Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

SDED: A novel filter method for cancer-related gene selection

Wenlong Xu[1], Minghui Wang[2], Xianghua Zhang[1], Lirong Wang[1], Huanqing Feng[1,*]

[1]Department of Electronic Science and Technology, University of Science and Technology of China, Hefei 230027, China; [2]The College of Life Science and Bio-engineering, Beijing University of Technology, Beijing 100022, China;
Huanqing Feng* - E-mail: hqfeng@ustc.edu.cn; Phone: 86 551 3601800; * Corresponding author

received December 19, 2007; revised March 13, 2008; accepted April 04, 2008; published April 11, 2008

Abstract:
Gene selection is to detect the most significantly expressed genes under different conditions expression data. The current challenge in gene selection is the comparison of a large number of genes with limited patient samples. Thus it is trivial task in simple statistical analysis. Various statistical measurements are adopted by filter methods applied in gene selection studies. Their ability to discriminate phenotypes is crucial in classification and selection. Here we describe the standard deviation error distribution (SDED) method for gene selection. It utilizes variations within-class and among-class in gene expression data. We tested the method using 4 leukemia datasets available in the public domain. The method was compared with the GS2 and CHO methods. The Prediction accuracies by SDED are better than both GS2 and CHO for different datasets. These are 0.8-4.2% and 1.6-8.4% more that in GS2 and CHO. The related OMIM annotations and KEGG pathways analyses verified that SDED can pick out more 4.0% and 6.1% genes with biological significance than GS2 and CHO, respectively.

Keywords: gene selection; filter method; support vector machine; SDED

Background:
DNA micro-array technology has enabled biologists to associate phenotypes with molecular genetics **[1, 2]**. It is commonly used to compare gene expression levels of different phenotypes (normal versus cancer). It enables the study of thousands of gene expression simultaneously. The difficulty is in interpreting expression data. Genes with significant expression across the sample set are selected using sound statistical techniques. These discriminatory genes will help to classify different cancer subtypes **[3, 4]**. There are two categories of gene selection strategies namely, filter and wrapper **[1]**.

Many filter methods have been proposed by eliminating redundant genes. Golub *et al.* **[5]** (1999) provided a signal-to-noise statistic method for binary classification. Baldi and Long **[6]** (2001) proposed multivariate test statistic to identify differentially expressed gene combinations. Cho *et al.* **[7]** (2003) used a new statistic method considering within-class variation (CHO). Yang *et al.* **[8]** (2006) used a stable gene selection in micro-array data analysis (GS2). In wrapper methods, genes are tested in groups according to their performance in the classification model. Xiong *et al.* **[9]** suggested a method to select genes through the space of feature subsets using classification errors. Guyon *et al.* **[10]** proposed a gene selection approach utilizing Support Vector Machines (SVM) based on recursive feature elimination.

Both categories of gene selection strategies have their disadvantage. Although GS2 is a stable method, calculations are too complex and the biological meaning is difficult for annotation. The CHO method considers within-class information and it loses the among-class information. The wrapper methods use exponentially increasing dimensions of the feature space for large gene sets. Thus, the wrappers are computationally intractable for high-dimensional gene data **[1]**. The inherent linear nature is their disadvantage and it makes it difficult to identify important genes in wrapper methods **[11]**. Here, we propose a statistical measurement to better score genes with subtle expression patterns. It incorporates the within/among class variations in gene expression data.

Methodology:
Datasets
MLL dataset:
We used the MLL dataset from the KORSMEYER Laboratory **[12]**, which containing 72 samples in three classes: (1) acute lymphoblastic leukaemia (ALL); (2) acute myeloid leukaemia (AML); and (3) mixed-lineage leukaemia (MLL), which has 24, 28, 20 samples, respectively. Each sample contains 12,582 gene expression values.

ALL-AML dataset:
The ALL-AML dataset is obtained from the cancer program of BROAD Institute **[13]**. It consists of 7129 gene expression profiles of two acute cases of leukaemia: (1) acute lymphoblastic leukaemia (ALL, 47 samples) and (2) acute myeloblastic leukaemia (AML, 25 samples). The ALL dataset is obtained from B-cell (ALL-B, 38 samples) and T-cell (ALL-T, 9 samples) and the AML is obtained from bone marrow

(AML-BM, 21 samples) and peripheral blood (AML-PB, 4 samples) samples. Due to the bipartition of each component, it can be treated both as a three-class dataset (ALL-B, ALL-T and AML) and as a four-class dataset (ALL-B, ALL-T, AML-BM and AML-PB). Here, the three-class version is referred to as ALL-AML-3 and the four-class version as ALL-AML-4.

