, with a lower computational complexity of $O(NP)$, has been applied in previous feature selection studies [4, 5].

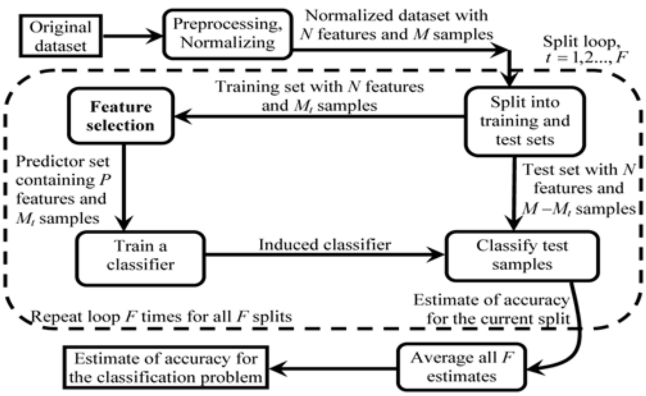

Fig. 1. *F*-splits procedure

2.5 Over-Optimistic Estimate of Accuracy

In several previous studies on feature selection for microarray datasets [3, 4, 5], feature selection techniques have been applied *once* on the *full* dataset before leave-one-out-cross-validation (LOOCV) procedure is employed to evaluate the classification performance of the resulting predictor sets. We denote this evaluation procedure the **I**nternal LOO**CV** (ICV) procedure. ICV is known to produce selection bias, which leads to an overly-optimistic estimate of accuracy [11].

To avoid this pitfall, we propose the use of different splits of the dataset into training and test sets and *repeating feature selection for each of the splits*. During each split, our feature selection techniques will be applied only on the training set of that particular split. It is very important that no information from the test set is ever 'leaked' into the process of forming the predictor set (which is precisely what happens during the ICV procedure). Classifier trained on the predictor set and the training samples will then be used to predict the class of the test samples of the current split. The test set accuracies obtained from each split will be averaged to give an estimate of the classification accuracy. We call this procedure of accuracy estimation the F-splits procedure (F being the number of splits used) (Figure 1).

In addition to accuracy, we used an approximation of the area under the Receiver Operating Characteristic (ROC) curve (AUC) as a performance evaluation parameter. The approximation used is the modified macro-average of class accuracies (MAVG-MOD) recommended in [12] for multiclass problems employing crisp classifiers.

$$\text{MAVG-MOD} = \left(\frac{1}{K} \sum_{k=1}^{K} a_k^{\tau} \right)^{\frac{1}{\tau}} \tag{5}$$

For best performance, the value of τ has been determined as 0.76 [12]. The class accuracy for class k is represented by a_k.

3 Results

3.1 Benchmark Datasets and Evaluation Procedures

Several multiclass microarray datasets are used as benchmark datasets (Table 1). The GCM dataset [2] contains 14 tumor classes: breast, prostate, lung, colorectal, lymphoma, bladder, melanoma, uterus, leukemia, renal, pancreas, ovarian, mesothelioma and CNS (central nervous system). For NCI60 [13], only 8 tumor classes (breast, CNS, colon, leukemia, melanoma, ovarian, renal and non-small-cell lung cancer) are analyzed; the 2 samples of the prostate class are excluded due to the small class size. In the 5-class lung dataset [14], 4 classes (lung adenocarcinoma, squamous cell lung carcinoma, pulmonary carcinoid and small-cell lung cancer) are subtypes of lung cancer; the fifth class comprises of normal samples. The MLL dataset [15] contains 3 subtypes of leukemia: ALL, MLL and AML. The AML/ALL dataset [16] also contains 3 subtypes of leukemia: AML, B-cell and T-cell ALL. Datasets are preprocessed and normalized based on the recommended procedures in [10] for Affymetrix and cDNA microarray data.

Different degrees of importance were placed on antiredundancy measure by varying the values of α from 0.1 up to 1. Optimal predictor sets ranging from sizes P=2,3,…,100 were formed in the different runs.

