Cancer classification based on gene expression using neural networks

H.P. Hu, Z.J. Niu, Y.P. Bai and X.H. Tan

School of Science, North University of China, Taiyuan, Shanxi, China

Corresponding author: H.P. Hu
E-mail: hhp92@163.com

Genet. Mol. Res. 14 (4): 17605-17611 (2015)
Received June 29, 2015
Accepted September 25, 2015
Published December 21, 2015
DOI http://dx.doi.org/10.4238/2015.December.21.33

ABSTRACT. Based on gene expression, we have classified 53 colon cancer patients with UICC II into two groups: relapse and no relapse. Samples were taken from each patient, and gene information was extracted. Of the 53 samples examined, 500 genes were considered proper through analyses by S-Kohonen, BP, and SVM neural networks. Classification accuracy obtained by S-Kohonen neural network reaches 91%, which was more accurate than classification by BP and SVM neural networks. The results show that S-Kohonen neural network is more plausible for classification and has a certain feasibility and validity as compared with BP and SVM neural networks.

Key words: Colon cancer; Gene expression; S-Kohonen neural network; BP neural network; SVM neural network

INTRODUCTION

With the rapid development of science and technology, great changes have taken place in the human diet structure. In addition, various external factors also pose threats to the human health. Cancer is one of the major diseases threatening human life and health. The correct classification of tumor type is important for patient diagnosis and treatments.

Two major challenges are encountered in cancer classification: class discovery and class prediction by gene expression monitoring. In cases such as acute leukemia, cancer susceptibility is associated with changed in gene expression. The method of weighted voting genes is used to classify leukemia into two subtypes AML and ALL (Golub et al., 1999).

Artificial neural networks are also applied for survival rate prediction. In breast cancer, area under the ROC curve (AUC) can be used as a measurement of accuracy when evaluating survival rate estimates for 5, 10, and 15 years in patients (Lundin et al., 1999).

Ovarian cancer, colon cancer, and leukemia are researched by extracting feature genes using the signal-to-noise ratio, analyzing the gene expression data with SVM, and setting up predicting models of tumors (Carsten et al., 2003).

A multiple-random validation strategy was used for prediction of cancer outcome with microarrays by use of statistical analysis (Michiels et al., 2005).

Interestingly, it was found that expression of certain genes unrelated to cancer such as of postprandial laughter and of skin fibroblast localization was significantly associated with breast cancer outcomes (Venet et al., 2011). Neural network and logistic regression were used to predict distant metastasis (DM) of colorectal cancer (CRC) patients (Biglarian et al., 2012).

In early breast cancer prediction, digital mammography ensures patients' early detection through localization of suspicious areas with benign/malignant microcalcifications using Gabor filter with discrete cosine transform and benign/malignant classification using neural network (Valarmathi et al., 2013). Gene expression is related to the process of glycosylation between breast cancer subtypes, and is applied to predict the survival of breast cancer patients (Potapenko, 2015).

Screening the differential gene expression patterns between bladder carcinoma patients and normal subjects by use of empirical Bayes methods of the linear models, co-gene-expression is obtained (Bi et al., 2015). This allows discovery of complex gene networks that play important roles in bladder cancer research.

Colon cancer places as the 4-6th most common malignant tumors, and its incidence is still rapidly rising. Aside from acute leukemia, breast cancer, and bladder cancer, colon cancer is also connected to changes in gene expression.

We found that based on gene expression, colon cancer patients with UICC II who were treated by elective standard oncological resection is classified into two cases: relapse and no relapse after surgery by use of S-Kohonen, BP and SVM neural networks. By comparison, results show that S-Kohonen neural network is more effective for cancer classification.

MATERIAL AND METHODS

Data sources

Research and analysis of gene expression is an important research topic in bioinformatics. Therefore, classification models for probability of UICC II cancer recurrence and identification

based on gene expression can be important for treatment and disease prognosis. In this paper, gene expression data with colon cancer came from database of NCBI NLM NIH. The database is composed of gene expressions of 53 colon cancer patients: 13 relapse cancer patients and 40 no relapse cancer patients. Every colon patient has 54,675 gene features. By experiment, 500 gene features are randomly chosen. Thus, 53 500-dimension samples with color cancer data were obtained. Samples (53) were normalized and randomly distributed. A no-relapse to relapse ratio of 4:1 was found, as shown in Table 1.

Table 1. Distribution of samples of colon cancer with UICC II stage.

Sets	No relapse count	Relapse count	Total
Training	33	10	43
Test	7	3	10

Kohonen neural network

Topological structure of Kohonen neural network is shown in Figure 1.

