Neural and MTS Algorithms for Feature Selection

*Chao-Ton Su and **Te-Sheng Li

*Department of Industrial Engineering and Management
National Chiao Tung University, Hsinchu, Taiwan
**Department of Industrial Engineering and Management
Minghsin Institute of Technology, Hsinchu, Taiwan

Abstract

The relationships among multi-dimensional data (such as medical examination data) with ambiguity and variation are difficult to explore. The traditional approach to building a data classification system requires the formulation of rules by which the input data can be analyzed. The formulation of such rules is very difficult with large sets of input data. This paper first describes two classification approaches using back-propagation (BP) neural network and Mahalanobis distance (MD) classifier, and then proposes two classification approaches for multi-dimensional feature selection. The first one proposed is a feature selection procedure from the trained back-propagation (BP) neural network. The basic idea of this procedure is to compare the multiplication weights between input and hidden layer and hidden and output layer. In order to simplify the structure, only the multiplication weights of large absolute values are used. The second approach is Mahalanobis-Taguchi system (MTS) originally suggested by Dr. Taguchi. The MTS performs Taguchi's fractional factorial design based on the Mahalanobis distance as a performance metric. We combine the automatic thresholding with MD; it can deal with a reduced model, which is the focus of this paper. In this work, two case studies will be used as examples to compare and discuss the complete and reduced models employing BP neural network and MD classifier. The implementation results show that proposed approaches are effective and powerful for the classification.

Key words: Feature selection, Artificial neural networks; Backpropagation; Mahalanobis distance; Automatic thresholding; Mahalanobis-Taguchi system.

1. Introduction

In most cases, many data (such as medical examination data) are characterized by multi-dimensional information with ambiguity and variation, which make it difficult to explore the relationships among them. The traditional approach to building an expert system requires the formulation of rules by which the input data can be analyzed. The formulation of such rules is quite difficult with large sets of input data. To resolve the difficulty, artificial neural network (ANN) has been applied as an alternative to traditional rule-based expert system. ANNs can be well trained by input data (e.g. examination results) and output response (e.g. signs or symptoms). Moreover, ANN has been applied to various pattern classifications in many fields. Giacinto et al. (2000) combined the neural and statistical algorithms for supervised classification of remote-sensing images. Sutter and Jrus (1997) used ANN to classify and quantify organic vapors. A neural network was trained by Pfurtscheller et al. (1996) to classify electroencephalogram (EEG) patterns in a real-time fashion. Kwak and Lee (1997) illustrated the capacity of ANN to classify and predict the health status of HIV/AIDS patients. In short, ANN has demonstrated its capability of pattern classification including diagnosis of diseases. Hence, ANN has been found to be more helpful than a traditional approach in dealing with the multi-dimensional data.

Even though it can basically approximate any function, an ANN still has a few problems such as time-consuming convergence, overfitted training, high complexity in computation and trained NNs are black boxes from the designer's point of view (Tsukimoto, 2000). The advanced computer hardware has contributed to the substantial improvement in the speed and ease of computation. However, the other problems are closely related to the neural network structures and training algorithms. Several algorithms (Tsukimoto, 2000) have been developed by researchers trying to understand the neural network structure. Through knowing the structures, deleting the redundant connections and extracting the rules, neural network users can learn in advance what the neural networks have discovered and how the neural networks predict. Therefore, users can apply the neural networks to some critical problems. In this study, we will analyze and evaluate the complete and reduced neural network models applicable to the multi-dimensional data. The reduced neural network will be obtained from an feature selection procedure. The basic idea of this procedure is to compare the multiplication weights between input and hidden layer and hidden and output layer. After eliminating the unimportant input nodes, the neural network still possesses the robust potential for classification.

On the other hand, the Mahalanobis distance (MD) is one of the minimum distance classifiers. In contrast to the Euclidean distance classifier, MD also considers the correlation among the multi-dimensional variables. MD is a very sensitive and useful way to determine the similarities among a group of data and detect any unknown data or outlier from a large data set. As MD has been known for some time, in fact, MD was successfully applied to spectral discrimination in analytical chemistry and pattern recognition in computer vision. Brown et al. (1998) used Mahalanobis distance metric based on multi-dimensional vector to evaluate the performance of three 100-compound spectra classifications. Shah and Gemperline (1990) qualitatively identified raw materials by near infrared (NIR) spectroscopy using a Mahalanobis distance classification method. Kato et al. (1999) proposed the asymmetric Mahalanobis distance as a fine classification technique for pattern recognition of handwritten Chinese and Japanese characters.

The Mahalanobis-Taguchi system (MTS) suggested by Dr. Taguchi, combining the Mahalanobis distance and Taguchi method, was used in the area of quality engineering (Taguchi, 1998). MTS can deal with a reduced model which determines the significant factors in the experiments by comparing the signal to noise (S/N) ratio between different levels. It is shown that the MTS is a robust approach by giving the noise to the training multi-dimensional data.

