A Post-filtering Gene Selection Algorithm Based on Redundancy and Multi-gene Analysis

Liangliang Wang, Jinwen Ma

Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, 100871, China

jwma@math.pku.edu.cn

Abstract

This paper proposes a post-filtering gene selection algorithm to discover informative genes of a tumor. With the identified genes via some conventional statistical test methods on the microarray data, the proposed algorithm utilizes a correlation measure and an approximate Markov blanket from the statistical tests to reduce redundant ones. It further evaluates the goodness of classification from the point of view of multi-gene action instead of individual gene action. To test the effectiveness of the post-filtering gene selection algorithm

statistics [2][3][4]. Usually, these conventional methods select the top k genes. But it has been shown that the genes selected by individual gene evaluations are often highly correlated. To maintain a high prediction accuracy of the classifier, we must find uncorrelated but still highly informative genes. Consequently, the Markov blanket was introduced to reduce the redundant genes from the selected ones [1][5][6]. Since the Markov blanket is purely theoretical, it was implemented by some approximates. However, these methods are sensitive to the order of genes according to their individual ranks. That is, these top genes have greater influence than other genes. As a result, it may eliminate too many genes and thus some useful information may be lost. Except individually selected genes, valuable information can also be discovered by evaluating the classification of combinations of genes [7].

To better tackle the problems mentioned above, we propose a post-filtering gene selection algorithm to select the informative genes of a tumor via a microarray data set by combining the multi-gene analysis with individual gene selection. Since non-parametric statistical tests have been shown to be efficient on informative gene selection [2][3][4], we will use statistical tests to select the genes for the redundancy analysis. The post-filtering gene selection process will reduce redundant genes by analyzing the correlation among the selected genes as well as their approximate Markov blankets based on the statistical tests. Moreover, the multi-gene analysis is conducted to keep the genes which play an important ro

where $\bar{X}_1$ and $\bar{X}_2$ are the two sample means, n_1 are n_2 are the sizes of the two samples, and $S = (S_1 + S_2)/(n_1 + n_2 - 2)$, where S_1 and S_2 are the two sample covariance matrices.

3. If the two covariance matrices are not the same according to the above testing result, the 2-sample Hotelling T^2 statistic is given by:

$$T^2 = (\bar{X}_1 - \bar{X}_2)' \left(\frac{S_1}{n_1(n_1 - 1)} + \frac{S_2}{n_2(n_2 - 1)}\right)^{-1} (\bar{X}_1 - \bar{X}_2). \quad (3)$$

In fact, as the sample size is large, the Hotelling's T^2 statistic approximately subjects to a Chi-square distribution; otherwise, it approximately subjects to an F distribution [8].

In statistics, the p value of a statistical test provides a measure of the strength of evidence that the sample is in favor of the null hypothesis. It is the probability of getting a result as extreme or more extreme than the one observed if the proposed null hypothesis is correct. A p value closing to zero means that the null hypothesis is false, and a difference is very likely to exist between the two populations. On the other hand, a large p closing to 1 implies that there is no detectable difference through the sample used. By the definition of p value, we can use it to measure the goodness of classification on a gene set. Actually, our informative gene selection method uses a significance level α on the statistical tests. That is, the genes are selected as the informative ones if the p values of the statistical tests are less than α.

With above preparations, we can now present the post-filtering gene selection algorithm. By conducting certain statistical tests on the gene expression profiles, we assume that a gene set with a smaller α value contains more information about the class than a gene set with a larger p value. We introduce a kind of approximate Markov blanket based on statistical tests as follows:

Definition 1 (Approximate Markov blanket) *For two gene sets GS_i and GS_j $(i \neq j)$, GS_i forms an approximate Markov blanket for GS_j $(i \neq j)$ iff $p(GS_i) < p(GS_j)$ and $p(GS_i) < p(GS_i \cup GS_j)$.*

Here $p(GS)$ is the p value of a 2-sample statistical test on the homogeneity of two types of the expression profiles throughout the gene set GS. If the gene set contains only one gene, we take a 2-sample univariate statistical test; otherwise, we take a 2-sample multivariate statistical test. The null hypothesis is that the two populations of the tumor and normal samples are identical in statistics. Our aim is to check whether the expression profiles on a gene set are different between the normal and tumor samples. If GS_i forms an approximate Markov blanket for GS_j, we consider that GS_i can provide more information about the classification than the combination of GS_i and GS_j.