For each dataset we implemented two different evaluation procedures: the 10-splits procedure, and the ICV procedure employed in previous tumor classification studies [3, 4, 5]. The difference between the estimates of accuracy obtained from the two procedures offers us an insight into the effect of selection bias which occurs when feature selection is not repeated for different splits or subsets of the dataset.

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

Dataset	Type	N	K	Training:Test set ratio (10-splits procedure)
GCM	Affymetrix	10820	14	144:54
NCI60	cDNA	7386	8	40:20
Lung	Affymetrix	1741	5	135:68
MLL	Affymetrix	8681	3	48:24
AML/ALL	Affymetrix	3571	3	48:24

The DAGSVM classifier is used throughout the performance evaluation for both 10-splits and ICV procedures. The DAGSVM is an SVM-based multi-classifier which uses substantially less training time compared to either the standard algorithm or Max Wins, and has been shown to produce accuracy comparable to both of these algorithms [17].

3.2 Best Estimates of Accuracy for the Benchmark Datasets

The best estimate of accuracy obtained from each dataset is shown in Table 2. Where a draw occurs in terms of the estimate of accuracy, the α value giving the smaller predictor set size is proclaimed as the optimal α. Comparisons with previously reported results will only be made for the 2 datasets which have been known to produce low realistic estimates of accuracy (<90%), the GCM and NCI60 datasets [1].

For the GCM dataset, with a predictor set containing no more than 94 genes at most, an accuracy of 80.6–84.3% is achievable with our predictor set scoring method when the value of α is set within the range of 0.2–0.3. This is a significant improvement compared to the 78% accuracy obtained, using all available 16000 genes, in the original analysis of the same dataset [2]. However, strict comparison cannot be made against this 78% accuracy of [2] and the 81.5% accuracy (using 84 genes) achieved in [18] since the evaluation procedure in both studies [2, 18] is based on a single (the original) split. We can make a more appropriate comparison, however, with a comprehensive study on various rank-based feature selection techniques [1]. The study uses external 4-fold cross validation to evaluate classification performance. In [1], the best accuracy for the GCM dataset is 63.3%, when *no* feature selection is applied prior to classification!

For the NCI60 dataset, the best accuracy of 74% from the 10-splits evaluation procedure occurs at $\alpha = 0.3$, and is better than the best accuracy obtained from the two studies employing a similar evaluation procedure [1, 10]. In [10], the best averaged accuracy is around 63% (using the top 30 BSS/WSS-ranked genes), whereas the study in [1] performs slightly better with best accuracy of 66.7% (150 genes) achieved using the sum minority rank-based feature selection technique [1]. For the estimate of accuracy from the ICV procedure, the ICV estimate of 96.7% for our predictor set scoring method is significantly higher than the best ICV estimate in [4] for the continuous-valued version of their predictor set scoring method (80.6%).

Table 2. Best accuracy estimated from the 10-splits and ICV procedures, followed by the corresponding differential prioritization factor and predictor set size

Dataset	10-splits	ICV
GCM	80.6%, α=0.2, 85 genes	84.3%, α=0.3, 94 genes
NCI60	74.0%, α=0.3, 61 genes	96.7%, α=0.2, 89 genes
Lung	95.6%, α=0.5, 31 genes	96.1%, α=0.4, 43 genes
MLL	99.2%, α=0.6, 12 genes	98.6%, α=0.6, 4 genes
AML/ALL	97.9%, α=0.8, 11 genes	98.6%, α=0.6, 5 genes

From Table 2 and Figure 2, it can be seen that the optimal value of α, (i.e. the value of α where the best accuracy is obtained) is not necessarily 0.5 (denoting equal priorities for maximization of relevance and maximization of antiredundancy) or 1 (in which our feature selection technique becomes rank-based selection technique). As mentioned previously in Section 2.2, for a given predictor set scoring method and a set of definitions of $F(i)$ and $R(i,j)$, the optimal range of α is most likely dataset-dependent, as shown by the results.