This paper first describes two classification approaches using back-propagation (BP) neural network and Mahalanobis distance (MD) classifier, and then proposes two classification approaches for multi-dimensional feature selection. The first approach proposed is an feature selection procedure from the trained back-propagation (BP) neural network. The second approach is Mahalanobis-Taguchi system (MTS) which combines the automatic thresholding approach with MD as a performance metric. We will illustrate the effectiveness of the proposed approaches in complete and reduced models by using the real-world medical exam data and industrial product data.

2. Methodologies

2.1 Back-probagation neural networks

Neural networks emerged as an attractive alternative to pattern classification. The feedforward, multi-layer neural networks used for pattern classification are typically trained via BP neural networks (Giacinto et al., 2000). Once trained, a BP network can be evaluated very quickly, which is an advantage during the optimization phase.

The BP neural networks consist of layers of neurons interconnected in such a way that information is stored in the weights assigned to the connections. Network learning is designed to determine an appropriate set of connection strengths that facilitates the activation of these processing units to achieve a desired state that mimics a given set of sampled patterns. The output level of a neuron is determined by a sigmoid function.

The BP modeling algorithm begins with a random set of weights. The input vector is first normalized within a reasonable range (say between -3 and 3), and then the initial weight of each neuron connection is used to calculate the activation of each neuron. Next, the calculated output vector is compared with the measured output. The network attempts to minimize the quadratic error sum between output neurons. Minimization is approaching via the gradient descent approach, by which the weights are adjusted in the direction of decreasing error. The basic weight update expression is known as the generalized delta rule. In this manner, the BP provides a recursive procedure to adjust the connection weights of all neurons in the network. The procedure is iterated until the desired convergence is achieved. The root mean square error (RMSE) or classification accuracy can be applied as a decision criterion for the network convergence.

Procedure 1: Induction of a BP classifier

Phase I: Training process
Step 1: Collect a set of observed data.
Step 2: Divide the data into training and testing data sets.
Step 3: Set the training parameters (e.g., learning rate and momentum).
Step 4: Train the different neural network structures.
Step 5: Select a trained network with the highest classification accuracy.
Phase II: Classification process
Step 1: Obtain the unknown input data.
Step 2: Present the data to the trained network that is selected from step 5 in phase I.
Step 3: Obtain the classification results.

2.2 Feature selection from the trained BP neural network

Although neural networks have been widely applied in many different fields and solved some problems, they are still considered as black box. The number of input and output nodes depends on the complexity of the problem. In this manner, the structure of the neural networks will also become more complicated and require more time to converge in the training process. Hence, the unimportant input nodes should be neglected. The simplified structure of neural networks can improve the interpretability and predictability of the network. However, only a few researchers proposed some relevant algorithms in literature. Sarle (2000) proposed a feature selection method that is based on the weights between layers. The essence of this method is to compare the multiplication weights between input and hidden layer and hidden and output layer. In order to simplify the structure, only the multiplication weights with large absolute values are used. This method is employed in our study. The weights between input and hidden layers and the weights between output and hidden layers may produce more significant impact than the others, therefore, users should take both weights into consideration. The equation of the sum of the absolute multiplication values for each node *i* is illustrated as follows:

$$Node_i = \sum_j \sum_k \left| W(X_i, H_j) \bullet W(H_j, O_k) \right| \qquad (1)$$

where X_i is *ith* input node

H_j is *jth* hidden node

O_k is *kth* output node

$W(X_i, H_j)$ is the weight between *ith* input node and *jth* hidden node

$W(H_j, O_k)$ is the weight between *jth* hidden node and *kth* output node

The algorithm, which can select the important input nodes in a BP network, is illustrated in procedure 2 as follows:

Procedure 2: Feature selection for a neural network

Step 1: Calculate the sum of the absolute multiplication values of weights between input and hidden layers and hidden and output layers for each input node.

Step 2: Sort the values obtained from Step 1 in a descending sequence and select a cutoff value.

Step 3: Find the corresponding input features which are larger than cutoff value selected from Step 2.

Step 4: Train the neural network by the selected input features and compare the classification results with that of all the original input features. If the classification result of selected input feature is satisfactory, then stop; otherwise back to Step 2 to select a new cutoff value.

2.3 Mahalanobis distance classifier

The Mahalanobis distance is not only a highly sensitive means of classifying multi-dimensional data but also an effective approach to determining the similarities between a set of variables from an unknown sample and a set of variables measured from a collection of known samples. In describing each training class, the Mahalanobis distance not only takes the variance-covariance matrix of each training class into account but also uses a hyperellipsoid whose boundary is defined by the standard deviation away from the class centroid. The Mahalanobis distance is a measure of the distance from an individual point to the centroid of a population in multi-dimensional space based on the assumption of a multivariate normal distribution $N(\mu, \Sigma)$ for the population. The Mahalanobis distance is expressed as follows:

$$D^2(x_i) = (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \qquad (2)$$

where x_i is a vector representing a set of responses or quality characteristics of a sample; μ is a vector representing the centroid of the population; and Σ is the variance-covariance matrix of the population. In practice, the estimated mean vector and variance-covariance matrix from a sample size of n are used to replace μ and Σ. An unknown sample is classified as a normal individual if its normalized Mahalanobis distance is less than the cutoff value, i.e., within the range of the normal data used in constructing the classification model. In contrast to the Euclidean distance metric, the Mahalanobis distance is scaled in each dimension by the standard deviation of the scores in that dimension. Therefore, the Mahalanobis distance is given equal weight in each dimension and is not necessarily dominated by the features having large range of values.

