AN INTEGRATED APPROACH TO FEATURE SELECTION AND CLASSIFICATION FOR MICROARRAY DATA WITH OUTLIER DETECTION

Y.Y. Leung* and Y.S. Hung

Department of Electrical and Electronic Engineering, University of Hong Kong,
CYC Building, Pokfulam Road, Hong Kong
**Email:yyleung@eee.hku.hk*

Microarray data classification remains a challenging problem due to the curse of dimensionality (i.e., large number of features and small sample size), making feature selection a key element of most classification studies. Some approaches, such as the filter-wrapper approach, integrate feature selection and classification into a combined problem. We extend the filter-wrapper method to a multiple-filter-multiple-wrapper (MFMW) approach whereby multiple statistics are used to reduce the original set of genes to a manageable size, and multiple classifiers are then used to iteratively select a small set of genes by a voting scheme based on classification accuracy. It can be shown that the MFMW method returns superior performance compared with existing methods. However, the classification results are inevitably sensitive to outliers (i.e., samples that are mislabeled) in the dataset, and therefore it is essential to address the issue of outliers. In this paper, we further propose to integrate outlier detection into the MFMW method by taking advantage of the information already available in the votes cast by the multiple classifiers. Hence, a "three-in-one" algorithm is developed to perform gene selection, sample classification and outlier detection simultaneously. In this algorithm, we pay special attention to maintain a stable set of selected genes, by using an L1-norm support vector machine to remove redundant genes. The performance of the integrated MFMW approach will be illustrated by means of synthetic and real data, including the Leukemia (7129 genes, 72 samples) and Colon (2000 genes, 62 samples) datasets.

1. INTRODUCTION

Classification has been one of the key problems in microarray data analysis. However, classification based on gene expression data is not easy due to the characteristics of the data: high dimensionality and small sample size. This problem makes conventional machine learning tools not suitable for use. The reduction in performance of the algorithm for datasets with many features is known as the curse of dimensionality.[1] To overcome the curse of dimensionality, we need to extract genes that are truly relevant to the disease. This problem of identifying relevant genes is known as gene selection. Gene selection can provide faster, more cost-effective models with better classification performance, while at the same time allow deeper understanding of the biology of the data.[2]

There are two main types of gene selection methods: filters and wrappers. Filters select genes simply based on the statistical scores of genes. They do not take into account the gene interactions when selecting genes. On the other hand, gene sets selected by wrappers are usually with best discriminative potential for a fixed classifier selected.[3] Though wrappers take into account the interactions between gene subset, finding such an optimum gene set requires high computational cost.

Studies have confirmed that there might never be a 'best' single approach (in the sense of using a single gene selection method, whether it is a filter or a wrapper, with any classifier) for revealing patterns of gene expression.[4] Filter-wrapper methods have been proposed to find an optimum balance between the precision of biomarker discovery and the computation cost, by taking advantages of both filter methods' efficiency and wrapper methods' high accuracy.[5] Different filter-wrapper methods have been proposed, yet different gene sets are selected by different studies based on the same dataset. The (lack of) reproducibility of gene lists have been noted in some recent publications[6-7] and this issue is referred to as 'stability'. Moreover in some datasets, classification performance seldom reaches perfect when using different gene selection tools and classifiers.[8] This indicates the possibility of the presence of outlying samples in the data.

2. MULTIPLE-FILTER-MULTIPLE-WRAPPER MODEL

Due to the limitations of the current filter-wrapper methods, a multiple-filter-multiple-wrapper (MFMW) approach has been proposed which combine multiple filters and multiple wrappers into a single model. Each of several different filters (e.g. Signal-to-noise ratio, *t*-statistics, Area under the Receiver-Operating-Characteristics curve) is first used to select a certain number of genes (say 200) out of several thousand present in the microarray data respectively. The gene lists obtained by different filters are then combined to provide a merged filtered subset of genes of manageable size. The use of multiple filters can help to ensure that useful biomarkers are unlikely to be screened out in the initial filter stage.