For the GCM dataset, the highest 10-splits accuracy, 80.6% is obtained when α is set to 0.2. If we had limited ourselves to equal priorities for the maximizations of relevance and antiredundancy (i.e. equivalent to setting α to 0.5), we would have achieved only a 75.9% accuracy, and if we had been content with rank-based techniques (i.e. equivalent to setting α to 1), we would have fared worse, with only a measly 65.6% if the maximum size of the predictor set is set to 100.

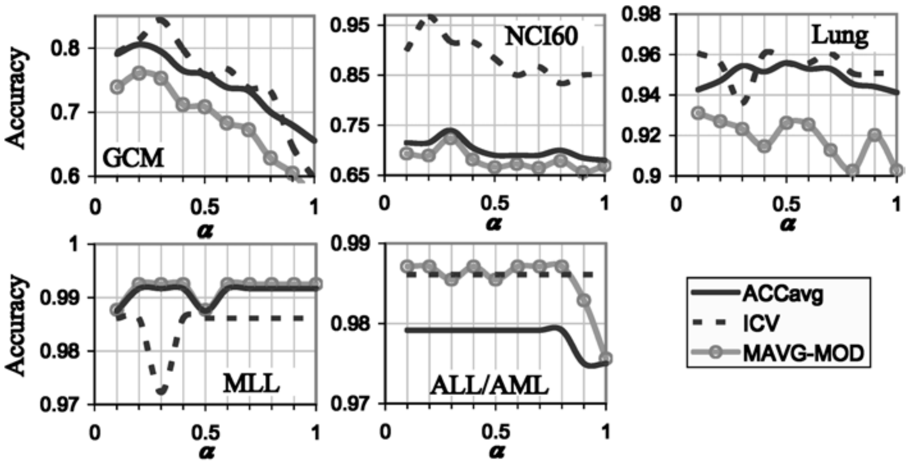

Fig. 2. MAVG-MOD and best accuracy estimate from the 10-splits procedure (ACC_{avg}) and best accuracy estimate from the ICV procedure respectively plotted against α

The same can be said for the NCI60 dataset. Using $\alpha = 0.3$ we obtain the best accuracy of 74% based on the 10-splits evaluation procedure, whereas at $\alpha = 0.5$ and $\alpha = 1$, much lower accuracies (69% and 68% respectively) are achieved.

The lung dataset is the only dataset tested where the best 10-splits accuracy is obtained when α is set to 0.5. Rather than presenting a conflict against the results from the other datasets, in Section 4.2 we will prove that in case of the lung dataset 0.5 merely happens to be the optimal α due to the characteristics of the lung dataset itself, in the same way that the values of the optimal α for each of the other datasets are influenced by the characteristics of the respective datasets.

For the MLL dataset, we have an accuracy of 99.2% at $\alpha = 0.6$. When the predictor set scoring method gives equal priorities to relevance and redundancy ($\alpha = 0.5$), the accuracy achieved drops to 98.7%. However, when the selection is rank-based ($\alpha = 1$), the same accuracy of 99.2% is obtained, but using *twice* the number of genes (24 instead of 12 genes) compared to the predictor set scoring method run with $\alpha = 0.6$. This is not surprising, considering that at $\alpha = 1$, genes are selected based on relevance without regards to redundancy, hence the bigger size of predictor set due to the inclusion of redundant genes.

For the AML/ALL dataset, we get an accuracy of 97.9% at $\alpha = 0.8$ using an 11-gene predictor set. At $\alpha = 0.5$, the same accuracy is achieved but using twice the number of genes (20 genes). Accuracy drops to 97.5% when the value $\alpha = 1$ is used.

With the single exception of the lung dataset, Figure 2 shows that our alternative performance evaluation parameter, MAVG-MOD, demonstrates the same trend against α as accuracy does, i.e., the peak of the accuracy curve always coincides with the peak of the MAVG-MOD curve. Even for the lung dataset, the peak of the MAVG-MOD at $\alpha = 0.5$ is only slightly lower than the peak at $\alpha = 0.1$ (by 0.005). Looking at the class accuracies for this dataset, we found the underlying reason: the class accuracies of the 4 classes with the best class accuracies peak at $\alpha = 0.1$, whereas only one class, the worst-performing class (squamous cell lung carcinoma) has its best accuracy at $\alpha = 0.5$. Therefore, being capable of producing the highest class accuracy for the worst-performing class, $\alpha = 0.5$ is still the optimal value of α for the lung dataset.