When calculating the Mahalanobis distance for a classification problem, the distributions of MD are commonly overlaying with each other. Thresholding method is critical because it would clearly divide the two different classes. This study adopts an automatic thresholding method based on a histogram-modeling method (Guo and Pandit, 1998) and a histogram-based statistical measure. Automatic multi-thresholding is achieved by maximizing the inter-class variance of the histogram clusters defined by the methodology of data-dependent systems. Let n_i denote the number of D^2 equal to i. The total number of data set in a Mahalanobis space N is then equal to the sum of n_i. Thus, the probability of $D^2 = i$ is defined as $p_i = n_i/N$. For a two-class model, the MD space is divided into two classes C_1 and C_2, representing samples of acceptable examples and samples of unacceptable examples by a threshold t: $C_1=\{1,2,...,t\}$ and $C_2=\{t+1, t+2,..., M\}$. The discriminating criterion maximizes the inter-class variance σ^2_B with respect to t. Otsu (1979) derived the optimal threshold t^*.

$$\sigma^2_B(t^*) = \max \sigma^2_B(t) \tag{3}$$

The range of variable t is restricted to

$$S^* = \{t; n_1^H < t < n_2^L\} \tag{4}$$

where n_1^H and n_2^L are the boundary D^2 of the histogram clusters. The maximum σ_2^B decreases when n_1^H and n_2^L restrict the range of variable t. The inter-class variance is expressed as

$$\sigma_B^2 = w_1(\mu_1 - \mu_T)^2 + w_2(\mu_2 - \mu_T)^2 \tag{5}$$

where the probabilities of class occurrence are obtained by

$$w_1 = \Pr(C_1) = \sum_{i=1}^{t} p_i ; \quad w_2 = \Pr(C_2) = \sum_{i=t+1}^{M-1} p_i \tag{6}$$

and the class mean and the total mean are defined as

$$u_1 = \sum_{i=1}^{t} ip_i / w_1; \qquad u_2 = \sum_{i=t+1}^{M} ip_i / w_2; \qquad u_T = \sum_{i=1}^{M} ip_i \tag{7}$$

According to the above formula, the optimal threshold can be easily obtained in a short time. The detailed classification procedure is summarized as follows:

Procedure 3: Induction of a MD classifier

Phase I: Training process

Step 1: Collect a set of data obtained from multiple items (including normal and abnormal conditions).

Step 2: Normalize the individual data under normal conditions.

Step 3: Calculate the variance-covariance matrix of the normalized data.

Step 4: Calculate the MD space.

Step 5: Plot the distribution of MD space.

Step 6: Determine the threshold of the MD space, t^*.

Phase II: Classification process

Step 1: Obtain the unknown input data.

Step 2: Normalize the data based on the means and variance under normal conditions.

Step3: Calculate the Mahalanobis Distance D^2.

Step4: Obtain the classification results, i.e. if $D^2 > t^*$, then this pattern belongs to an abnormal set. Otherwise, the pattern belongs to a normal set with similar properties.

2.4 Mahalanobis-Taguchi System

The Mahalanobis-Taguchi System (MTS) is a method that combines Mahalanobis Distance (MD) and Taguchi's Method. Basically, MTS utilizes the Orthogonal Array (OA) to arrange the design of experiments, the MD as a signal factor involved in the computation of S/N ratio, and the S/N ratio as the criteria for selection of main factors in the experiment. MTS has two main objectives, one of which is to deal with the multi-dimensional classification, and the other is to reduce the number of multiple variables in the experiment. From the practical viewpoint, if an experiment can be conducted with fewer factors but reach the satisfactory results, it will significantly decrease the cost, labor and time taken in the experiment. In the following, we will present the MTS procedure step by step.

Procedure 4: Induction of a MTS classifier

Step 1: Collect n normal data, which are characterized by K-dimensional items.

Step 2: Calculate the D^2 for each data.

Step 3: Let be the signal in Taguchi's dynamic system, i.e.

$$M_i = \sqrt{D_i^2},\ i = 1 \cdots n.$$

Step 4: Divide K items into L items and $(K\text{-}L)$ items; L items need to be further studied in Orthogonal Array and $(K\text{-}L)$ items represent the absolutely necessary items due to theoretic consideration or learned from previous experience.

Step 5: Select an appropriate OA and assign the L items into the column of OA. In the OA table, use two levels for each factor; 1 means not using this factor and 2 means using this factor in the experiment.

Step 6: Calculate the MD space for each row of OA. In case of all 1's in the row, it means that all the factors are not used in the experiments and we will calculate the MD space characterized by the other *(K-L)* items. In contrary, if both1's and 2's exist in the same row, we will use the factors corresponding to 2's column plus *(K-L)* items to create MD space.