We summarize the post-filtering gene selection algorithm in Table 1. There are three predefined parameters. α_0 is the significance level of the statistical test to determine the genes that will be considered by the post-filtering algorithm. The larger α_0 is, the more genes will be processed. α is also a significance level of the statistical tests which restricts the goodness of classification through the gene set. β is a threshold value of the correlation measure. When β is too small, we will remove more genes at risk of losing some useful information. On the contrary, if β is too large, our approximate Markov Blanket based on the statistical tests is not powerful enough to remove sufficiently redundant genes. Usually, we set β in [0.5, 0.8]. *UTest* in the algorithm is a kind of univariate statistical test, and *MTest* is a kind of multivariate statistical test.

Table 1: The post-filtering gene selection algorithm

1. Input : $G, L, \alpha_0, \alpha, \beta$
2. Output : GS_{best}
3. Order the gene sets:
4. *if* $s == 1$
5. Calculate every $P(G_i) = UTest(G_i)$ for all $G_i \in G$
6. *if* $(P(G_i) < \alpha_0)$, append G_i to G_{list};
7. Order G_{list} in ascending p values;
8. *if* $s > 1$ *and* $s <= maxs$
9. $G_{list} = NULL$;
10. *begin*
11. Calculate $\min P(G_i; G_j) = \min(Mtest(G_i; G_j))$ for all $G_i, G_j \in RG_{list}$
12. remove G_i and G_j from RG_{list}, append $G_{i;j} = G_i \cup G_j$ to G_{list};
13. *end until*$(RG_{list} == NULL)$;
14. Redundancy analysis :
15. $G_j = getFirstElement(G_{list})$;
16. *if* $P(G_j) > \alpha$, append G_j to RG_{list}
17. *else*
18. *begin*
19. $G_k = getNextElement(G_{list}, G_j)$;
20. *begin*
21. *if* $(C(G_j; G_k) > \beta)$, remove G_k from G_{list}
22. *elseif* $(P_j < P_{j;k})$, remove G_k from G_{list}, append G_k to RG_{list};
23. $G_k = getNextElement(G_{list}, G_k)$
24. *end until*$(G_k == NULL)$
25. $G_j = getNextElement(G_{list}, G_j)$
26. *end until*$(G_j == NULL)$;
27. Add G_{list} to GS_{best};
28. $s = s + 1$, go to order the gene sets.

The post-filtering gene selection algorithm involves *maxs* iterations ($s = 1, \cdots, maxs$). Every iteration is composed of two sequential steps. 1. Order the gene sets: for the first iteration, this step calculates the p values of a univariate statistical test (line 5) for every gene from G, selects genes whose p values are less than α_0, and appends these genes to G_{list}, a list of gene sets, in a ascending order of their p values. In the following iterations, this step calculates a multivariate statistical test (line 11) for every pair gene sets from RG_{list}, a list of gene sets that are temporarily removed during the previous iteration. Then, this step removes the pair with minimum p value from RG_{list} and appends them as one element to G_{list}. The process repeats until RG_{list} is empty. 2. Redundancy analysis by evaluating classification of a combination of 2^{s-1} genes: the filtering is based on a predefined threshold β of the correlation measure (line 21) and the approximate Markov blanket. Through this step, every newly selected gene set satisfies three criteria: 1). The p value of the statistical test is less than α. 2). It is not highly correlated with other gene set selected already. 3). It does not have an approximate Markov blanket in the selected gene sets. Those genes in the sets satisfying all the three criteria will be added into GS_{best} as the result of the s-th iteration. Those gene sets violating criterion 1 or 3 will be added into RG_{list} and be processed in the $(s+1)$-th iteration which will evaluate larger gene sets by combining two smaller ones from RG_{list}, and select those gene sets satisfying the same three criteria. We can stop the algorithm when no more informative

genes can be found in an iterative way. Alternatively, for simplicity, the algorithm stops when it reaches the user-specified maximum number of iterations.

