Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-191

Learning Differentially Expressed Gene Pairs in Microarray Data

Xiao-Lei XIA[a], Sinead BROPHY[b], and Shang-Ming ZHOU[b,1]

[a] *School of Mechanical and Electrical Engineering, Jiaxing University, Jiaxing, P. R. China, 314001*

[b] *Farr Institute of Health Informatics Research, Swansea University Medical School, Swansea, SA2 8PP, UK*

Abstract. To identify differentially expressed genes (DEGs) in analysis of microarray data, a majority of existing filter methods rank gene individually. Such a paradigm could overlook the genes with trivial individual discriminant powers but significant powers of discrimination in their combinations. This paper proposed an impurity metric in which the number of split intervals for each feature is considered as a parameter to be optimized for gaining maximal discrimination. The proposed method was first evaluated by applying to a synthesized noisy rectangular grid dataset, in which the significant feature pair which forms a rectangular grid pattern was successfully recognized. Furthermore, applying to the identification of DEGs on colon microarray data, the proposed method demonstrated that it could become an alternative to Fisher's test for the prescreening of genes which led to better performance of the SVM-RFE method.

Keywords. Differentially expressed genes, Microarray data, Gene interactions, Machine learning

1. Introduction

In the analysis of microarray data, one of the most important tasks is the identification of differentially expressed genes (DEGs). Due to the large number of genes, univariate ranking methods have been widely employed, which can be divided into two categories: parametric approaches and model-free ones. The former category, epitomized by the *t*-test and ANOVA, assumes an underlying distribution that the samples are drawn from. Model free methods, in contrast, circumvent the assumption about data generation. To detect the DEGs, some methods have used a subset of the training sets, in which permutation of samples was implemented in order to prevent the false positive error from accumulating due to multiple testing. These methods include approaches bounding the "Family-Wise Error Rate" (FWER) which is the overall chance of one or more false positives and those controlling the "False Discovery Rate" (FDR) which is the expected percentage of false positives among the genes deemed as differentially expressed [10]. To identify important genes, some researchers have proposed gene selection methods, such as the gene pair selection approach [3], correlation-based approaches [11], Markov blanket filtering [6], minimum redundancy maximum

[1] Corresponding author, Farr Institute of Health Informatics Research, Swansea University Medical School, Swansea, SA2 8PP, UK; E-mail:s.zhou@swansea.ac.uk.

relevance [5] and uncorrelated shrunken centroid [12]. The rationale behind this scheme is that a good feature subset is highly correlated with the class and uncorrelated with each other[11]. The above algorithms as filter methods perform feature selection independently of a classifier. In contrast, wrapper feature selection methods use a classifier to evaluate a feature subset from which a classifier is trained. A number of heuristic search strategies are thus proposed, such as, estimation of distribution, sequential search, genetic algorithms [9], as well as a incremental augmenting search scheme preceded by univariate gene ranking. In gene selection, embedded methods, on the other hand, use the intrinsic property of a specific classifier to evaluate feature subsets, such as random forest induced approaches [4] and algorithms measuring the importance of genes by weight vectors respectively yielded by Support Vector Machines (SVM), known as the SVM-Recursive Feature Elimination (RFE) algorithm [7] and logistic regression [8].

However, the majority of current methods for the detection of DEGs tend to rank genes individually. The problem is that such a paradigm is very likely to dismiss the genes whose discriminant powers are trivial individually but significant jointly, as exemplified by the rectangular grid dataset contaminated by with different noise levels. Thus, in this paper we proposed an impurity metric which can efficiently identify gene pairs with good generalization performances. The novelty of this metric is that the values of each feature are split into intervals while the number of split intervals is considered as a parameter to be optimized for gaining maximal discrimination. The significance of a feature pair is measured by the sum of correctly classified training samples across all the grids on the plane. The more significant a feature pair is, the larger the sum is. The method was compared with the traditional Fisher's ratio test in the contexts of identification of DEGs. Experiments on the colon dataset [1] demonstrated that our proposed method could become an alternative to Fisher's test for the prescreening of genes which led to better performance of the SVM-RFE method.

2. Methods

Our method starts with splitting each feature into multiple intervals. Then, a pair of features divide the data into a specific number of rectangular grids. For the training samples, these grids contain either the samples from multiple classes, or the samples from a sole class, or no samples. Assuming the total number of training samples within a grid is n, among which m samples are from the positive class. The number of correctly classified samples is the maximum of m and $(n - m)$. The significant feature pair is defined as the pair that optimise the problem Eq. (1):

$$\max_{(i,j)} \sum_{k=1}^{(\#vi)^2} \Delta(i,j,k) \tag{1}$$

where i and j are the indexes of the feature pair. Representing the number of value interval for each feature by $\#vi$, $\Delta(i, j, k)$ is the number of correctly-classified samples of the k-th grid among the $(\#vi)^2$ grids that features i and j divides the data into. So the significance of a feature pair is measured by the sum of correctly-classified training samples across all the grids. The more significant a feature pair is, the larger the sum is. The least significant a feature is, the closer the sum is to half of the total number of training samples. The rationale behind the algorithm lies in our novel perspective on the binary classification process, with which the input space is partitioned into a specific number of disjoint grids. Each grid carries a specific label which indicates the class of all, or the majority, of the inclusive training samples.

