Neural Processing Letters **19:** 211–226, 2004.

Application of Probabilistic Neural Networks to the Class Prediction of Leukemia and Embryonal Tumor of Central Nervous System

CHENN-JUNG HUANG[1] and WEI-CHEN LIAO[2]
[1]*Institute of Learning Technology, National Hualien Teachers College, 123 Huo-His Rd., Hualien, Taiwan 970, Taiwan. e-mail: cjhuang@mail.nhltc.edu.tw*
[2]*Department of Natural Science Education, National Taitung University, Taitung, Taiwan 950, Taiwan*

Abstract. Accurate diagnosis and classification is the key issue for the optimal treatment of cancer patients. Several studies demonstrate that cancer classification can be estimated with high accuracy, sensitivity and specificity from microarray-based gene expression profiling using artificial neural networks. In this paper, a comprehensive study was undertaken to investigate the capability of the probabilistic neural networks (PNN) associated with a feature selection method, a so-called signal-to-noise statistic, in the application of cancer classification. The signal-to-noise statistic, which represents the correlation with the class distinction, is used to select the marker genes and trim the dimension of data samples for the PNN. The experimental results show that the association of the probabilistic neural network with the signal-to-noise statistic can achieve superior classification results for two types of acute leukemias and five categories of embryonal tumors of central nervous system with satisfactory computation speed. Furthermore, the signal-to-noise statistic analysis provides candidate genes for future study in understanding the disease process and the identification of potential targets for therapeutic intervention.

Key words. Acute leukemia, class prediction, embryonal tumor of central nervous system, feature selection, gene expression data, probabilistic neural network

1. Introduction

Successful cancer treatment depends on choosing the right regimen for a given patient. How to diagnose cancer subtypes accurately becomes one of the biggest challenges in clinical cancer research, since separate treatment strategies are adopted for different tumors. The diagnosis of cancer has traditionally been made on the basis of non-molecular criteria such as tumor tissue type, pathological features, and clinical stage. A recent study reported by Golub *et al.* [1, 2], the first microarray-based and bioinformatic-orientated approach for identifying and classifying tumor types, moves cancer diagnosis away from visually based systems to molecular based systems. They employed a signal-to-noise statistic to select a small set of genes and developed a scheme named weighted voting to distinguish acute lymphoblastic leukemia (ALL) from acute myeloid leukemia (AML); the recognition rate they obtained was 94.1%. The report of Golub *et al.* [1, 2]

made a revolution for the cancer classification and diagnosis. They created a novel criteria for molecular classification of cancer by using microarray gene expression analysis.

Several algorithms have been used to analyze publicly accessible data sets on cancer research in the literature [3–10] motivated by the report of Golub *et al* [1]. By using the ALL/AML data set, Toure *et al*. [3] applied the multilayer perceptron network (MLP) [11] to predict the class of leukemia cancer and gave 58% accuracy on test data. Ryu *et al*. [4] experimented with the MLP, support vector machine (SVM) [12], and the k-nearest neighbor (KNN) [13] as the classifiers, and the best classification rate they achieved was 97.1% if gene is selected via the combination of Pearson's correlation analysis and the MLP. Su *et al*. [5] employed the modular neural networks and the best 75% correct classification was reached. Xu *et al*. [6] adopted the ellipsoid ARTMAP [14] to analyze the data and the best result was 97.1%. For other data sets, Ben-Dor *et al*. [7] experimented with several approaches to analyze the Colon and Ovarian cancer data sets, and achieved 88.7% recognition rate with a clustering algorithm [15] and 89.3% with the boosting method [16] correspondingly. Kuramochi *et al*. [8] investigated 249 gene function classes in the MIPS database with the SVM and the KNN respectively, and obtain the highest precision rate of 48.6%. Azuaje [9, 10] used the growing cell structures (GCS) and the simplified fuzzy ARTMAP neural learning model [17, 18] to distinguish normal subjects from those with diffuse large B-cell lymphoma (DLBCL) respectively, and best recognition rate of 76% was achieved. Pomeroy *et al*. [19] used the KNN to classify five types of embryonal tumors of central nervous system (CNS) and achieved 83.3% accuracy rate.

Although most of the above-mentioned algorithms demonstrated that neural networks possess the adequate ability to analyze and classify the data with binary classes, it is essential to find a model that can improve both the prediction accuracy and the computationally expensive training and recall phases of neural networks in the process of massive data sets. What is more, no effort has been made on investigation of applying neural network approaches to predict multiple tumor types such as embryonal CNS tumors [19] so far. To address these challenges, we perform a study on the suitability of the probabilistic neural networks (PNN) [20–25] as a classifier with a set of informative patterns selected from four data sets via an appropriate feature selection method prior to the training and testing stages. The comparison of several representative feature selection methods was given by Ryu *et al*. [4], such as information gain, mutual information, Pearson's and Spearman's correlation coefficients, and the signal-to-noise, statistic, etc. They reported that the SN statistic could select the most informative genes for the classification of ALL/AML in average. Thus, we compare the SN statistics with three other well-known feature selection methods, which are the Chi-Square statistic [26], the Relief method [27], and the Correlation-based Feature Selection method [28], in this study, and then use four data sets to evaluate the PNN and the feature subset composition. For the ALL/AML data