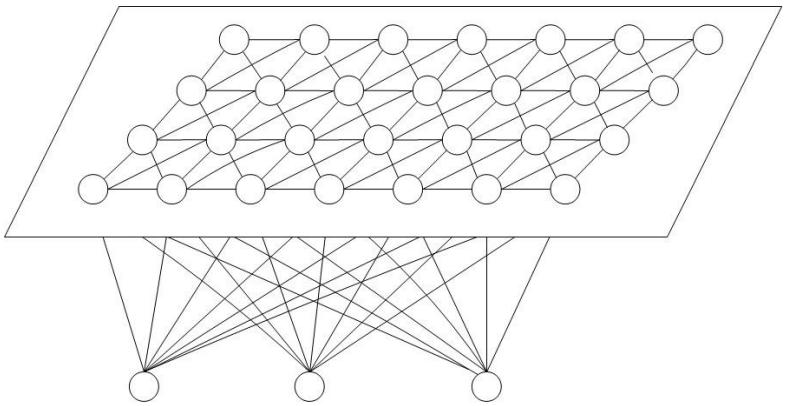

Figure 1. Topological structure of Kohonen neural network.

Kohonen neural network is a self-organizing neural network proposed by Dr. Teuvo Kohonen of University of Helsinki Finland, whose learning process is unsupervised and whose weights are adjusted by self-organizing map to recognize the features in order to obtain automatic clustering. Kohonen neural network is a two-layer feedforward neural network: the first layer is the input layer, whose nodes are equal to the dimension of the input sample, and the second layer is the competitive layer (that is, output layer) whose nodes' distribution is a 2-dimension array. There are whole connections of adjustable weights between input nodes and output nodes. Thus, Kohonen neural network is a map from n-dimension output space to 2-dimension output plate, and keep the same topological feature.

Based on Kohonen neural network, the added layer (that is, output layer) following the competitive layer makes the neural network achieve the goals of prediction, classification and so on. The obtained network is named the S-Kohonen neural network. The adjustment of learning

rate in algorithm immediately affects clustering precision and convergence. Output of S-Kohonen neural network is 1 or 2. In this study, the learning rate in S-Kohonen neural network was chosen to make clustering better and convergence faster, whose formula is the exponential function:

$$\alpha(t) = \alpha_0 e^{(\frac{t}{T})} \quad \text{(Equation 1)}$$

$$\eta(t) = \eta_0 e^{(\frac{t}{T})} \quad \text{(Equation 2)}$$

where $a(t)$ is the learning rate of connected weights between input layer and competitive layer, a_0 is initial, $\eta(t)$ is the learning rate between competitive layer, and the output layer, η_0 is initial, t is the t-th iteration, T is the total number of iterations. $a(t)$ decreases with the increasing iterations.

BP neural network

BP (Back Propagation) neural network proposed by Rumelhart and McCelland in 1986 is a multilayer feedforward network, which used to be the widest neural network. BP neural network consists of three layers: input layer, hidden layer and output layer. Each layer influences next layer. The learning process of BP algorithm is supervised. If the error of the outcome in the output layer is larger than the goal-error, the BP algorithm is turned into the back propagation. Topological structure of BP neural network is shown in Figure 2.

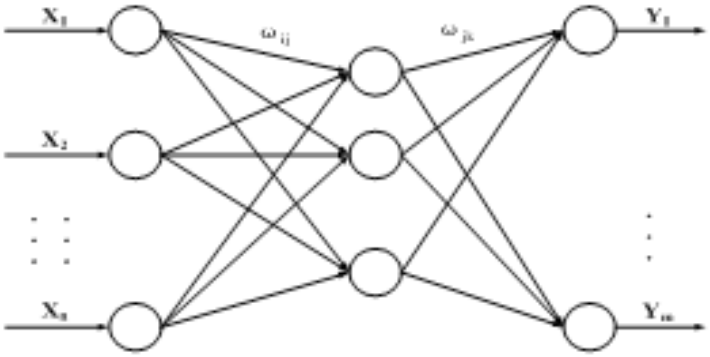

Figure 2. Topological structure of BP neural network.

In Figure 2, $X_1, X_2, \cdots, X_n$ are the input values of the network and $Y_1, Y_2, \cdots, Y_m$ are the predictive values of the network. There is a relation of function map between n independent variables and m dependent variables.

Support vector machines (SVM)

Support vector machine (SVM), firstly proposed by Vapnik, is a new-style machine learning method for pattern classification and nonlinear regression. In this method, classification hyperplate is set up as a decision surface, which maximizes edge isolation between the positive samples

and the negative samples. The classification method of SVM is divided into two classes: the direct approach and the indirect approach.

