Symbolic and Subsymbolic Machine Learning Approaches for Molecular Classification of Cancer and Ranking of Genes

W. Dubitzky, M. Granzow, D. Berrar, S. Bulashevska, C. Conrad, D. Gerlich, R. Eils*
Division “Intelligent Bioinformatics Systems” (H0900), German Cancer Research Center, 69120 Heidelberg, Germany
*corresponding author: r.eils@dkfz-heidelberg.de

Background: Classification of human tumors into distinguishable entities is preferentially based on clinical, pathohistological, enzyme-based histochemical, immunohistochemical, and in some cases cytogenetic data. This classification system still provides classes containing tumors that show similarities but differ strongly in important aspects, e.g. clinical course, treatment response, or survival. Thus, information obtained by new techniques like cDNA microarrays that are profiling gene expression in tissues might be beneficial for this dilemma. Microarray experiments, however, provide the scientific community with an immense amount of data. Without appropriate analysis tools significant perceptions hidden in the pool of data might not be recognized. Therefore, methods capable of handling large data sets of thousands of attributes are demanded.
Method: Based on microarray gene expression, we investigate two popular machine learning techniques in the context of molecular classification of cancer, identification of most informative genes and predicition of clinically relevant parameters. The techniques in question are (1) decision trees (symbolic approach) and (2) artificial neural networks (subsymbolic approach). As a basis for our comparative study we have chosen two of the most popular algorithms in machine learning software, namely the decision tree/rule induction algorithms C5.0 and the well-known backpropagation algorithm for multilayer perceptrons (MLP), a specific architecture of artificial neural networks (ANN) [2,3,4]. For both algorithms we used the proprietary implementation realized in the data mining tool Clementine from SPSS [5]. Decision trees are advantageous in situations where the complexity is relatively low (small number of variables and low degree of dependencies between variables) and the variables are directly interpretable by humans (numeric variables such as age, cholesterol, and symbolic variables such as gender, tumor stage etc.). Artificial neural networks on the other hand have been found useful in situations where there are many interacting variables (e.g., images) and non-linear behavior of the underlying phenomena. We used all expression data (except the control data) without further processing

1. to determine, compare, and explain the classification performance of both methods based on n-fold cross-validation procedure and the commonly used lift measure [3]
2. to analyze the entire set of 72 cases and determine the genes that are most relevant for the classification of the underlying tumor classes.

Summary of results:

For ANN classification, each MLP was composed of one input, two hidden and one output layer. The average classification accuracy over all 6 cross-validation runs was 84.35%. Further analysis showed that although for each of the three neural net runs the ALL tumor was classified with a higher accuracy (92.76%) than the AML class (54.74%) the lift measure for the AML class scored higher in each of the test runs. Hence, the model showed a higher sensitivity/selectivity with regard to the AML class. The best classification performance of the C5.0 decision tree method was obtained on the basis of 20-fold boosting (combination of multiple definitely different models). In this case the average classification accuracy over all 6 cross-validation runs was 92.98%. as compared to 84.09% without boosting. For most of the cross-validation subsamples, boosting was able to identify multiple complementary models, thus indicating multiple genes and expression levels related to differentiating AML and ALL.

Conclusions: The comparison of the two methods suggests that (1) both can be used directly (no further preprocessing or discretization) with high dimensional inputs (> 7000 genes) for molecular tumor classification and gene identification, (2) the C5.0 decision tree seems to be the preferred classification model as it (a) showed higher precision and sensitivity levels, (b) provides an output format that is easy to interpret by humans (symbolic rules), and (c) was faster to train than the neural model. However, sensitivity analysis for ranking and identifying high-impact variables was found easier to use, as it provided a direct ranking of the genes. A list of the fifty most relevant genes based on all 72 cases will be published in the final paper.

References:

[1] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Science 286(5439): 531-537, 1999.

[2] Werbos, P. J.: Beyond Regression, Doctoral Dissertation, Appl. Math., Harvard University, November 1974.

[3] Rumelhart, D. E. et al.: Parallel Distributed Processing, Vol. 1, MIT Press Cambridge, 1986.

[4] J.E. Dayhoff, “Neural Network Architectures: An Introduction”, Thomson Computer Press, 1996.

[5] SPSS: http://www.spss.com/datamine/, and Clementine User Group: http://www.spss.com/clement