set, we first train the PNN with a set of 38 leukemia samples and evaluate with another set of 34 test samples. Next, 100 iterations of stratified 10-fold cross validation [29, 30] were used to evaluate the PNN with the 72 ALL/AML samples, and three embryonal CNS tumors data sets, which consist of 42 samples of the embryonal CNS multi-class tumors, 34 medulloblastoma morphology samples, and 60 samples of medulloblastoma treatment outcome, respectively. The experimental results show that the PNN can achieve higher classification rate in the acute leukemias and the embryonal CNS multi-class tumor data sets in the literature. Meanwhile, this study reveals that the performance of the PNN classifier is greatly affected by the correlation with the class distinction in the data set. The number of the marker genes can be reduced to a certain extent if the gene's typical expression in one class is quite different from its typical expression in the other.

The remainder of the paper is organized as follows. The feature selection method for choosing effective predictive genes in our work is introduced in Section 2. Then Section 3 gives a brief introduction for the architecture of the PNN network. Section 4 examines the simulation results of the classifier operated on four data sets. Section 5 interprets the marker genes selection results and discusses its biological relevance. Conclusions are made in Section 6.

2. Feature Selection

We used the four datasets collected in [1, 19] for training and testing of our classifiers. The four datasets consist of 72, 42, 34, and 60 samples, respectively, wherein each sample contains 7129 gene expression numbers. The preprocessing of marker genes selection from the dataset is thereby important because each data set is highly dimensional and many genes in the data set are irrelevant to distinction of the cancer class [1].

The feature selection methods can be divided into two broad categories in the literature [1,4,26–28,31], i.e. filter methods and wrapper methods. Filter methods select predictive subset of the features using heuristics based on characteristics of the data, whereas wrapper methods make use of the classifier actually used to evaluate the accuracy of feature subsets. Wrapper methods generally result in better performance than filter methods because the latter suffers from the potential drawback that the feature selection principle and the classification step do not necessarily optimize the same objective function. However, for the case with gene expression data where the number of features is large in this study, wrapper methods are far too expensive to be used because each feature set considered must be evaluated with the trained PNN classifier. Moreover, the repeated application of cross validation on the same data set might result in finding a feature subset that performs well on the validation data alone. For these reasons, wrapper methods will not be considered in this study since filter methods are much faster than wrapper methods and therefore are better suited to high dimensional data sets.

2.1. SIGNAL-TO-NOISE STATISTIC

Several representative feature selection algorithms have been investigated in this work. Slonim *et al.* [2], Pomeroy *et al.* [19], and Ryu *et al.* [4] tested with several gene selection methods and reported that the best performance was obtained with the signal-to-noise statistic defined by

$$F = \left|\frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}\right|, \tag{1}$$

where μ_1 denotes the mean expression level and σ_1 represents the standard deviation of expression for the samples in class 1, respectively. μ_2 and σ_2 are defined similarly for the samples in class 2. It is obvious that Equation (1) tries to pick up the genes with the feature of wider class separation and the smaller spread around class means.

2.2. CORRELATION-BASED FEATURE SELECTION

It was reported in [32, 33] that Correlation-based Feature Selection (CFS) [28] demonstrated its feasibility to determine the discriminatory features. Based on the hypothesis that a good feature subset is one that contains features highly correlated with the class, yet uncorrelated with each other, the CFS uses a correlation-based heuristic to evaluate subsets of features:

$$Merit_s = \frac{k \cdot \bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}}, \tag{2}$$

where k denotes the number of features for a feature subset S, r_{cf} is the average feature–class correlation and r_{ff} stands for the average feature–feature inter-correlation. In order to calculate the $Merit_S$ score, numeric features are discretized and then the degree of association between discrete features are calculated using symmetrical uncertainty

$$\mathrm{SU} = 2.0 \cdot \left(\frac{H(Y) - H(Y|X)}{H(X) + H(Y)}\right), \tag{3}$$

where X and Y denote discrete features, and $H(Y)$ H and $H(Y|X)$ give the entropy of Y before and after observing X

$$H(Y) = -\sum_{y \in Y} p(y)\log_2 p(y), \tag{4}$$

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x)\log_2 p(y|x), \tag{5}$$

where $p(x)$is the probability for the variable x.

The CFS algorithm uses a heuristic search strategy to explore the feature subset space. The search starts with an empty set of features and generates all possible single feature subsets. The subset with the highest $Merit_S$ score is chosen, and is expanded by adding single features as before. This process is repeated with a stop-

ping criterion of five consecutive subsets expanded with no improvement to the $Merit_S$ score.