III. Experimental Results and Comparisons

We tested the effectiveness of our post-filtering gene selection algorithm on the two real data sets through the support vector machine as follows:

Colon cancer data set: The colon

n results as our evaluation result. In this experiment, we set $\alpha_0 = \alpha$ and used the 2-sample rank sum test. From Table 2, the best prediction accuracy achieved by using the post-filtering algorithm was 98.4%, where only one prediction error happened in the LOOCV experiment.

In order to illustrate the influence of the correlation coefficient threshold β under the same significance level, we gave the classification accuracies with different correlation coefficient thresholds. To utilize all the information of a microarray data set, we selected genes on the whole data set before conducting LOOCV in this and the following experiments. We conducted LOOCV experiments using 2-sample t tests after selecting genes on the whole colon data set. From Table 3, we can find out that our algorithm achieved better prediction accuracies when β is 0.5 or 0.6 on the colon data set.

Table 3. The influence of the threshold β on the prediction accuracy under the same significance level

M & P	β	0.3	0.4	0.5	0.6	0.7	0.8
$\alpha_0 = 0.03$	1	85.4%/5	91.9%/4	90.3%/8	95.2%/13	93.5%/23	93.5%/62
$\alpha = 0.03/315$	2	88.7%/7	95.2%/10	96.8%/18	96.8%/27	93.5%/41	93.5%/94
$2-sample\ t\ test$	3	90.3%/11	95.2%/10	96.8%/18	96.8%/31	95.2%/53	93.5%/102

When we set $\alpha_0 = 0.015$ and used the 2-sample Kolmogorov-Smirnov (KS) test to do the conventional gene selection on the colon data set, 177 genes were selected and the corresponding prediction accuracy is 93.5%. These genes were to be further filtered by the post-filtering algorithm under several different significance levels. From the results shown in Table 4, the post-filtering algorithm can achieve better prediction accuracies by removing redundant genes. As a matter of fact, all the best prediction accuracies reached 98.4% under these significance levels shown in Table 4.

Table 4. The LOOCV results on the colon data set with the *KS* test.

M & P	α	0.001	0.003	0.005	0.007	0.009
$\alpha_0 = 0.015$	$s=1$	93.5%/5	93.5%/6	93.5%/6	95.2%/7	95.2%/7
$\beta = 0.6$	$s=2$	96.8%/19	95.2%/20	98.4%/22	98.4%/23	98.4%/23
$KS\ test$	$s=3$	98.4%/27	98.4%/24	98.4%/26	96.8%/27	96.8%/27

α_0 set a significance level which was only used in the first iteration of the algorithm. We fixed $\alpha = 0.005$ and used the 2-sample Kolmogorov-Smirnov test to do the experiment on the colon data set. We gave the prediction accuracies over the different α_0 in Fig.1. We calculated the average and the maximum prediction accuracies of 3 iterations of the post-filtering algorithm. Meanwhile, Figure 1 also illustrated the prediction accuracies before filtering. By comparison, best prediction accuracies were achieved when α_0 was 3-5 times of α on the colon data set, which was shown in Fig. 1.

We also conducted an experiment on the leukemia data set. Because of the large scale, we took the 4-fold cross validation [4] instead of LOOCV on this data set. From Table 5, we can find that the post-filtering algorithm reached a perfect result. The two classes can be totally separated with only 9 genes.

Fig. 1. The LOOCV results with different α_0 on the colon data set.