4 Discussion

4.1 Selection Bias

Selection bias is reflected in the difference between the accuracy estimated from the 10-splits procedure and the accuracy obtained from the ICV procedure for each predictor set size (*P*=2,3,…,100). By contrasting between the results from the 10-splits and the ICV procedures, it is clear that the apparently better performance reported previously [4] is a product of selection bias.

In order to quantify selection bias susceptibility for a dataset, we use the ratio of the number of features to the median class size, N/CS_{50} (Table 3). Greater number of total features (N) and smaller class sizes (CS_{50}) mean higher likelihood for a search method to find the predictor set which fits the data, thereby increasing the probability of over-fitting. In *F*-splits evaluation procedure, over-fitting can be easily detected through the resulting low accuracy estimate from the classification of test sets uninvolved in feature selection. In ICV evaluation procedure, over-fitting naturally results in higher accuracy estimate since *all* samples have been involved in feature selection.

Table 3. Median class size, CS_{50}, ratio of N to CS_{50} and optimal α values for each dataset

Dataset	K	CS_{50}	N/CS_{50}	optimal α
GCM	14	12	901.7	0.2
NCI60	8	7.5	984.8	0.3
Lung	5	20	87.1	0.5
MLL	3	24	361.7	0.6
AML/ALL	3	25	142.8	0.8

The N/CS_{50} ratios for all datasets are correlated to the selection bias shown in Figure 3. The NCI60 dataset is the most susceptible to selection bias due to its high N/CS_{50} ratio of 984.8. The GCM dataset, which has the second highest overall selection bias, also has the second highest N/CS_{50} ratio (901.7). The remaining three datasets (lung, MLL and AML/ALL datasets), for each of which the N/CS_{50} ratio is below 400, have relatively smaller selection bias (near 0% for most of the tested values of α).

4.2 Optimal Differential Prioritization Factor and Dataset Characteristics

The relationship between the optimal value of α and the number of classes, K, for all benchmark datasets is illustrated in the left-side panel of Figure 4. The effect of the class size, represented by the median class size, CS_{50}, on the optimal value of α is shown in the right-side panel of Figure 4. It can be seen that the optimal value of α decreases as K increases, but increases as CS_{50} becomes larger.

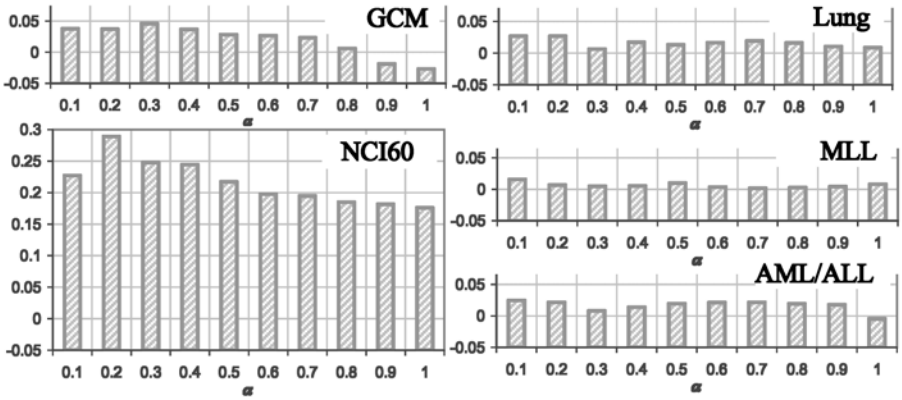

Fig. 3. Selection bias in GCM, NCI60, lung, MLL and AML/ALL datasets. Averaged bias among predictor sets of different sizes (P=2,3,…,100) plotted against α

The optimal value of α (Table 3) has a strong positive correlation to CS_{50} (0.90) but strong negative correlation to K (–0.89). This means that with **smaller** class sizes and **more** classes per dataset, the Pearson-moment-based antiredundancy plays increasingly important role in the search for the optimal predictor set than the BSS/WSS-based relevance (as reflected by the smaller optimal α). Conversely, maximizing antiredundancy becomes less important as K decreases – therefore supporting the assertion in [7] that redundancy does *not* hinder the discriminant power of the predictor set when K is 2.