2.3. CHI-SQUARE STATISTIC

Another commonly used feature selection method is chi-square statistic (χ^2) method [34–36]. This method evaluates each gene individually by measuring the chi-square statistics with respect to the classes. The gene expression numbers are first discretized into several intervals using an entropy-based discretization method. Then the chi-square value of each gene is computed by

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{\left(A_{ij} - \frac{R_i \cdot C_j}{N}\right)^2}{\frac{R_i \cdot C_j}{N}}, \tag{6}$$

where m denotes the number of intervals, k the counts of classes, N the total number of patterns, R_i the number of patterns in the ith interval, Cj the number of patterns in the jth class, and A_{ij} the number of patterns in the ith interval, jth class. The genes with larger chi-square statistic values are then selected as marker genes for classification.

2.4. RELIEF

Relief and its variants were successfully applied to input feature selection in power system static security assessment [37] and attribute selection for the decision tree machine learning method [38, 39]. The key idea of Relief is to reinforce the similarities among instances in the same class, while decline the similarities among instances in different classes. Given an instance, Relief first searches for its two nearest neighbors, one from the same class, and the other from a different class. Then this method updates the weighting of each feature via the computation of difference of the feature values for two nearest neighbors. The feature weighting approach is used to estimate attributes based on how well their values distinguish among the instances that are near to each other. The extension of Relief called Relief further relaxes the limitations of original Relief algorithm and is able to manage multiclass problems and tolerate incomplete and noisy data.

3. Probabilistic Neural Networks

The motivation of using the PNN network in the classification of cancer is inspired by the characteristic of the feature selection strategy mentioned in the previous section. The gene selection strategy given in Equation (1) attempts to pick up the gene expressions with the feature of better class distinction and relatively correlated with the class they belong to. It is well known that the key advantage of the PNN is that only one epoch of training is necessary, whereas its major drawback is that an enormous amount of memory might be needed to store all training samples and thus slow down the computation of recall process [20–25]. Accordingly, the proper way to

speed up the computation of both the training and recall processes for the PNN is to keep both the number and the dimension of the training samples as small as possible. Meanwhile, by using the Parzen window estimator to estimate the probability distributions of the class samples, the PNN models the Bayesian classifier [24] and tries to minimize the expected risk of classifying patterns in the wrong class. The PNN can thus perform well if we obtain a small set of test data with large value of class separation metric as defined in Equation (1). In addition, a single PNN is capable of handling multi-class problem, such as the embryonal CNS multi-type tumor data set. This is opposite to the so-called one-against-the-rest or one-per-class approach taken by some classifiers [40], such as the SVM, which decompose a multi-class classification problem into dichotomies, and each chotomizer has to separate a single class from all others.

Figure 1 illustrates the architecture of a PNN network with two hidden layers. The number of the nodes in the pattern units as shown in Figure 1 is identical to the counts of the training samples, and the synaptic weight $w_{ij}^{(p)}$ in the input-to-pattern connections is

$$w_{ij}^{(p)} = x_i^{(j)}, \tag{7}$$

where $x_i^{(j)}$ denotes the ith node input of the jth sample at the input layer.

As for the weight between the pattern and summation units $w_{jk}^{(s)}$ can be expressed as

$$w_{jk}^{(s)} = \begin{cases} 1 & \text{if } T_k^{(j)} = 1, \\ 0 & \text{else,} \end{cases} \tag{8}$$

where the value of $T_k^{(j)}$ is 1 only when sample j is associated with class k, and 0s elsewhere.

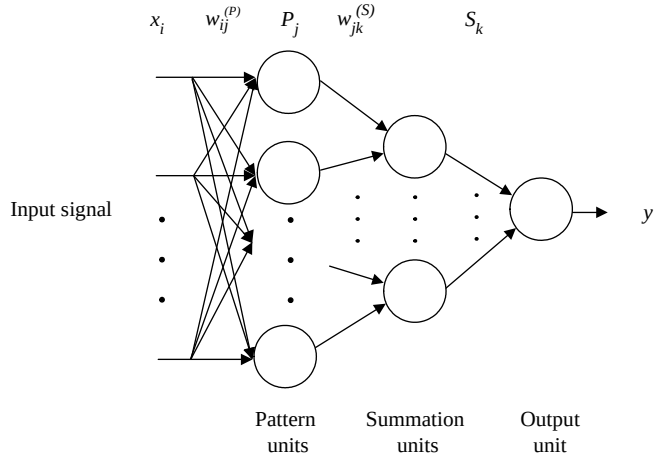

Figure 1. Architecture of a probabilistic neural network.

After the instantaneous training process as shown in Equations (7) and (8), the classification of input patterns can be initiated by computing the net input to the pattern units as follows:

$$n_j^{(p)} = \sqrt{\sum_i (w_{ij}^{(p)} - x_i)^2}. \tag{9}$$

Then the output of the pattern units is computed as

$$P_j = \exp\left(-\frac{n_j^{(p)}}{2\sigma^2}\right), \tag{10}$$

where σ is a smoothing parameter corresponding to the standard deviation of the Gaussian distribution. Note that if the input is close to one or several training vectors of a single class, it is represented by one or several outputs at the pattern units that are close to 1.