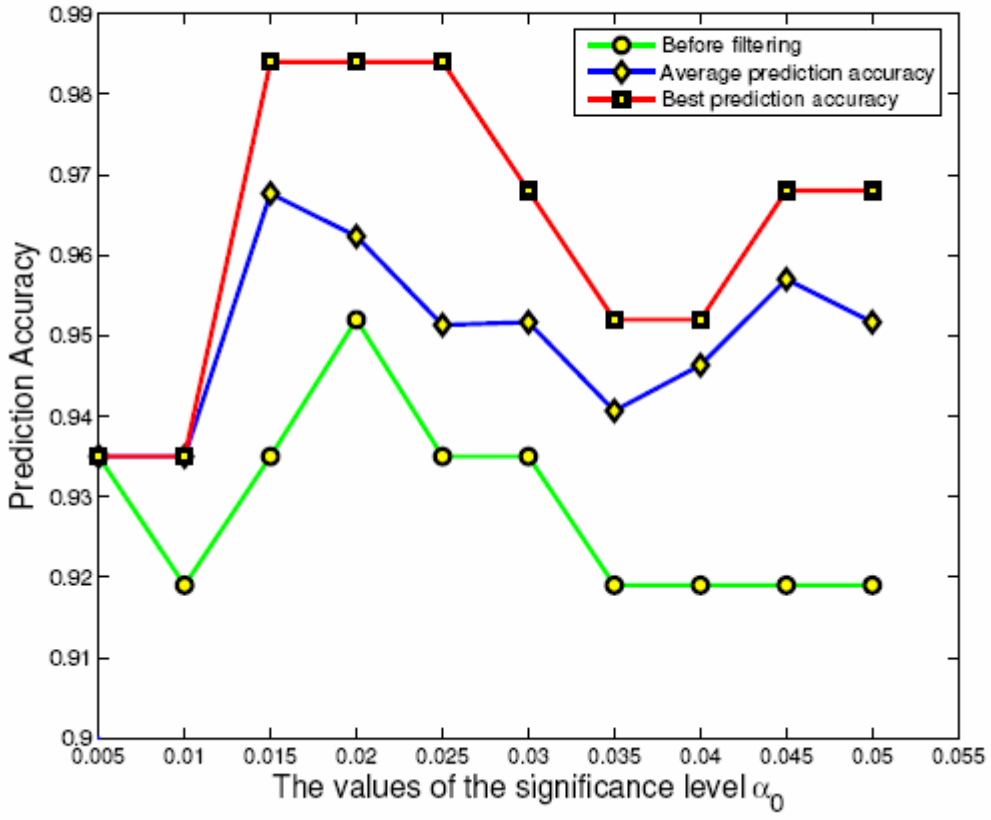

Table 5. The 4-fold cross-validation results on the leukemia data set with the rank-sum test.

M & P	α	0.0000001	0.000001	0.00001	0.0001	0.001
$\beta = 0.5$ *Rank sum*	$s = 0$	98.6%/53	98.6%/93	98.6%/163	95.8%/293	77.8%/516
	$s = 1$	98.6%/7	100%/9	100%/13	100%/19	100%/21
	$s = 2$	100%/9	100%/11	100%/21	100%/49	98.6%/79
	$s = 3$	100%/9	100%/11	100%/21	100%/57	98.6%/107

IV. Further Discussions and Remarks

From the above experiments, we can find that the

generalization of the constructed classifier. As a result, we should seek a balance between the two aspects at selecting α. From the results of the experiments, 0.005 or 0.01 as the significance level α works rather good on both the colon and leukemia data sets. As for the selection of α_0, we should balance the computing complexity and the required information of classification. Generally, it should be 3-5 times of α.

By using statistical tests, our post-filtering method for informative gene selection is not merely a kind of algorithm. It can be considered as a basic framework. Actually, many other statistical tests and methods can be also applied to this approach. To form an approximate Markov Blanket, we can use any statistical test with a nu

[6] E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," *Proceedings of the 18th International Conference of Machine Learning (ICML'01)*, Massachusetts, USA, June 28-July 1, 2001, pp. 601-608.

[7] T. Bo and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," *Genome Biol.*, vol.3, pp. RESEARCH0017.1--0017.11, 2002.

[8] K. Fang, "Applied Multivariate Statistical Analysis" (in Chinese), Shanghai: East China Normal University Press, 1989, pp. 105-141.

[9] D. S. Dudoit, J. Fridyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumor using gene expression data," *Univ. of California, Dept. of Statistics, Tech Report*, no.576, 2000.

[10] C. Hsu, C. Chang, and C. Lin, "A Practical Guide to Support Vector Classification," *National Taiwan University, Dept. of Computer Science and Information Engineering* posted in the web: www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.