Step 7: Based on the MD space and the responses, calculate the S/N ratio for each row in the OA as shown below.

$$\gamma = \sum_{i=1}^{n} M_i^2 \tag{8}$$

$$\beta = \frac{\sum_{i=1}^{n} M_i y_i}{\gamma} \quad (y_i \text{ is the response, } i=1\cdots n) \tag{9}$$

$$S_\beta = \frac{(\sum_{i=1}^{n} M_i y_i)^2}{\gamma} \tag{10}$$

Total sum of square $\mathrm{S_T} = \sum_{i=1}^{n} \mathrm{y_i^2}$ (11)

Error variation $S_e = S_T - S_\beta$ (12)

Error variance $V_e = \dfrac{S_e}{n-1}$ (13)

S/N ratio $\eta = 10\log \dfrac{\frac{1}{\gamma}(S_\beta - V_e)}{V_e}$ (14)

Step 8: Plot the factor effects and determine the important items in the experiment.

Step 9: Use Procedure 3 to obtain the MD space, determine the threshold and reach the final classification results.

3. Illustration

The above four methods are compared in terms of classification capability in the following two case studies. One is medical examination data set used to diagnose liver malfunctions, and the other is glass identification data set used to classify different type of glass.

3.1 Example One: Liver Disease Diagnosis

3.1.1 Problem description

Thirty features including sex, age, Neutrophil, Lymphocyte, Monocyte, Eosinophil, ..., etc. were collected from a hospital located in Taipei, Taiwan. These items are characterized by multi-dimensional information about the current health status of patients and have difficulty in diagnosing other diseases based on such a large amount of information. Until now, the relationship between the medical examination data and liver malfunction symptom is still ambiguous. As mentioned earlier, disease diagnosis can be considered as a classification task. Thus, our proposed approaches are employed to classify whether an individual has liver disease.

In this study, medical examination data were collected from 952 individuals including those who had normally functioning or malfunctioning liver patients for almost one year. After the data collection, we chose 89 individuals who were labeled as normal individuals and 79 individuals labeled as liver malfunction patients based on the medical history of the patients judged by doctors. Cumulatively, the general examination data of these 168 people were used as inputs in this study.

3.1.2 Complete modeling by Procedure 1

First, 30 examination items (including age and gender) about the 54 normal individuals and 47 patients with liver disease were used for supervised BP neural network training. On the other hand, the testing data set also included the testing results of 35 normal individuals and 32 liver patients from the model created by the training data set. Thirty items of attributes for each individual were used as the input, and the output was a node representing the situation whether an individual was a liver patient or not. In this manner, the structure of BP neural network can be expressed as 30-X-1, where X denotes the number of hidden nodes. In this study, NeuralWorks Professional II package (Neural Ware, Inc. 1992) was used to perform computation so that the structure could reach the highest classification rate. After trial and error, the optimal structure for the network was 30-6-1 and its classification rate was 96.04%, i.e., 4 out of the 101 individuals were misclassified. Correspondingly, the classification rate for testing data set was 89.55%, i.e., 7 out of the 67 individuals were misclassified.

3.1.3 Reduced modeling by Procedure 2

The above section indicates that the BP neural network structure 30-6-1 is the optimal structure in this case due to the highest accuracy rate in testing data set. The weights of 30-6-1 neural network structure are utilized to calculate the sum of absolute multiplication values of weights between input and hidden layer and hidden and output layer for each node. The results are listed in Table 1. However, the cutoff value in the computation results is determined by the data distribution of the whole data set. In this case, 9.5 can be considered as the cutoff value in Table 2 since the classification accuracy is lower than that of the complete model by less than 10%. Thus, 15 physical examination items will remain in the BP neural networks, and the other 15 examination items will be deleted from the

Table 1. Computation results by equation (1)

Code	Items	Results	Code	Items	Results	Code	Items	Results	Code	Items	Results	Code	Items	Results
A	Gender	6.301	G	Eosinophil	3.704	M	MCH	6.187	S	GPT	**59.144**	Y	Chol.	7.862
B	Age	**14.937**	H	Basophil	**14.648**	N	MCH	7.117	T	r-GT	**19.501**	Z	TG	**26.625**
C	WBC	7.376	I	RBC	1.995	O	Platelet	7.903	U	T-Bil.	7.096	A'	HDL-C	2.884
D	Neutrophil	**10.017**	J	Hemoglobin	7.274	P	GLUAC	**16.511**	V	D-Bil.	**12.239**	B'	Cr	9.251
E	Lymphocyte	**16.692**	K	Hematocrit	3.600	Q	ALK-P	**18.997**	W	T-Protein	**18.027**	C'	BUN	**13.512**
F	Monocyte	**10.561**	L	MCV	7.360	R	GOT	**52.149**	X	Albumin	3.703	D'	Uric Acid	**18.377**

BP networks.