Grid search is employed to find the optimal settings of the parameters. In model selection, each classifier is trained with the selected hyper-parameters, in which the model performance is evaluated on validation data. The grid search selects the settings of the hyper-parameters that achieved the highest score in the validation procedure.

The proposed metric is similar, in terms of its methodology, to the well-known metric of gini impurity. But gini impurity is developed in the framework of decision trees and thus the range of each feature is split into two intervals, while our method allows for multiple value intervals for each feature.

3. Results of Prescreening Genes for Microarray Data

The proposed method was employed as a filter method to prescreen genes in microarray data for the identification of significant feature pair, in which the SVM with selected genes was used to classify the samples. The performance of our proposed method was compared with that of Fisher's ratio test in the SVM-RFE in terms of the prediction accuracy of the SVM established upon the selected genes. In our method, each feature pair was assigned a score as the total number of correctly-classified training samples. Assuming n genes for a microarray dataset which results in $n(n-1)/2$ pairing scores, five alternative ranking strategies were proposed:

Strategy 1: *A gene's rank is decided by the mean of the* $(n-1)$ *scores from pairing the gene with the remaining* $(n-1)$ *gene respectively.*

Strategy 2: *Among all the* $n(n-1)/2$ *pairs, find the genes pairs whose scores are among the highest. The union of these gene pairs is used as the candidate gene set.*

Strategy 3: *Select the genes pairs with the highest scores. Then, with a gene being included in one particular gene pair, the gene pairs with lower scores will be removed. The union of the resultant set of gene pairs is taken as the candidate set.*

Strategy 4: *Among all the* $n(n-1)/2$ *pairs, find the genes pairs with the lowest scores. The union of these gene pairs is excluded from further analysis.*

Strategy 5: *Select the genes pairs with the lowest scores. Then, with a gene being included in one particular gene pair, the gene pairs with higher scores will be kept. The union of the resultant set of gene pairs is excluded from further analysis.*

For microarray data with normal features less than 100 samples, it is highly recommended to employ the bootstrap resampling technique with replacement for an unbiased estimate [2]. The overall classification performance is the average of the performances on the resampled sets.

In this study, the colon dataset contains the expression values of 2000 genes with highest minimal intensity from 62 tissues. The identity of the 62 tissues is given in file tissues. There are 22 normal tissues and 40 cancerous tissues [1]. The data was subjected to base 10 log transformation, followed by each scaled to the value range of $[-1, 1]$. The five strategies and Fisher's ratio were respectively applied to the data to reduce the number of candidate genes to 1000. The three subplots for Figure 1 from left to right corresponds to the setting of $\#vi$ at 2, 3 and 4 respectively for our proposed method. In each subplot, the six different filtering methods were highlighted in different colors.

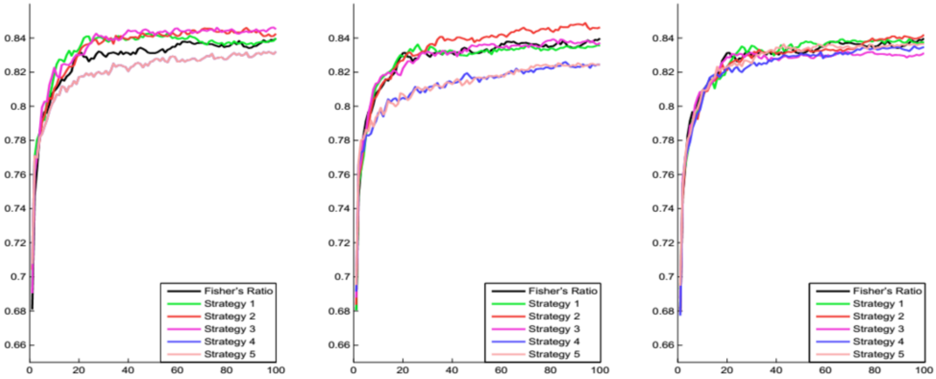

Figure 1. The performance of the SVM as a function of the number of DEGs with the regularization parameter $C = 10$: The x-axis is the number of DEGs and the y-axis represents the accuracy of the SVM classification.

Table 1 gives the top 48 pair of DEGs selected by Strategy 2 which corresponds to the red solid line in the leftmost subplot of Figure 1. In Table 1, the 1st column corresponds to the pairs with the top 12 ranking and the last three columns list a dozen of pairs respectively rank 13-th~24-th, 25-th~36-th,37-th~48-th.