The use of multiple wrappers (e.g. Weighted Voting, *k*-Nearest Neighbor, Support Vector Machine) is intended to enhance the reliability of the classification by establishing consensus among several classifiers. The consensus is achieved by the application of unanimous voting for deciding the overall classification output based on the outputs of the classifiers. In the case where a unanimous vote cannot be reached, the classification output is regarded as indecisive (denoted as ‘X’). Suppose we chose three wrappers in our MFMW model. Table 1 illustrates the voting results for all possible combinations of the outputs of three classifiers in a two-class (with labels ‘A’ and ‘B’) classification problem where the sample has a true class label ‘A’. Out of these combinations of classifier outputs, only two will produce a unanimous vote of class A or B, one of which is right and the other is wrong (with predication status ‘R’ and ‘W’ respectively). The other six produce an indecisive outcome (denoted as ‘I’). By unanimous voting, each sample can be categorized as either ‘R’, ‘I’ or ‘W’. This information can be used for gene selection in the MFMW model. The number of ‘I’ and ‘W’ prediction statuses across all samples will be used to determine the usefulness of the set of genes. A prediction status ‘W’ implies that all classifiers misclassify the sample under consideration, and is therefore particularly undesirable. Our first objective in gene selection is therefore to minimize the number of ‘W’. Beyond that, the next goal will be to reduce the number of ‘I’. The score of ‘W’ and ‘I’ will be taken together as the consensus score. If there is more than one gene that is given the same consensus score, L1-norm Support Vector Machine (SVM) is used to select the set of most informative genes. L1-norm SVM was shown to be able to automatically select relevant genes when there are redundant noises present.[5] Hence, the final genes selected can be considered to be more robust with a mixture of characteristics that fit several wrappers, and are therefore better qualified as biomarkers. Furthermore, since the MFMW model already incorporates the characteristics of multiple filters and wrappers, it is no longer necessary to try different filter-wrapper combinations in order to search for a suitable combination that yields the highest classification accuracy. For interested readers, please refer to for details. [9]

Table 1. All possible combinations (C1-C8) of outputs of three classifiers CF1, CF2 and CF3 in a two-class classification problem.

	C1	C2	C3	C4	C5	C6	C7	C8
True class label	A	A	A	A	A	A	A	A
CF1	A	A	A	A	B	B	B	B
CF2	A	A	B	B	A	A	B	B
CF3	A	B	A	B	A	B	A	B
Classification Output	A	X	X	X	X	X	X	B
Prediction Status	R	I	I	I	I	I	I	W

The classification results are inevitably sensitive to outliers (i.e., samples that are mislabeled) in the dataset. To address the issue of outliers, we propose to integrate outlier detection into the MFMW method by taking advantage of the information already available in the votes cast by the multiple classifiers, yielding a “three-in-one” algorithm that performs gene selection, sample classification and outlier detection simultaneously. We make use of external Leave-one-out cross-validation (LOOCV) to perform outlier detection. LOOCV is chosen as the model estimator as study shows that the LOO estimate produces a practically unbiased estimate of the expected error rate if the samples are statistical independent.[10] Each time, one sample is being left out and the remaining samples are used for building the MFMW model. After selecting a gene set, it is then tested on the left-out sample and its class label is determined. We treat the sample as an ‘outlier’ if all the classifiers employed cannot give a correct classification label to this particular sample. After repeating the procedure for all samples, those ‘outliers’ are removed and the same external LOOCV is performed on this new reduced dataset with a smaller number of samples. This procedure is repeated iteratively until no more samples are given a wrong class label by all classifiers.

The final gene set is obtained from the last iteration stage when no more samples are excluded. Suppose we have performed LOOCV *n* times. We take all *n* gene lists together and rank the genes by their frequency of occurrence out of *n* times, thus giving a measure for the relative importance of a gene for final class prediction. [7] A gene is most certain to be relevant if it is selected most of the time.

3. DATASETS

The proposed MFMW model was evaluated by means of two DNA microarray datasets, namely LEU72, [11] and COL62. [12]

LEU72 dataset– 72 samples were analyzed with Affymetrix oligonucleotide arrays. We combine the original training and testing dataset provided. In total, there are 47 Acute Lymphoblastic Leukemia (ALL) and 25 Acute Myeloid Leukemia (AML) samples with 7129 probes (6817 genes).

COL62 dataset – Gene expression in 40 tumor and 22 normal colon tissue samples were analyzed with Affymetrix oligonucleotide arrays. 2000 out of around 6500 genes were selected based on the confidence in the measured expression levels.

4. EXPERIMENTAL RESULTS

Our proposed MFMW model was experimented using LEU72 and COL62 data. Results are shown as follows.

For LEU72 dataset, after performing LOOCV for 72 times, the classifiers didn't output a correct class label only when leaving out sample 66. Sample 66 was then removed. At the 2nd iteration stage, we were left with 71 samples. None of the 71 models predict a wrong class label for the remaining sample. The iteration process stopped and we conclude that only one outlier (sample 66) present in the LEU72 data. Table 2 summarizes the results for LEU72 data.

Table 2. Details showing which samples are removed as outliers in each iteration of the MFMW for the LEU72 data.