At the summation units, each node represents an individual class. The output of each node can be expressed as

$$S_k = \frac{1}{\sum_j w_{jk}^{(S)}} \sum_j w_{jk}^{(S)} \cdot P_j. \tag{11}$$

Then the output layer classifies the input vector into a specific one class if that class had the maximum output value from the corresponding node at the summation units:

$$y = \arg\max_k S_k. \tag{12}$$

The time complexity of the recall process for the PNN is $O(N \cdot \tau)$, where N denotes the dimension of the input patterns, and τ represents the counts of the training samples. Thus the PNN classifier turns out to be infeasible if there is no collocated feature selection method to trim the size of the feature vectors for a massive data set.

4. Experimental Results

As mentioned in Section 2, we experimented with four data sets in this work. They are the leukemia, embryonal CNS tumor, medulloblastoma morphology, and medulloblastoma treatment outcome data sets, respectively. All the four data sets are composed of 7219 gene expressions. As many of the gene expression levels in the four data sets are most likely irrelevant to the classification of cancer [1, 2, 19], and an input vector of high dimension will seriously slow down the operating speed of the PNN, so we preprocess the raw expression data by applying the feature selection methods described in Section 2 on each data set separately. We generate different subsets of 10, 50, 100, 200, and 500 marker genes respectively for each data set by using the signal-to-noise statistic, chi-square statistic, and Relief methods in turn, while select all the features from each data recommended by the CFS method as mentioned in Section 2. Then 100 iterations of stratified 10-fold cross-validation [29, 30] were used to evaluate

the composite of the PNN model and the marker gene subsets for the samples of each data set in turn. Notably, the tissue samples used in the test set for cross-validation are not used in the selection of the marker gene subsets to avoid the notorious selection bias problem [30], and stratified 10-fold cross-validation is adopted here to effectively decrease the variance due to the small number of training samples. The dataset is randomly divided into 10 sets with approximately equal size and class distributions in stratified 10-fold cross-validation. For each "fold", the classifier is trained using all but one of the 10 groups and then tested on the unseen group. This procedure is repeated for each of the 10 groups. Table 1 shows the average 10-fold cross-validation classification result of the four data sets by using the PNN classifier and the four feature selection methods to select the marker genes. The signal-to-noise statistic outperforms the other three feature selection methods for all the four data sets.

As mentioned in Section 1, different classifiers were used to analyze the ALL/AML data set by the researchers recently. We thus set up the same testing environment to compare the performance of the PNN and the signal-to-noise statistic composite with others reported in the literature. Within the 72 leukemia samples, 38 samples are used for training the PNN model by the leave one out cross-validation (LOOCV) method [30], again feature selection is not performed on the complete data set to avoid bias selection. The smoothing parameter σ given in Equation (10) is tuned to generate the fewest possible errors during the cross-validation stage. Then, the trained classifier is applied to the other 34 independent leukemia samples to produce the prediction

Table 1. 10-Fold cross-validation classification result of the four feature selection methods.

	Data set			
Feature selection method	Leukemia (%)	Embryonal CNS tumor (%)	Medulloblastoma morphology (%)	Medulloblastoma treatment outcome (%)
Signal-to-noise statistic	95.4	84	94.1	69.2
χ^2 statistic	94	79.9	92.3	55.8
Relief	93	70.7	92.6	60.7
CFS	94.6	73.1	88.6	60.3

Table 2. Comparison of best prediction rate for the independent leukemia data set.

Type of classifier	Accuracy rate (%)
Weighted voting [1, 2]	94.1
MLP [4]	97.1
SVM [4]	97.1
KNN [4]	97.1
Mult-domain gating network [5]	75
Ellipsoid ARTMAP [6]	97.1
PNN	100

result. Table 2 shows the prediction results of the 34 leukemia samples using the trained PNN and other classifiers in the literature. The PNN classifier achieves 100% accuracy when the 50-gene predictors are derived in cross-validation tests by means of the signal-to-noise statistic feature selection method as done in [1, 2].

In Table 3, the 10-fold cross-validation evaluation result of the PNN classifier for the four data sets with different sizes of the marker genes is averaged over 100 iterations. The test result shows that the PNN classifier attains similar recognition rates on the four data sets with different subset sizes, except when the size of the marker gene subset is 10 for the Embryonal CNS tumor data set. We believe that the classification of the Embryonal CNS tumor data set requires more genes is because that the data set consists of five classes, and the PNN can not perform well if only two genes are used to represent each class. Note that the problem of overfitting does not likely happen in our PNN model since only two parameters needed to be adjusted. One is the smoothing parameter σ as given in Equation (1), and the other is the size of the marker gene subset. The tuning of the smoothing parameter is stopped whenever the trained PNN classifier generates the fewest errors in cross validation tests such that the PNN does not adapt too well to the distinctive characteristics of the training set. As for the feature subset size, it is demonstrated in Table 3 that the subset size does not affect performance of the PNN classifier significantly, provided that at least tens of marker genes are selected via the signal-to-noise statistic.

The reports given in [4, 19] show that the KNN classifier outperforms other models for the classification of the four data sets when the marker genes are selected via the signal-to-noise statistic. We thus give 10-fold cross-validation tests for the four data sets using the PNN and the KNN classifiers and the best results obtained

Table 3. 10-Fold cross-validation results of the PNN classifier with different subset size.