The next step is to retrain the neural network by the remaining 15 input nodes bolded in Table 2. The training results for the 15 input nodes are listed in Table 3. From Table 3, the neural network structures 15-5-1, 15-9-1 and 15-10-1 have better classification accuracy in both training and testing data set. Therefore, these three structures are suitable for classification of liver disease. It is shown that the input nodes reduced from thirty to fifteen nodes do not decrease but increase the classification accuracy. Therefore, Procedure 2 can indeed find the more important input nodes in a large set of data.

3.1.4 Complete modeling by Procedure 3

The MD classifier employed medical examination data from 54 normal individuals and 47 liver patients (training data) to create the Mahalanobis space. For example, the vector (x_1, x_2, x_3,···, x_{30}) = (0.619, 1.054, -0.063,···, -0.226) represents the first person's exam data, which has been normalized by mean and standard deviation of each examination item. The Mahalanobis distance D^2 is calculated by equation (2), where Σ^{-1} denotes the variance-covariance matrix of all examination items. After calculating the remaining data set, all the D_i^2's of the training set form a specific distribution. Figure 1 plots the distribution of normal individuals and potential liver patients.

Table 2. Accuracy rate of classification for reduced BP neural network structures

Structure	Training	Testing
15-3-1	0.9703	0.9254
15-5-1	0.9802	0.9254
15-7-1	0.9703	0.9254
15-9-1	0.9802	0.9254
15-10-1	0.9802	0.9254
15-12-1	0.9505	0.8955
15-15-1	0.9703	0.9254

Note: Learning rate = 0.25; Momentum = 0.95

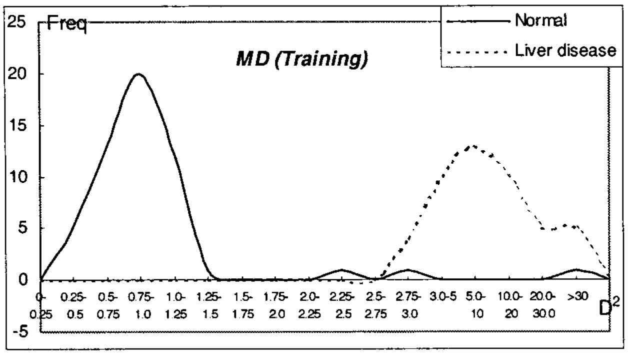

Figure 1. Distribution of D^2 based on training data

Finally, automatic thresholding is determined by maximizing the inter-class variance.

After calculation, the threshold was set at 2.5 and the other attempts are listed in Table 4. According to this table, the maximum variance between classes was 7.051 while the threshold was set at 2.5. The testing data were the same as the data used for testing in BP neural network. Consequently, the classification rate is 95.52%, and type I error and type II error are 0% (0/35) and 9.38% (3/32), respectively, for the testing data set. As a result, Procedure 3 outperforms Procedure 1.

3.1.5 Reduced modeling by Procedure 4

Firstly, the medical examination data of normal individuals is used to compute the signal of the multi-dimensional information system. The signal is measured by the square root of D_i^2's. Secondly, we determine the items required by the diagnosis of liver disease. Among the thirty items, there are 16 items (including A, B, D, E, K, L, O, P, Q, R, S, T, Y, Z, A′ and C′ items) suggested by doctors or existing for theoretic considerations. Under such a condition, the remaining 14 items that are arranged in $L_{16}(2^{15})$ Orthogonal Array as shown in Table 5 should be further studied. In this table, it shows that all the 1's in the first row represents all the control factors that will not be used. Hence, the experimental response will be calculated in terms of the other sixteen items. The other rows except

Table 3. Automatic thresholding of MD distribution based on training data

Threshold	Type I error	Type II error	Total Error	Correct rate	VarianceB
1.5~2.25	3	0	3	0.9703	6.468
2.5	2	0	2	0.9801	7.051
2.75	1	4	5	0.9505	6.003

Table 4. $L_{16}(2^{15})$ Orthogonal Array and the S/N ratio

Item	C	F	G	H	I	J	M	N	U	V	W	X	B′	D′		
No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	S/N
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	7.638
2	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	5.525
3	1	1	2	2	2	2	2	1	1	1	1	2	2	2	2	6.053
4	1	1	2	2	2	2	2	2	2	2	2	1	1	1	1	4.664
5	1	2	1	1	1	2	2	1	1	2	2	1	1	2	2	6.076
6	1	2	1	1	1	2	2	2	2	1	1	2	2	1	1	4.580
7	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1	3.830
8	1	2	2	2	2	1	1	2	2	1	1	1	1	2	2	5.872
9	2	1	1	1	2	1	2	1	2	1	2	1	2	1	2	5.964
10	2	1	1	1	2	1	2	2	1	2	1	2	1	2	1	4.959
11	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1	4.198
12	2	1	2	2	1	2	1	2	1	2	1	1	2	1	2	5.757
13	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	3.549
14	2	2	1	1	2	2	1	2	1	1	2	2	1	1	2	6.068
15	2	2	2	2	1	1	2	1	2	2	1	2	1	1	2	5.474
16	2	2	2	2	1	1	2	2	1	1	2	1	2	2	1	4.110

the first row in $L_{16}(2^{15})$ OA Table, are composed of the 1's and 2's on the same row. Their responses will be calculated with the remaining sixteen items plus their corresponding 2's control factors on the row.

