PLS-based Characteristic Selection and Identification of Gene Expression Profiles

Yong Zeng[1], Xiaohui Wu[1], Meishuang Tang[2], Guoli Ji[1]

[1]*Department of Automation, Xiamen University, Xiamen, Fujian, 361005, China, glji@xmu.edu.cn*

[2]*Modern Educational Technical and Practical Training Center, XiamenUniversity, Xiamen 361005, China*

In genetics, gene expression is one of the most fundamental level at which the genotype gives rise to the phenotype. While, the gene expression profiles are always characteristics of high-dimensional, small sample, strong relevance, and high noise. The gene expression analysis always including the differential analysis, class prediction (supervised learning), class discovery (unsupervised), and pathway analysis. Here we focus on the class prediction, in which searches for informative genes predict significant phenotype membership. Although, there are also several classification methods have been proposed, such as classification and regression trees (CART), K-nearest neighbors (KNN), probabilistic neural network (PNN), Weighted Voting, and Support Vector Machines (SVM), few of them can handle the high dimensional and small samples (HDSS) problem efficiently. Hence, we proposed a partial least squares (PLS) based gene-selection method, which synthesizes genetic relatedness and is suitable for multicategory classification.

Partial least squares(PLS) regression is a statistical method to find a linear regression model by projecting the predicted variables and the observable variables to a new space. Given the high-related character between genes, we took the joint distribution of gene into account, and a new filter-based method of global gene selection, where each specific gene is extracted based on all sample genes in the input domain, is proposed. Furthermore, using explanation difference of independent variables on dependent variable (class), we defined three indicators, which are independent-variable explanation gain (IEG), dependent variable explanation gain (DEG), and variant importance in projection (VIP). Based on the aforementioned preparation, the PLS-based global gene selection algorithm was presented. It can detect those genes with a relatively small main effect, but with a strong interaction effect. The procedure of our PLS-based global gene selection mwthods are shown in the TABLE1.

TABLE 1. The procedure of PLS-based Gene Selection Algorithm

Step 1	*nfac*=0, k=0, max_nfac=g(number of category), max_k=100.
Step 2	Calculate the value of three PLS-based indicators(VIP, IEG, DEG) for each feature in the traning set.
Step 3	Select the top k values on PLS-index in the training set for SVMs classify learning.
Step 4	Classify the testing set by SVMs using the top k selected features, and calculate the recognition rate.
Step 5	k=k +1, if k<max_k, goto Step 3.
Step 6	Nfac=nfac+1; if nfac<=max_nfac, repeat Step 2 to Step 5.
Step 7	Maximize the classifier accuracy and at the same time minimize k from the max_nfac*max_k results.

In addition, a comparison with the stat-of-the-art methods was also implemented based on the Benchmark database, and several indexs were defined to measure the performance of the method. We may safely draw the conclution that our algorithm is computationally efficient especially for high-dimensional dataset, and it can be applied to both two-category classification and multi-category classification problems without limitation.

Furtthermore, we set out to introduce the multi-pertubation mechanisms to improve the fidelity of the analysis result for small samples, and to apply our methods to the salinity adapability study of three-sipine stickleback under different salinity environments, which will help to figure out informative genes for specific salinity level or tissue on genome-wide scale, and the result of which will be of huge interest to aquatic ecological physiologists.

Acknowledgements: This project was funded by the National Natural Science Foundation of China (Nos.61174161, 61201358 and 61203176), the Natural Science Foundation of Fujian Province of China (No. 2012J01154), the specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20120121120038), the Key Research Project of Xiamen City of China (No. 3502Z20123014), the Fundamental Research Funds for the Central Universities in China (Xiamen University: Nos. 2011121047、201112G018 and 201212G005), and the Fundamental Research Fund for the university student Creative and Entrepreneurship training program in China(Xiamen University: No. XDDC201210384063).