	Data set			
Number of genes	Leukemia (%)	Embryonal CNS tumor (%)	Medulloblastoma morphology (%)	Medulloblastoma treatment outcome (%)
10	95.2	63.7	86.8	63.8
50	95.3	80.5	89.8	68.0
100	95.4	84.0	86.7	69.2
200	93.0	86.1	84.5	68.9
500	92.8	80.7	82.9	66.2

Table 4. 10-Fold cross-validation classification result for the PNN and the KNN classifiers.

	Data set			
Feature selection method	Leukemia (%)	Embryonal CNS tumor (%)	Medulloblastoma morphology (%)	Medulloblastoma treatment outcome (%)
PNN	95.4	86.1	89.8	69.2
KNN	94	79.6	85.2	66.4

for each classifier are given in Table 4. We can see that the PNN achieve much better performance than the KNN for the embryonal CNS multi-class tumors data set, and slightly outperforms the KNN for the three other data sets.

Since we rank and select genes based on the signal-to-noise statistic in our experiments, we further investigate how the value of the signal-to-noise statistic for each of the four data sets affects the classification result of the PNN. Table 5 shows the average signal-to-noise statistic for the top 10 genes from each data set. It can be seen that the value of signal-to-noise statistic for the top 10 genes from the four data sets is highly correlated with the classification results as given in Table 3. The feature of wider separation between different classes increases the prediction rate of the PNN, while a quite small value of signal-to-noise statistic clearly degrades the performance of the medulloblastoma treatment outcome as shown in Table 4. Notably, the signal-to-noise statistic for the embryonal CNS multi-type tumor data set is obtained by first dividing one specific tumor type and all other tumor types into two groups, and then computing the mean and the standard deviation of the two groups, respectively. It looks like that the signal-to-noise statistic for the multi-class data sets might be more significant than the binary class ones although the prediction rate for the multi-class data sets is slightly lower.

In summary, we can infer from Table 3 that the PNN will not perform well if only 10 marker genes are selected as test data as suggested in [3] for the reason that the PNN needs larger input dimension to provide the classifier more useful information to behave properly. Secondly, the signal-to-noise statistic given in Equation (1) indeed effectively collects the subset of the most informative genes demanded by the PNN classifier whereas leaves the unrelated data behind. The selection of the data set not only raises the classification accuracy rate successfully, but also shortens the computation time of the PNN to a great extent, which is the most burdensome complaint for most neural networks. Moreover, the achievement of the higher prediction rate in Table 4 encourages the usage of the PNN as the classifier for the data sets with multi-categories.

5. Biological Relevance

Gene expression profiling by microarray technology has been successfully applied to classification and diagnostic prediction of cancers. However, there are several technical and biological issues need to be addressed [42–45].

Table 5. Average signal-to-noise statistic of top 10 genes for the four data sets.

Data set	Average of $\lvert\frac{\mu_1-\mu_2}{\sigma_1+\sigma_2}\rvert$
Leukemia	1.2713
Embryonal CNS multi-type tumors	1.8045
Medulloblastoma morphology	0.9070
Medulloblastoma treatment outcome	0.4684

The first issue was whether the microarray data exhibited a high degree of noise. One might expect that because the high experimental uncertainties of expression measurements result in a loss of confidence. However, the noise data generated by microarray array system can be improved by other additional experimental techniques and statistical methods [42]. As the technology advances, microarray experiments are becoming less expensive, which makes the use of multiple arrays feasible. Therefore, well-designed microarray studies will have high power while controlling type I error risk. On the other hands, several computing methods were reported to reduce the noise of microarray data. Comparing with other models, PNN is proved to be one of the artificial neural networks which is less sensitive to noise. In our study, combining gene selection and PNN is a proper approach to extract enough informative microarray data for AML/ALL biological model construction.

The next issue was that microarray data often exceeded the sample space dimensionality by a factor of 100 or more. To address this problem, feature selection and dimension reduction techniques are both used in microarray data analysis. For example, if two genes are similarly co-regulated, then they provide the same basic information, and the removal of one of these genes does, in general, not result in a loss of information for a classifier [43].

Finally, the third issue was that microarray data analysis should fulfill the biologist's requirement for sound mathematical confidence measures [42, 43]. In medical expert's opinion, a disease classifier should not only provide an easy-to-interpret measure of confidence for its decisions, but also take into account asymmetrical misclassification costs for false positive and false negative classifications. The PNN model provides sound statistical confidences for its decisions, and it is able to model asymmetrical misclassification costs. PNN have shown excellent classification performance in other applications, each pattern neuron to the outcome of the network is explicitly defined and accessible, and has a precise interpretation [44]. Recent analysis also suggests that microarrays can identify unanticipated similarities and diversity in individual patients and thus may be useful in augmenting risk-group stratification in the future [45].