Table 1. The top 48 pairs of DEGs selected by Strategy 2

1st dozen	2nd dozen	3rd dozen	4th dozen
⌊Hsa.692, Hsa.549⌋	⌊Hsa.1972, Hsa.2645⌋	⌊Hsa.467, Hsa.562⌋	⌊Hsa.6080, Hsa.957⌋
⌊Hsa.823, Hsa.37937⌋	⌊Hsa.541, Hsa.692⌋	⌊Hsa.878, Hsa.8125⌋	⌊Hsa.4689, Hsa.652⌋
⌊Hsa.8147, Hsa.698⌋	⌊Hsa.678, Hsa.9235⌋	⌊Hsa.7877, Hsa.8223⌋	⌊Hsa.5398, Hsa.459⌋
⌊Hsa.831, Hsa.608⌋	⌊Hsa.832, Hsa.7728⌋	⌊Hsa.36694, Hsa.21562⌋	⌊Hsa.6039, Hsa.5392⌋
⌊Hsa.853, Hsa.8374⌋	⌊Hsa.580, Hsa.6472⌋	⌊Hsa.7498, Hsa.579⌋	⌊Hsa.2361, Hsa.6288⌋
⌊Hsa.688, Hsa.462⌋	⌊Hsa.2950, Hsa.36689⌋	⌊Hsa.951, Hsa.41208⌋	⌊Hsa.33965, Hsa.2610⌋
⌊Hsa.442, Hsa.662⌋	⌊Hsa.5971, Hsa.2715⌋	⌊Hsa.81, Hsa.421⌋	⌊Hsa.4252, Hsa.1454⌋
⌊Hsa.451, Hsa.692⌋	⌊Hsa.9972, Hsa.2097⌋	⌊Hsa.8052, Hsa.6814⌋	⌊Hsa.41282, Hsa.2996⌋
⌊Hsa.2357, Hsa.2928⌋	⌊Hsa.61, Hsa.45658⌋	⌊Hsa.24944, Hsa.57⌋	⌊Hsa.24506, Hsa.1410⌋
⌊Hsa.821, Hsa.6317⌋	⌊Hsa.316, Hsa.3331⌋	⌊Hsa.773, Hsa.1254⌋	⌊Hsa.3007, Hsa.29913⌋
⌊Hsa.42186, Hsa.8214⌋	⌊Hsa.36952, Hsa.960⌋	⌊Hsa.2654, Hsa.2821⌋	⌊Hsa.3306, Hsa.612⌋
⌊Hsa.3305, Hsa.6317⌋	⌊Hsa.8175, Hsa.627⌋	⌊Hsa.2471, Hsa.404⌋	⌊Hsa.3135, Hsa.3083⌋

It can be seen from Figure 1 that, with the setting of $\#vi = 2$, strategies 1, 2 and 3 all outperformed the Fisher's ratio. And with $\#vi = 2$, Strategies 2 and 3 still remained superior to Fisher's ratio. With $\#vi = 4$, the performances of the six methods were more or less the same although strategy 1 was slightly the best.

4. Discussions

The proposed metric assessed the importance of gene pairs in terms of the sum of correctly classified training samples across all the grids. Although standard grid search suffers from the curse of dimensionality, in this study the number of split intervals for each gene does not need to be too big. By rule of thumb, for a training dataset with n samples, the feasible range of #vi could be [2, √n]. In this study, the $\Delta(i, j, k)$ increases when the number of grids goes up from 4 (corresponding to $\#vi = 2$) to 16(corresponding to $\#vi = 4$), while decreasing for the number of grids from 16 (corresponding to $\#vi = 4$) to 36 (corresponding to $\#vi = 6$). So as a rule of thumb, the optimal $\#vi$ should be either $\#vi = 4$ or $\#vi = 5$.

5. Conclusions

In order to identify the significant feature pairs, this paper proposed an impurity metric to assess the significance of a feature based on the "purity" of the samples for the resultant subproblems. The advantage of the proposed method is that the number of value intervals for a feature is treated as a parameter to be optimized. In the identification of DEGs for microarray data as a filter strategy, the proposed method has demonstrated better performance than Fisher's ratio method.

References

[1] U. Alon, N. Barkai, D. A Notterman, K. Gish, S. Ybarra, D. Mack, and A. J Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999.

[2] C. Ambroise and G.J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10):6562, 2002.

[3] T. Bo and I. Jonassen. New feature subset selection procedures for classification of expression profiles. *Genome Biology*, 3(4):0017, 2002.

[4] R. Díaz-Uriarte and A. de Andrés. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.

[5] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology, 3(2):185–206, 2005.

[6] O. Gevaert, F.D. Smet, D. Timmerman, Y. Moreau, and B.D. Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, 22(14), 2006.

[7] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422, 2002.

[8] S. Ma and J. Huang. Regularized roc method for disease classification and biomarker selection with microarray data. *Bioinformatics*, 21(24):4356–4362, 2005.

[9] C.H. Ooi and P. Tan. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, 19(1):37–44, 2003.

[10] V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[11] E.J. Yeoh, M.E. Ross, S.A. Shurtleff, W.K. Williams, D. Patel, R. Mahfouz, F.G. Behm, S.C. Raimondi, M.V. Relling, A. Patel, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–143, 2002.

[12] K. Yeung and R. Bumgarner. Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome Biology*, 4(12):R83, 2003.