In this paper, the one-to-many method of the second class was adopted to set up models samples so that k SVM is built by samples of k classes.

SIMULATION EXPERIMENT

Data processing

In this study, experimental data with the same proportion are used in the above three methods. The classification prediction models are built by three neural networks such that the results are comparable. A simulation experiment was carried out to verify the validation of these three neural networks. Gene expression data of UICC II colon patients were normalized, and classified as either no relapse or relapse. Gene features (100-1000) from 53 samples were used in the three neural networks mentioned above. The results are shown in Figure 3.

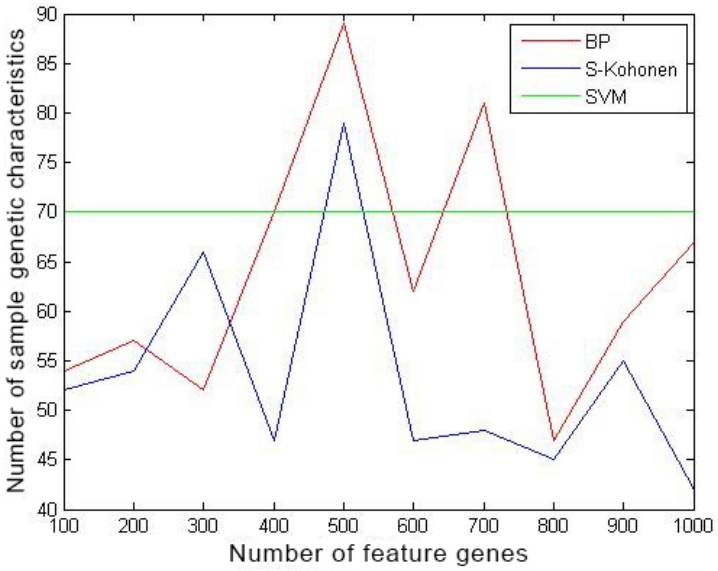

Figure 3. Classification accuracy of the three neural networks with the number of feature genes.

EXPERIMENT RESULTS

Samples (53) with 500 chosen gene features chosen were used for classification prediction by S-Kohonen neural network, BP neural network, and SVM neural network. Some experiments were performed repeatedly in term of randomness of initial weights. Based on the characteristic of the data, the nodes of input layer were 500 and the nodes of output layer were 2. The first, third, and fifth experiment results by S-Kohonen neural network were taken to obtain the classification charts, as shown in Figure 4. By analyzing the classification charts, the structure of S-Kohonen neural network was increasingly stable. The winner nodes of different category were distributed by blocks.

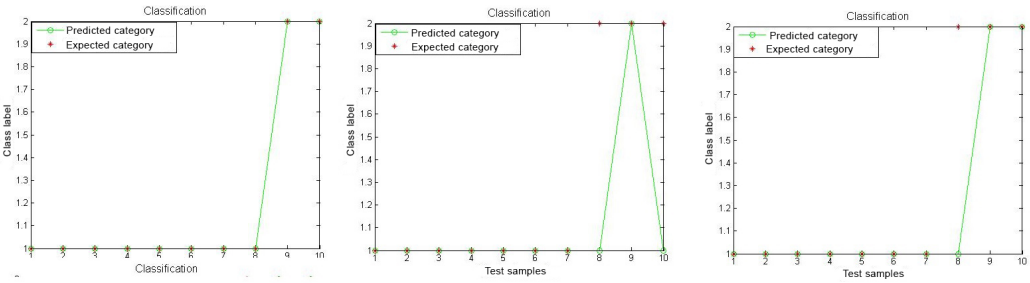

Figure 4. First, third, and fifth classification charts by use of S-Kohonen neural network.

Nodes (25) in hidden layers were chosen in the BP neural network. The first, third, and fifth experimental results by BP neural network were taken to obtain the classification charts, as shown in Figure 5. The results show that the classification results were arbitrary. Some experimental results by SVM neural network were taken to obtain the same classification accuracy, as shown in Figure 6.