The S/N ratios are calculated as shown in the last column of Table 5. Figure 2 illustrates the average S/N ratios used or not used for each item. From Figure 2, all the S/N ratios of not used items are higher than the ones of used items. Hence, all of these 14 items can be deleted from the liver disease diagnosis. The remaining 16 items of 101 individuals are conducted by Procedure 3. Figure 3 plots the distribution of D^2 for the 16 remaining items of training data. The threshold value herein is set at 2.0 when the algorithm previously described. The

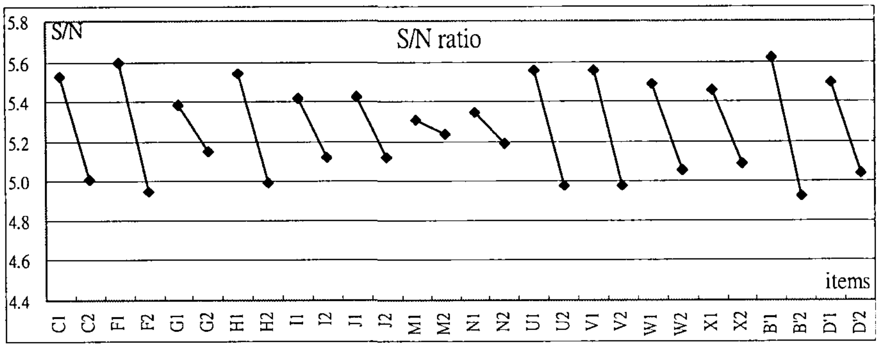

Figure 2. S/N ratio for the items studied (1:not used; 2: used)

Table 5. Automatic thresholding of MD distribution based on training data

Threshold	Type I error	Type II error	Total Error	Correct rate	VarianceB
1.5	2	4	6	0.9405	46.73
2.0	3	2	5	0.9505	52.53
2.5	1	5	6	0.9405	47.02

performance of reduced model by MTS is: the classification rate of testing data is 92.53% and type I error and type II error are 8.57% (3/35) and 6.25% (2/32), respectively. Consequently, the system has reduced 14 items from the original data set, but the classification accuracy still keeps almost the same rate as the complete MD model.

3.2 Example Two: Glass Identification

3.2.1 Problem description

The second data set was tested on glass identification database from the University of California, Irvine repository of machine learning data set (German et al., 1987). The data set consisted of 10 attributes of 214 glass instances. All attributes are continuously valued. The goal of the classifier was to determine, based on the attributes of glass, whether the glass is belong to window glass or non-window glass. The attributes of glass in this case consisted of refractive index, sodium, magnesium, silicon, ..., etc. The number of instances in training set contained 111 window glass samples and 31 non-window glass samples. On the other hand, the testing set consisted of 52 window glass samples and 20 non-window glass samples.

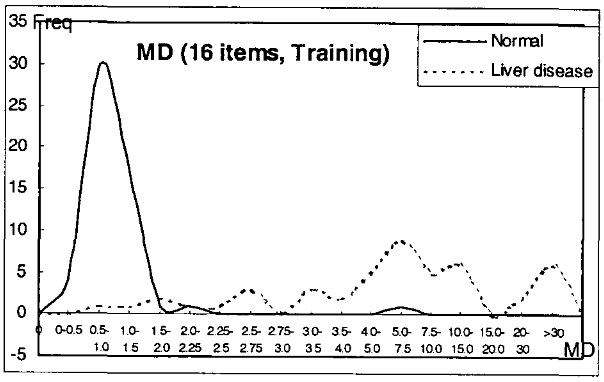

Figure 3. Distribution of D^2 based on the training data in the reduced model

Table 6. The performance of four classification approaches

Methods		Procedure 1 (BP Neural Network)	Procedure 2 (Feature selection)	Procedure 3 (MD classifier)	Procedure 4 (MTS)
Input nodes		30 items	15 items	30 items	16 items
Model/Criteria		30-6-1	15-5-1 15-9-1 15-10-1	t*=2.5	t*=2.0
Classification	Training	96.04%	98.02%	98.02%	95.05%
Rate	Testing	89.55%	92.54%	95.52%	92.54%

3.2.2 Complete modeling by Procedure 1

Same as the previous example, 9 attributes (except the first Id #) of the 142 training samples were used for supervised neural networks training. Nine items of attributes for each sample were used as the input, and the output was a node indicating whether an individual was a window glass or not. In this case, after trying the different neural structures, the optimal structure for the network was 9-5-1 and its classification rate was 97.18%, i.e., 4 out of the 142 samples were misclassified. Correspondingly, the classification rate for testing data set was 94.44%, i.e., 4 out of the 72 samples were misclassified.