The selection of the data set not only raises the classification accuracy rate successfully, but also identifies genes whose expression correlated with biological features of childhood leukemia. Table 6 shows that several marker genes were identified. Zyxin is the highly correlated gene associated with acute myelogenous leukemia. Zyxin is the prototype of LIM domain proteins that are localized primarily at focal adhesion plaques. Zyxin may regulate gene transcription by interaction with transcription factors. In some cases, misregulation of nuclear functions of zyxin proteins appears to be associated with the pathogenic process of leukemia. The amount of Zyxin expression has independent prognostic significance and could be used in diagnosis to help the difficult evaluation of the malignancy potential of leukemia [46]. Furthermore, cytogenetic studies showed that a novel t(3;11)(q28;q23) chromosomal translocation in a patient who developed leukemia, and the LPP

Table 6. Top 50 genes selected by signal-to-noise metric (10-fold cross-validation tests) to be most highly associated with each tumor type.

Data set	Top 50 genes
Leukemia (ALL/AML)	ZyxinX95735, DF D compon, CD33 CD33 a, CST3 Cystat, PRG1 Proteo, CTSD Cathep, APLP2 Amylo, PPGB Protec, LYN V-yes-1, GLUTATHIONE, FAH Fumaryl, ATP6C Vacuo, MYL1 Myosin, SPTAN1 Spec, IGB Immunog, CD19 geneM8, Terminal tr, GB DEF PROTEASOME, CCND3 Cycli, MB-1 geneU0, TCF3 Transc
Embryonal CNS multi-type tumors	D13900_atMi, D83735_atAd, M12125_atSk, D26070_at"T
Medulloblastoma morphology	D14530_at40, HG3543-HT37, X53331_atMG, X65724_atND, X67951_atPA, U63842_atNe, X64330_atAT, HG1980-HT20
Medulloblastoma treatment outcome	U08998_atTA

(lipoma preferred partner) gene on 3q28 was identified as the MLL fusion partner. LPP contains substantial identity to the focal adhesion protein, zyxin [47].

Our work also explored PNN model and the associated feature selection strategy to boost the classification accuracy for the embryonal CNS tumors data sets. Differential expression of a number of genes previously known to be involved in the pathogenesis of brain tumors and several novel genes been identified by our results and other similar reports [48]. However, the CNS embryonal tumours represent a heterogeneous group of tumours about which little is known biologically, and the pathogenesis are more complex than those of leukemia. A more comprehensive study is undertaken to construct the disease progressing model at the transcriptional level.

The most important goal of gene expression experiments is to build sparse and interpretable classification models which hopefully lead to a better understanding of the disease mechanism on a molecular level. Since information on the complete genome sequence is available, gene expression profiles can be used in the classification of gene interaction patterns, and the identification of cis-regulatory elements and sets of co-regulated genes. By comparing the genome-wide effects of Zyxin family proteins on treated yeast, murine and human cells, we will find functional similarity and specificity of co-expression genes to construct transcriptional model for ALL/AML and present data using phylogenetic-type tree clustering algorithm (manuscript in preparation).

On the other hand, the transcriptional expression profiles can be used to classify drugs and their mode of action, if a drug interacts with and inactivates a specific cellular protein, the phenotype of the drug-treated cell should be very similar to the phenotype of a cell in which the gene encoding the protein has been genetically inactivated, usually through mutation. Those key molecules for disclosing the molecular mechanisms of AML may be the potential targets for chemotherapeutic agents development [49, 50]. We will join the gene expression

levels to a database with measures of cancer susceptibility to anticancer agents, to see how the baseline RNA expression levels in the cell lines correlated with the inhibition of growth of these same cell lines to thousands of anticancer agents (manuscript in preparation).

6. Conclusion

In order to predict the class of cancer, we have demonstrated the usefulness of the PNN model using a marker genes selection method based on their correlation with the class distinction. The experimental results show that the PNN network yields 100% recognition accuracy in classifying the type of leukemia cancer with the subset of 50 most informative genes. The yield of the precise prediction is mainly contributed by the characteristic of manifest class distinction and the smaller spread around class means possessed by the selected test data.

Not only this work reveals that the PNN is well suited for the classification if the chosen training and test data possess large value of signal-to-noise statistic, but the PNN is faster than other neural networks, such as the MLP, when the data size keeps in the range of tens to hundreds because no iterative procedures are needed in the training process of the PNN. Furthermore, we also show that the PNN also performs well when it is used to classify the data sets with multiple tumor types, such as embryonal CNS tumors. Noticeably, the PNN model provides some novel insights into the analysis and interpretation of microarray data. The gene selection method identified several genes whose expression correlated with biological features of childhood leukemia. It is going to undertake the biological model construction and cancer drugs screening which is inspired by this study.

Our future work will focus on exploring other appropriate machine learning models and the associated feature selection strategy to boost the classification accuracy for the data sets with lower signal-to-noise statistic, such as the medulloblastoma treatment outcome data set, and those with multi-categories, such as embryonal CNS tumors.

Acknowledgements

This research was partially supported by National Science Council under Grant NSC 92-2213-E-026-001.

References

1. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, **286** (1999), 531–537.
2. Slonim, D. K., Tamayo, P., Mesirov, J. P., Golub, T.R. and Lander, E. S.: Class prediction and discovery using gene expression data. In: *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, 2000, pp. 263–272.