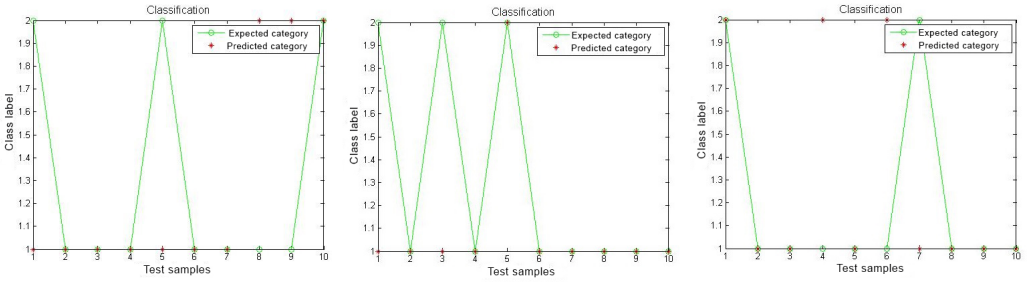

Figure 5. First, third, fifth classification charts used by the BP neural network.

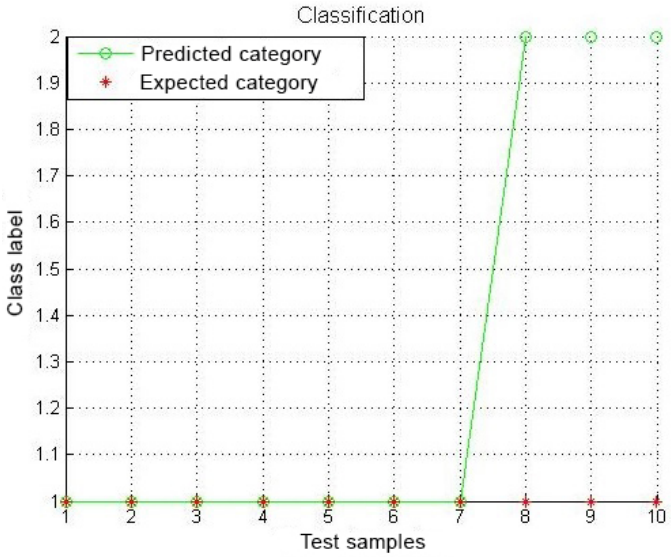

Figure 6. Classification chart of SVM neural network.

Results from ten repeated experiments were taken, and the right count of classification for each is listed in Table 2. Predicted classification by S-Kohonen neural network were in accordance with the expected classification, which reached an accuracy of 91%. Comparing BP neural network and SVM neural network with S-Kohonen neural network, the classification accuracies of BP neural network and SVM neural network were 66 and 70%, respectively. These results are shown in Table 2.

Table 2. Classification results of S-Kohonen, BP, and SVM neural network.

	1st time	2nd time	3rd time	4th time	5th time	6th time	7th time	8th time	9th time	10th time	Classification accuracy (%)
S-Kohonen	9	8	10	9	8	10	9	9	10	9	91
BP	6	6	8	7	7	8	4	8	4	8	66
SVM	7	7	7	7	7	7	7	7	7	7	70

DISCUSSION

In the paper, S-Kohonen BP, and SVM neural networks were used to classify stage UICC II colon patients into either the no relapse group or the relapse group. Trained set and test set chosen were equivalent each time. However, the classification results were different. The results showed that Kohonen neural network is more suitable for classification of colon cancer based on gene expression as compared to BP and SVM neural network, and maintains better classification accuracy and stability. Thus, it can meet the practical requirements of classification and prediction.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation (#61275120) of China.

REFERENCES

Bi DB, Ning H, Liu S, Que X, et al. (2015). Gene expression patterns combined with network analysis identify hub genes associated with bladder cancer. *Comput. Biol. Chem*. 56: 71-83.

Biglarian A, Bakhshi E, Gohari MR and Khodabakhshi R (2012). Artificial Neural Network for Prediction of Distant Metastasis in Colorectal Cancer. *Asian Pac. J. Cancer Prev.* 13: 927-930.

Golub TR, Slonim DK, Tamayo P, Huard C, et al. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* 286: 531-537.

Lundin M, Lundin J, Burke HB, Toikkanen S, et al. (1999). Artificial neural networks applied to survival prediction in breast cancer. *Oncology* 57: 281-286.

Michies S, Koscielny S and Hill C(2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 365: 488-492.

Peterson C and Ringnér M. (2003). Analyzing Tumor Gene expression profiles. *Artif. Intell. Med.* 28: 59-74.

Potapenko IO, Lüders T, Russnes HG, Helland Å, et al. (2015). Glycan-related gene expression signatures in breast cancer subtypes; relation to survival. *Mol. Oncol.* 9: 861-876.

Valarmathi P and Radhakrishna V (2013). Tumor Prediction in Mammogram using Neural Network. *Global J. Inc.* 3: 19-24.

Venet D, Dumont JE and Detours V (2011). Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. *PLoS Comput. Biol.* 7: e1002240.