	Total # of samples in the dataset	Suspected outlier (sample number)
1st iteration	72	66
2nd iteration	71	NIL

We take all the genes selected in the 71 gene lists of 2nd iteration and count the frequency of occurrence of these genes among the 71 gene lists. There are a total of 42 genes. Table 3 shows the ten genes with highest frequency count. Only the first 4 genes are taken into our final gene set as the frequency count of the remaining are far too small.

Table 3. 10 genes with highest frequency count for all 71 gene lists in the 2nd iteration MFMW model of LEU72 data.

Gene id	Frequency Count	Gene id	Frequency Count
760	59	**2288**	6
6169	54	**4847**	6
1829	47	**3183**	5
3847	41	**4291**	5
6215	8	**1779**	4

Similarly the MFMW model was applied to the COL62 data. Samples which are suspected outliers were removed at each iteration. In the first iteration, there were four such samples. These four samples (T33, T36, T37, N20) were removed and then in 2nd iteration, we were left with 58 samples. Samples T2 and T30 were removed in the 2nd iteration while samples N2, N8 and N18 were removed in the 3rd iteration. In the 4th iteration, no more samples were removed and the dataset was left with 53 samples. Our result is in coincidence with existing outlying-detection studies on detecting outliers in the COL62 data.[13-15] We summarize these results in Table 4.

Table 4. Details showing which samples are removed as outliers in each iteration of the MFMW for the COL62 data.

	Total # of samples in the dataset	Suspected outlier (sample number)
1st iteration	62	T33, T36, T37, N20
2nd iteration	58	T2, T30
3rd iteration	56	N2, N8, N18
4th iteration	53	NIL

We take all the genes selected in the 53 gene lists of 4th iteration and count the frequency of occurrence of these genes among the 53 gene lists. There are a total of 15 genes. Table 5 shows the ten genes with highest frequency count. We only take the first 7 genes into our final gene set as it's hard to determine which genes to be selected further down the list as they have the same frequency count.

Table 5. 10 genes with highest frequency count for all 53 gene lists in the 4th iteration MFMW model of COL62 data.

Gene id	Frequency Count	Gene id	Frequency Count
1635	29	**267**	7
929	28	**365**	5
249	15	**26**	2
1884	14	**391**	2
1772	12	**493**	2

5. DISCUSSIONS AND CONCLUSIONS

Our proposed 'three-in-one' MFMW approach is shown to be able to perform gene selection, sample classification and outlier detection simultaneously. With the help of LOOCV, genes selected in the training data are tested on the left-out sample, the class label of which can be determined and which in turn help to identify whether the left-out sample is an outlier or not. The gene set obtained is stable and small in size.

For validation, we have changed the class labels of our detected outliers to that of the opposite class. We then compare the classification accuracies of the original data with the new data containing these detected outliers of adverse class labels. These were done on both LEU72 and COL62 data using the final gene set. This set of stable genes can achieve 100% classification accuracy, as compared to 98.61% (LEU72) and 85.48% (COL62) on the original data.

References

1. Bellman RE. *Adaptive Control Processes.* Princeton University Press, Princeton, NJ. 1961.
2. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007; **23**(19): 2507-2517.
3. Hauskrecht M, Pelikan R, Valko M, Lyons-Weiler J. *Fundamentals of Data Mining in Genomics and Proteomics.* Springer, New York, USA. 2007: 149–172.
4. Quackenbush J. Microarray analysis and tumor classification. *N Engl J Med* 2006; **354**(23): 2463-2472.
5. Peng Y, Li W, Liu Y. A hybrid approach for biomarker discovery from microarray gene expression data for cancer classification. *Cancer Inform* 2007; **2**: 301-11.
6. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 2005; **21**(2): 171-178.
7. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 2005; **365**(9458): 488 – 492.
8. Quackenbush J. Computational analysis of microarray data. *Nat Rev Genet* 2001; **2**(6): 418-427.
9. Leung Y, Hung Y. A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2008. IEEE computer Society Digital Library. IEEE Computer Society.*
10. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. *J. Natl. Cancer Inst.* 2003; **95**(1): 14-18.
11. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* 1999; **286**(5439): 531 - 537.
12. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA* 1999; **96**(12): 6745 - 6750.
13. Lu X, Li Y, Zhang X. A simple strategy for detecting outlier samples in microarray data. *Control, Automation, Robotics and Vision Conference.* 2004; **2**: 1331- 1335.
14. Kadota K, Tominaga D, Akiyama Y, Takahashi K. Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification. *Chem-Bio Informatics J.* 2003; **3(1)**: 30–45.
15. A. Malossini, E. Blanzieri, R. Ng. Detecting potential labeling errors in microarrays by data perturbation. *Bioinformatics* 2006; **22**(17): 2114.