3. Toure, A. and Basu, M.: Application of neural network to gene expression data for cancer classification. In: *Proceedings of the 2001 International Joint Conference on Neural Networks*, Vol. 1, 2001, pp. 583–587.
4. Ryu, J. and Cho, S. -B.: Gene expression classification using optimal feature/classifier ensemble with negative correlation. In: *Proceedings of the 2002 International Joint Conference on Neural Networks*, Vol. 1, 2002, pp. 198–203.
5. Su, Min, Basu, M. and Toure, A.: Multi-domain gating network for classification of cancer cells using gene expression data. In: *Proceedings of the 2002 International Joint Conference on Neural Networks*, Vol. 1, 2002, pp. 286–289.
6. Xu, R., Anagnostopoulos, G. and Wunsch, D.: Tissue classification through analysis of gene expression data using A new family of ART architectures. In: *Proceedings of the 2002 International Joint Conference on Neural Networks*, Vol. 1, 2002, pp. 300–304.
7. Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z.: Tissue classification with gene expression profiles. In: *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, 2000, pp. 54–64.
8. Kuramochi M. and Karypis G.: Classification using expression profiles: A feasibility study. In: *Proceedings of the IEEE 2nd International Symposium on Bioinformatics and Bioengineering*, 2001, pp. 191–200.
9. Azuaje, F.: Gene expression patterns and cancer classification: A self-adaptive and incremental neural approach, In: *Proceedings of the 2000 IEEE EMBS International Conference on Information Technology Applications in Biomedicine*, 2000, pp. 308–313.
10. Azuaje, F.: Making genome expression data meaningful: Prediction and discovery of classes of cancer through a connectionist learning approach. In: *Proceedings of IEEE International Symposium on BioInformatics and Biomedical Engineering*, 2000, pp. 208–213.
11. Principe, J., Euliano, N. and Lefebvre, W.: *Neural and Adaptive Systems: Fundamentals Through Simulations,* John Wiley & Sons, Inc., 2000.
12. Moghaddam, B. and Yang, M.-H.: Gender classification with support vector machine. In: *Proceedings of 4th IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 306–311.
13. Dasarathy, B. V.: *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques,* IEEE Computer Society Press, 1991.
14. Anagnostopoulos, G. C. and Georgiopoulos, M.: Ellipsoid ART and ARTMAP for incremental unsupervised and supervised learning. In: *Proceedings of the 2001 International Joint Conference on Neural Networks*, Vol. 2, 2001, pp. 1221–1226.
15. Ben-Dor, A., Shamir, R. and Yakhini, Z: Clustering gene expression patterns, *J. Comput. Biol.*, **6** (1999), pp. 281–297.
16. Freund, Y. and Schapire, R. E: A decision-theoretic generalization of on-line learning and an application to boosting, *J. Comput and Syst Sci.* **55** (1997), pp. 119–139.
17. Fritzke, B.: A self-organizing network that can follow non-stationary distributions. In: *Proceedings of the 2001 International Joint Conference on Neural Networks*, 1997, pp. 613–618.
18. Kasuba, T.: Simplified fuzzy ARTMAP, *AI Expert*, 1993, pp. 19–25.
19. Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturia, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y. H., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, M. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. N., Mesirov, J. P. E. S., Lander and Golub, T., R.: Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature*, **415 (**2002), 436–442.
20. Specht, D. F.: Probabilistic neural networks and the polynomial adaline as complementary techniques for classification, *IEEE Trans. Neural Networks,* **1** (1) (1990), 111–121.