ALL dataset:
The ALL dataset by St. Jude Children's Research hospital **[14]** contains 248 samples in six classes of subtype ALL: (a) TEL, (b) Hyper, (c) T, (d) E2A, (e) MLL, and (f) BCR, which contains 79, 64, 43, 27, 20 and 15 samples, respectively. Every sample contains 12,625 gene expression values.

Data normalization
These 4 datasets were used in the analyses. Each sample was normalized to standard distribution - N(0,1) before scoring for gene selection. The expression of each gene was normalized based on the expression level in each sample.

SVM classifier
SVM is a powerful and popular machine-learning method and has been widely used in biological classification. The key idea of SVM is to maximize the margin separating the two classes while minimizing the total classification error. There were a number of kernels used in SVM models for decision plane computing and the radial basis function (RBF) kernel was chosen for our purpose. As for the design of multi-class SVM classifier, we used the one-versus-one method. The final prediction decision was given by the voting strategy: the predicted class is assigned to the one that has the maximum vote. If more than one class has the same maximum vote, the classifier will have to make a random prediction. It is known that proper selection of parameter is very important for SVM, so the grid search strategy by Chih-Jen Lin **[15]** was performed to find the best combination of parameters for each prediction process. The toolkit for SVM implementation we used in MATLAB was LIBSVM-Version 2.82 **[15]**.

Discussion:
Samples are first divided into testing and training data for each dataset. We used the training samples for scoring the genes. The quality of these top ranked x genes are selected based on two aspects, namely: (1) the classification accuracy; (2) relevance to relative inheritance or diseased association in related pathways.

Classification accuracies
We used the top ranked genes selected by a gene selection method, together with their expression values in the training dataset to build a classifier for each testing sample. We defined the classification accuracy as the percentage of correct decisions made by the classifier on the testing samples. We adopted the SVM classifier to compare the performance of SDED with GS2 and CHO. The classification accuracy was obtained through the leave one out cross validation (LOO_CV) process. One sample was taken as testing and the remaining were used as training data in LOO_CV. This is done for all samples and for every top ranked x (from 1 to 100 with $p < 0.01$) genes in the datasets.

Figure 1 shows the plot for classification accuracy of the SVM classifier based on SDED, GS2 and CHO on MLL dataset. The SDED method could achieve better results than GS2 (94.444%/91, 97.222%/48, 93.056%/36), CHO (88.889%/82, 95.833%/74, 93.056%/69) for MLL, ALL-AML-3 and ALL-AML-4 datasets. The SDED showed 97.222%/48, 98.611%/16, 97.222%/57, accuracy for these datasets even with less number of genes, respectively. The performance of SDED method (98.387%/96) was only comparable with GS2 (97.581%/68) and CHO (96.774%/87) in ALL dataset. In summary, the SDED filter method can perform about 0.8-4.2% and 1.6-8.4% better classification accuracies than GS2 and CHO, respectively.