3.2.3 Reduced modeling by Procedure 2

The above section indicates that the BP neural network structure 9-5-1 is the optimal structure in this case due to the highest accuracy rate in testing data set. The weights of 9-5-1 neural network structure are utilized to calculate the sum of absolute multiplication values of weights between input and hidden layer and hidden and output layer for each node. The results are listed in Table 7. However, the cutoff value in the computation results is determined by the data distribution of the whole data set. In this case, 14 can be considered as the cutoff value in Table 7 since the classification accuracy is lower than that of the complete model by less than 10%. Thus, 5 features will remain in the BP neural networks, and the other 4 features will be deleted from the BP networks.

The next step is to retrain the neural network by the remaining 5 input nodes bolded in Table 7. The optimal neural

Table 7. Computation results by equation (1) example 2

Code	Items	Results	Code	Items	Results	Code	Items	Results	Code	Items	Results	Code	Items	Results
A	**Na2O**	**15.83**	**C**	**MgO**	**16.38**	**E**	**Al2O3**	**33.40**	**G**	**SiO2**	**24.90**	**I**	**K2O**	8.52
B	**CaO**	2.77	**D**	**BaO**	12.89	**F**	**Fe2O33**	**14.76**	**H**	**Refractive**	7.79			

network structures 5-5-1.The classification results for the remaining 5 input features are 100% for training data and 100% for testing data, respectively. Therefore, BP features selection approach is suitable for classification of glass. It is shown that the input features reduced from nine to five features do not decrease but increase the classification accuracy. Therefore, Procedure 2 can indeed find the more important input features in a large set of data.

3.2.4 Complete modeling by Procedure 3

The Mahalanobis distance D^2 is calculated by equation (2), where Σ^{-1} denotes the variance-covariance matrix of all features. After calculating the remaining data set, all the D_i^2's of the training set form a specific distribution. Next, automatic thresholding is determined by maximizing the inter-class variance. After calculation, the threshold was set at 1.75. According to this threshold, the classification results are 90.14% for training data and 90.27% for testing data, respectively. As a result, Procedure 1 outperforms Procedure 3.

3.2.5 Reduced modeling by Procedure 4

Firstly, the glass data of window is used to compute the signal of the multi-dimensional information system. Secondly, we determine the items required by classification of glass. Among the nine features, there are 4 features (including A, B, C, F) existing for theoretic considerations. Under such a condition, the remaining 5 items that are arranged in $L_8(2^7)$ Orthogonal Array. The S/N ratios are calculated. The S/N ratios of D, E, G and H not used features are higher than the ones of used feature. Hence, these 4 items can be deleted from the glass classification. The remaining 5 items of 142 samples are conducted by Procedure 3. The distribution of D^2 can be plotted (not shown). The threshold value herein is set at 3.25 when the algorithm previously described. The performance of reduced model by MTS is that the classification results are 90.48% for training data and 94.44% for testing data, respectively. Consequently, the system has reduced 4 items from the original data set, but the classification accuracy outperforms the complete MD model.

Table 8. The performance of four classification approaches

Methods		Procedure 1 (BP Neural Network)	Procedure 2 (Feature selection)	Procedure 3 (MD classifier)	Procedure 4 (MTS)
Input nodes		9 items	5 items	9 items	5 items
Model/Criteria		9-5-1	5-5-1	t*=1.75	t*=3.25
Classification	Training	97.18%	100.00%	90.14%	90.84%
Rate	Testing	94.44%	100.00%	90.27%	94.44%

4. Comparison and discussion

Two BP neural networks (complete/reduced models) and two MD classification approaches (complete/reduced models) were trained and tested on the two case studies. The performance of these four classification approaches is summarized in Table 6 and Table 8. It is notable that the approaches employed in this study are robust and effective for the classification of liver diseases and glasses because the classification accuracy are higher than ninety percent for the testing data. Moreover, another interesting finding from this study is that the classification accuracy of the reduced BP neural network is several percent higher than that of the complete BP neural network. In addition, the classification accuracy of the reduced MD approach is less than that of the complete MD classification approach in example 1 but is higher than that of the complete MD approach in example 2. The following conclusions can be drawn from the above results:

1. Based on Table 6 and Table 8, obviously, no one approach can dominate any other approach in the illustrating case studies. In example 1, Mahalanobis distance complete model has the better performance than other approach. However, in the second example, BP feature selection approach outperforms than other approach. Therefore, no one classifier can definitely perform better than other classifier. But, all the approaches illustrated herein are really robust and effective, the classification accuracy are higher than ninety percentage.
2. Most importantly, the BP network can map the relationship between input features and output classification. The importance of each input feature can be illustrated by Procedure 2. According to the priority of the input feature, we can determine how many features should remain in the BP neural network to keep almost the same accuracy. In this study, the accuracy of testing data is 92.54% and 100% for the example 1 and example 2, respectively. Both accuracy of reduced models are higher than complete models. Our results demonstrate that the feature selection procedure is a powerful and robust approach for classifying the clinical diseases and glasses.
3. The Mahalanobis distance takes into account the variance-covariance matrix of each training class and describes each class by using a hyperellipsoid whose boundary is defined by the standard deviation away from the class centroid. The MD is an extremely sensitive method for classifying multi-dimensional data such as the medical examination data in this study. In addition, the decision on the automatic thresholding to adopt based on

the maximum variance between classes is important in MD classification because the threshold determines the misclassification. In this study, the percentage of correct classification by example 1 and example 2 is 92.54% and 94.44%, respectively. MTS can delete unimportant items in the multi-dimensional information system. The results also demonstrate that the MTS do slightly reduce some important information in the multi-dimensional data set. Hence, by adopting the Mahalanobis distance and Taguchi method as a measure metric, the MTS can be carried out by evaluating the tradeoff between classification accuracy and loss of critical information.