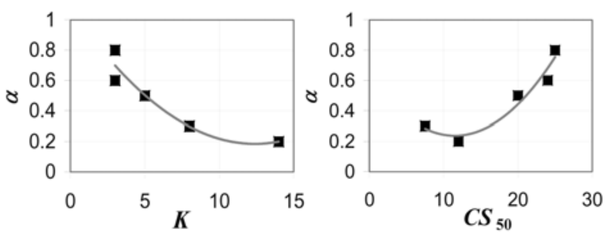

Fig. 4. Optimal values of α for all benchmark datasets plotted against K (left-side panel) and CS_{50} (right-side panel) of corresponding datasets

Larger number of multiclass datasets of diverse characteristics needed to be tested before a more definite rule can be determined regarding the optimal choice for the value of α – which we know, for now, most likely depends on at least two characteristics of the dataset, class size (CS_{50}) and the number of classes, K.

5 Conclusion

For majority of the datasets tested, the differential prioritization factor makes it possible to achieve an accuracy rate higher than the rates obtainable using an equal-priorities scoring method (α fixed at 0.5) or a rank-based selection technique (α fixed at 1). Therefore, instead of limiting ourselves to a fixed universal set of priorities for relevance and antiredundancy (α fixed to 0.5 or 1) for **all** datasets, a suitable range for α should be chosen based on the characteristics of the dataset of interest in order to achieve the optimal estimate of accuracy.

Estimate of accuracy from the ICV procedure, which has been popularly used for gene expression datasets due to its low computational cost, can be radically overly-optimistic, particularly when the N/CS_{50} ratio is large.

References

1. 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
2. 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
3. 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
4. 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
5. Yu, L., Liu, H.: Efficiently Handling Feature Redundancy in High-Dimensional Data. In: Domingos, P., Faloutsos, C., Senator, T., Kargupta, H., Getoor, L. (eds.): Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM Press, New York (2003) 685–690
6. Hall, M.A., Smith, L.A.: Practical feature subset selection for machine learning. In: McDonald, C. (ed.): Proc. 21stAustralasian Computer Science Conference. Springer, Singapore (1998) 181–191
7. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3 (2003) 1157–1182
8. Knijnenburg, T.A.: Selecting relevant and non-redundant features in microarray classification applications. M.Sc. Thesis. Delft University of Technology. http://ict.ewi.tudelft.nl/pub/marcel/Knij05b.pdf (2004)
9. Ooi, C.H., Chetty, M., Gondal, I.: The role of feature redundancy in tumor classification. In: He, M., Narasimhan, G., Petoukhov, S. (eds.): Proc. International Conference on Bioinformatics and its Applications (ICBA'04). World Scientific Publishing (2004) 197–208
10. Dudoit, S., Fridlyand, J., Speed, T.: Comparison of discrimination methods for the classification of tumors using gene expression data. JASA 97 (2002) 77–87
11. Ambroise, C., McLachlan, G. J.: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. 99 (2002) 6562–6566
12. Ferri, C., Hernández-Orallo, J., Salido, M.A.: Volume under the ROC Surface for Multi-class Problems. In: Lavrac, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.): Proc. of the 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia. Springer (2003) 108–120

13. 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
14. 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
15. 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
16. 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
17. Platt, J.C., Cristianini, N., Shawe-Taylor, J.: Large Margin DAGs for Multiclass Classification. Advances in Neural Information Processing Systems 12 (2000) 547–553
18. 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