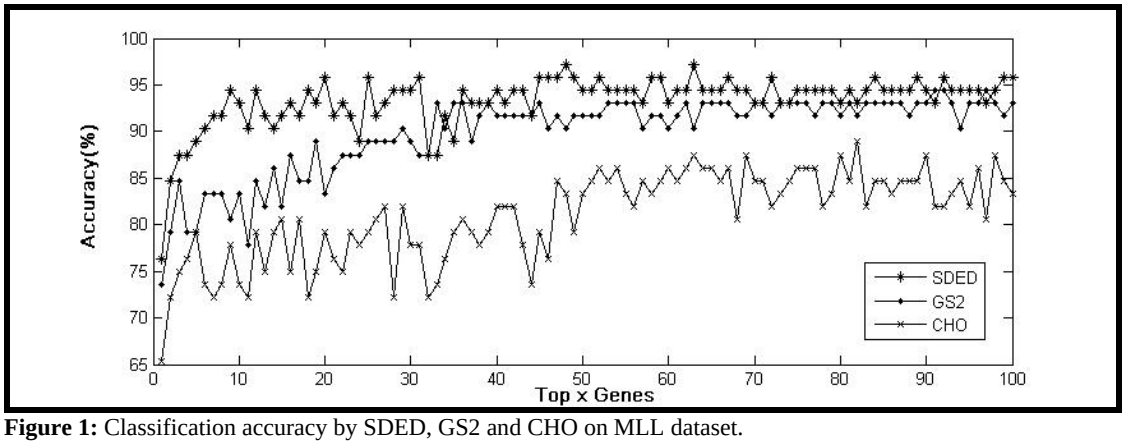

Figure 1: Classification accuracy by SDED, GS2 and CHO on MLL dataset.

Biological meaning
We examined genes and their association in pathways to demonstrate the biological significance and evidence of gene selection. The top 100 ranked genes were chosen ($p < 0.01$) for each method and dataset. The numbers of genes in the dataset that are found in

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

OMIM (Online Mendelian Inheritance in Man) and KEGG Pathways were listed in Table 1 (see supplementary material). The SDED method helped to select more genes compared to other methods in ALL_AML_3, ALL_AML_4 and ALL datasets. It selected about 4.0% (570/800 versus 538/800) and 6.1% (570/800 versus 521/800) genes with biological significance than GS2 and CHO, respectively.

Conclusion:

In this paper, we described an effective gene selection method named SDED. The method was tested using 4 leukaemia datasets and compared with the GS2 and CHO methods. The described SDED method achieved 0.8-4.2% and 1.6-8.4% better classification than GS2 and CHO, respectively. The related OMIM annotation and KEGG pathways analyses verified that SDED method can pick out more genes with biological significance.

Acknowledgement:

The authors are grateful to Dr. Ao Li, Dr. Peng Qiu and Dr. Yin Liu for their help with preliminary data analysis, results discussion, critical reading and English writing. This work was financially supported by Innovation Center for Postgraduates at HFNL (Hefei National Laboratory for Physical Sciences at the Microscale), USTC (no. C07-05).

References:

[01] Y. Liang & A. Kelemen, *Funct Integr Genomics*, 6: 1 (2006) [PMID: 16292543]
[02] T. Hanai, *et al.*, *J Biosci Bioeng.*, 101: 377 (2006) [PMID: 16781465]
[03] K. Hoffmann *et al.*, *BMC Cancer*, 6: 229 (2006) [PMID: 17002788]
[04] P. Qiu *et al.*, *Bioinformatics*, 2005, 21: 3114 (2005) [PMID: 15879455]
[05] T. R. Golub *et al.*, *Science*, 286: 531 (1999) [PMID: 10521349]
[06] P. Baldi & A. D. Long, *Bioinformatics*, 17: 509 (2001) [PMID: 11395427]
[07] J. H. Cho *et al.*, *FEBS Letters*, 551: 3 (2003) [PMID: 12965195]
[08] K. Yang *et al.*, *BMC Bioinformatics*, 7: 228 (2006) [PMID: 16643657]
[09] M. Xiong *et al.*, *Genome Research*, 11: 1878 (2001) [PMID: 11691853]
[10] I. Guyon *et al.*, *Machine Learning*, 46: 389 (2002)
[11] O. Alter *et al.*, *Proc Natl Acad Sci U S A*, 97: 10101 (2000) [PMID: 10963673]
[12] http://research.dfci.harvard.edu/
[13] http://www.broad.mit.edu/
[14] http://www.stjuderesearch.org/
[15] http://www.csie.ntu.edu.tw/~cjlin

Edited by P. Kangueane

Citation: Xu *et al.*, Bioinformation 2(7): 301-303 (2008)

Supplementary material

#Genes reported in OMIM				#Genes reported in KEGG pathways			
MLL	**ALL-AML-3**	**ALL-AML-4**	**ALL**	**MLL**	**ALL-AML-3**	**ALL-AML-4**	**ALL**
85	98	97	93	39	51	58	49
82	96	94	92	32	45	50	47
86	95	91	84	44	44	38	39

Table 1: Number of genes reported in OMIM, KEGG by SDED, GS2 and CHO