21. Patra, P. K., Nayak, M., Nayak, S. K. and Gobbak, N. K.: Probabilistic neural network for pattern classification, In: *Proceedings of the 2002 International Joint Conference on Neural Networks*, Vol. 2, 2002, pp. 1200–1205.
22. Huang, D. S. and Zhao, W.: A novel method for improving the classification capability of radial basis probabilistic neural network classifiers, In: *Proceedings of the 2002 International Joint Conference on Neural Networks*, Vol. 1, 2002, pp. 102–106.
23. Jin, X., Srinivasan, D. and Ruey, L. C.: Classification of freeway traffic patterns for incident detection using constructive probabilistic neural networks, *IEEE Trans. Neural Networks,* **12** (5) (2001), 1173–1187.
24. Menhaj, M. B. and Delgosha, F.: A soft probabilistic neural network for implementation of bayesian classifiers, In: *Proceedings of the 2001 International Joint Conference on Neural Networks*, Vol. 1, 2001, pp. 454–458.
25. Zhu, Y., Zhao, Y., Palaniappan, K., Zhou, X. and Zhuang, X: Optimal bayesian classifier for land cover classification using landsat tm data, *IEEE 2000 International Geoscience and Remote Sensing Symposium*, Vol. 1, 2000, pp. 447–450.
26. Liu, H. and Setiono, R.: Chi2: Feature selection and discretization of numeric attributes, In: *Proceedings of the IEEE 7th International Conference on Tools with Artificial Intelligence*, 1995, pp. 338–391.
27. Witten, I. H. and Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 2000.
28. Hall, M. and Holmes, G.: Benchmarking attribute selection techniques for discrete class data mining, *IEEE Trans. Knowledge Data Eng.* **15** (3) 2003, in press.
29. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection, In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 1137–1143.
30. Ambroise, C., and McLachlan, G. J.: Selection bias in gene extraction on the basis of microarray gene-expression data, In: *Proc. Natl. Acad. Sci.* **9** (10) (2002), 6562–6566.
31. Kohavi, R. and John, G. H.: Wrappers for feature subset selection, *Artif. Intell.*, **97** (1997), 273–324.
32. Liu, H., Li, J. and Wong, L.: A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns, *Genome Informatics*, **13** (2002), 51–60.
33. Skrypnyk, I.: Comparison of feature selection strategies for hearing impairments diagnostics. In: *Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems*, 2002, pp. 231–236.
34. Wu, S. and Flach, P. A.: Feature selection with labelled and unlabelled data. In: *Proceedings of ECML/PKDD'02 Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*, 2002, pp. 156–167.
35. Jung, S. H. Kang, S. H. and Ahn, C.: Chi-Square test for R $\times$ C contingency tables with clustered data, *J. Biopharmaceut. Stat.* **13** (2) (2003), 241–251.
36. Tay, F. E. H. and Shen, L.: A modified Chi^2 algorithm for discretization, *IEEE Trans. Knowledge Data Eng.* **14**, (3) (2002), 666–670.
37. Luan, W. P., Lo, K. L. and Yu, Y. X.: ANN-based pattern recognition technique for power system security assessment, In: *Proceedings of International Conference on Electric Utility Deregulation and Restruct. and Power Technologies*, 2000, pp. 197–202.
38. Zelic, I., Kononenko, I., Lavrac, N. and Vuga, V.: Induction of decision trees and bayesian classification applied to diagnosis of sport injuries, *J. Medi. Syst.* **21** (6) (1997), 429–444.
39. Kukar, M., Kononenko, I. and Silvester, T.: Machine learning in prognosis of the femoral neck fracture recovery, *Arti. Intell. Medi.* **8** (5) (1996), 431–451.

40. Vapnik, V. N. *The Nature of Statistical Learning Theory*, N.Y.: Springer, 1995.
41. Agresti, A. and Coull, B. A.: Approximate is better than Exact for interval estimation of binomial proportions, *Am. Statistician*, **52**, (1998), 119–126.
42. Szabo, A. Boucher, K. Carroll, W. L., Klebanov, L. B., Tsodikov, A. D. and Yakovlev, A. Y.: Variable selection and pattern recognition with gene expression data generated by the microarray technology, *Math. Biosci.,* **176** (2002), 71–98.
43. Chen, Y. Kamat, V. Dougherty, E. R. Bittner, M. L. Meltzer, P. S. and Trent, J. M.: Ratio statistics of gene expression levels and applications to microarray data analysis, *Bioinformatics,* **18** (2002), 1207–1215.
44. Bijlani, R., Cheng, Y., Pearce, D. A., Brooks, A. I., Ogihara, M.: prediction of biologically significant components from microarray data: Independently consistent expression discriminator (ICED), *Bioinformatics*, **1**, (2003), 62–70.
45. Moos, P. J., Raetz, E. A., Carlson, M. A., Szabo, A., Smith, F. E., Willman, C., Wei, Q., Hunger, S. P., Carroll, W. L.: Identification of gene expression profiles that segregate patients with childhood leukemia, *Clini. Cancer Rese*. **8** (2002), 3118–3130.
46. Wang, Y., Gilmore, T. D.: Zyxin and paxillin proteins: Focal adhesion plaque LIM domain proteins go nuclear, *Biochem. Biophys. Acta,* **17** (2003), 115–120.
47. Daheron, L., Veinstein, A., Brizard, F., Drabkin, H., Lacotte, L., Guilhot, F., Larsen, C. J., Brizard, A., Roche, J.: Human LPP gene is fused to MLL in a secondary acute leukemia with at(3;11) (q28;q23), *Genes Chromosomes Cancer*, **31** (2001), 382–389.
48. Hunter, S. B., Moreno, C. S.: Expression microarray analysis of brain tumors: What have we learned so far, *Frontier Biosci.* **1** (2002), 74–82
49. Okutsu, J., Tsunoda, T., Kaneta, Y., Katagiri, T., Kitahara, O., Zembutsu, H., Yanagawa, R., Miyawaki, S., Kuriyama, K., Kubota, N., Kimura, Y., Kubo, K., Yagasaki, F., Higa, T., Taguchi, H., Tobita, T., Akiyama, H., Takeshita, A., Wang, Y. H., Motoji, T., Ohno, R., Nakamura, Y.: Prediction of chemosensitivity for patients with acute myeloid leukemia, according to expression levels of 28 genes selected by genome-wide complementary DNA microarray analysis, *Mole. Cancer Ther*. **1** (2002), 1035–1042.
50. Cheok, M. H., Yang, W., Pui, C.-H., Downing, J. R., Cheng, C., Naeve, C. W., Relling, M. V. and Evans, W. E.: Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells, *Nature Genetics*, **34** (2003), 85–90.