A Novel Feature Selection Method using Evolving Supervised Clustering and Applications for Gene Expression Data Modeling

Nikola Kasabov, Liang Goh
Knowledge Engineering and Discovery Research Institute (KEDRI)
Auckland University of Technology
Private Bag 92006, Auckland 1020, New Zealand
nkasabov@aut.ac.nz; lgoh@aut.ac.nz

Many bioinformatics applications, such as the microarray gene expression data modeling, are characterized by a large number of variables (e.g. genes) and small number of samples. Finding the most suitable set of variables for the modeling task (e.g. classification, prediction, profiling) is an important procedure, as it will define to a greater extent the accuracy of the model and its applicability. Several methods have been used so far, that include: correlation analysis; signal to noise ratio analysis; t-test [1-6]. The approach proposed here is based on a supervised clustering, achieved in an evolving connectionist system (ECOS)[7]. The ECOS is developed as a classification model on gene expression data. It learns to classify the data into various classes. The classification model is based on clustering and the formed clusters are used to select the sets of most important variables for each of them. The proposed method applies ECOS on the whole variable space (e.g. 7,129 genes) and using the formed clusters selects the most significant genes for the task.

As a case study, we have used the lymphoma microarray data [5] which consists of 77 samples classified into two classes - Diffuse Large B-cell lymphoma (DLBCL) and Follicular lymphoma (FL).

ECOS are evolving connectionist systems that adapt their structure and functionality according to the data. An ECOS for classification clusters in a supervised way vector samples according to their ‘similarity’ in the Euclidean problem space through adjusting an influence field for each cluster. For the case study here, the whole data set of microarray lymphoma data is fed into the system, distances between the samples are calculated and rule nodes (prototypes, cluster centers) are created to cluster (partition) the input space of 7,129 genes by their expression values. The feature selection procedure consists of applying the signal to noise ratio method, not on the pair of data from the two classes, but on all possible combinations of pairs of clusters from each class. In our experiment there were 22 clusters for the first class and 10 for the second class of lymphoma data formed on the whole 7,129 gene expression data space. After the selection of a set of variables for each cluster and the union of these sets, a new ECOS was then trained on the selected variable set and tested through the leave one out method. The result is 89.61% accuracy. These results outperformed an ECOS trained on a selected set of variables with the use of the signal to noise ratio method applied in the traditional way – accuracy of 84.95%.

The proposed method combines the task of classification and the task of feature selection and uses the obtained clusters in an ECOS that learns a classification task to extract specific features for each of the clusters of class data. The method overcomes the problem of the signal to noise ratio method when data of the same class are spread in several clusters of the problem space and overlap with clusters of another classes.

Reference

[1] A. A. Alizadeh, M. B. Elsen, R. E. Davis, C. Ma, and et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503, 2000.

[2] L. D. Miller, P. M. Long, L. Wong, S. Mukherjee, L. M. McShane, and E. T. Liu, "Optimal gene expression analysis by microarrays," *Cancer Cell*, vol. 2, pp. 353-361, 2002.

[3] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, and et al., "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, pp. 426, 2002.

[4] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, and et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 15149, 2001.

[5] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, pp. 68-74, 2002.

[6] L. J. v. t. Veer, H. Dai, M. J. v. d. Vijver, Y. D. He, and et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530, 2002.

[7] N. Kasabov, "Evolving connectionist systems for adaptive learning and knowledge discovery: methods, tools, applications VO - 1," presented at Intelligent Systems, 2002. Proceedings. 2002 First International IEEE Symposium, 2002.