4. Both BP and MD (no matter the complete or the reduced models) approaches can classify the multi-dimensional data space, such as medical examination data and glass with multiple attributes. In this work, the faster identification of a disease implies a greater likelihood of finding a cure for the patient. Results in this study provide a valuable reference to doctors for diagnosis at the early stage of a disease. However, more advanced measurements and tests should be developed at the subsequent stages to perform more complicated classification of liver diseases.

5. Conclusion

By adopting BP (complete /reduced) and MD (complete /reduced) approaches, this study classifies the multi-dimensional examination data for diagnosis of a liver disease and glass classification. In the first example, the results show that the reduced BP network (15 items) is better than the complete BP network. The best way to elucidate the above results is the feature selection procedure that can actually classify the items into the important and unimportant classes. In contrast to the results of BP network, the complete MD classifier provides slightly more information than the reduced MD model (16 items) because MTS has lost some information during the procedure. In the second example, the results show that the reduced BP network (5 items) is better than all the other classifier. Correspondingly, MTS classifier also outperformed than the MD even MTS has reduced some features during the procedure. The analytical results indicate that these four classifiers are all robust and effective methods to classify the medical data and industrial product in this study. However, how many variables can be reduced in a MD model without serious impact on the classification accuracy is a subject for future research.

Acknowledgement

This paper was supported in part by the National Science Council, Taiwan, R.O.C., under NSC 91-2213-E-009-072

References

1. Andrews, R. and Diederich, J., 1996. Rules and Networks. Proceedings of Rule Extraction Trained Artificial Neural Networks Workshop, AISB.
2. Andrews, R., Diederich, J. and Tickle, A. B., 1995. Survey and critque of techniques for extracting rules from trained artificial neural networks. Knowledge-Based System 8, 373-389.
3. Antony, J., 2000. Multi-response optimization in industrial experiments using Taguchi's quality loss function and principal component analysis. Quality Reliability Engineering. International 16, 3-8.
4. Brown, C. W. and Lo, S. C., 1998. Chemical information based on neural network processing of Near-IR Spectra. Analytic Chemistry 70, 2983-2990.
5. Giacinto, G., Roli, F. and Bruzzone, L., 2000. Combination of neural and statistical algorithms for supervised classification of remote-sensing images. Pattern recognition letters 21, 385-397.
6. Guo, R. and Pandit, S. M., 1998. Automatic threshold selection based on histogram lodes and a discriminant criterion. Machine vision and applications10, 331-338.
7. Kato, N., Suzuki, M., Omachi, S., Aso, H. and Nemoto, Y., 1999. A handwritten character recognition system using directional element feature and asymmetric Mahalanobis distance. IEEE Transaction on Pattern Analysis and Machine Intelligence. 21, 258-263.
8. Kwak, N. K. and Lee, C., 1997. A neural network application to classification of health status of HIV/AIDS patients. Journal of Medical System 21. 87-97.
9. Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Transaction on Systems Man and Cybernetics. SMC 9, 62-66.
10. Pfurtscheller, G., Kalcher, J. and Nerper, C., 1996. On-line EEG classification during externally paced hand movements using a neural network-based classifier. Electroncephalogr. Clinical Neruophsiology 99, 416-425.
11. Sarle, W. S., 2000. How to measure importance of inputs?. SAS Institute Inc., Cary, NC, USA. ftp://ftp.sas.com./ put/neural/importance.html.
12. Sette, S., Boullart, L., Langenhove, L. V. and Kiekens, P., 1997. Optimizing the fiber-to-yarn production process with a combined neural network/genetic algorithm approach. Textile Research Journal 67, 84-92.
13. Shah, N. K. and Gemperline, P. J., 1990. Combination of the Mahalanobis

distance and residual variance pattern recognition techniques for classification of Near-Infrared reflectance spectra. Analytic Chemistry 62, 465-470.

14. Sutter, J. M. and Jurs, P. C., 1997. Neural network classification and quantification of organic vapors based on fluorescence data from a fiber-optic sensor array. Analytic Chemistry 69, 856-862.
15. Taguchi, Genichi, 1998. Mathematics for Quality Engineering. Journal of Quality Engineering Forum 6, 5-10.
16. Tsukimoto, H., 2000. Extraction Rules from Trained Neural Networks. IEEE Transactions on Neural Networks 11, 377-389.
17. Younis, K. S., Rogers, T. K. and DeSimio, M. P., 1996. Vector quantization based on dynamic adjustment of Mahalanobis distance. Aerospace and Electronics Conference Proceedings of IEEE of 1996 1.2, 858-862.