An investigation of neural networks in thyroid function diagnosis

Guoqiang (Peter) Zhang [a] and Victor L. Berardi [b]

[a] *Department of Decision Sciences, College of Business Administration, Georgia State University, Atlanta, GA 30303, USA*
E-mail: gzhang@kentvm.kent.edu
[b] *Department of Management, College of Business, Bloomsburg University, Bloomsburg, PA 17815, USA*

Received November 1997; revised March 1998

We investigate the potential of artificial neural networks in diagnosing thyroid diseases. The robustness of neural networks with regard to sampling variations is examined using a cross-validation method. We illustrate the link between neural networks and traditional Bayesian classifiers. Neural networks can provide good estimates of posterior probabilities and hence can have better classification performance than traditional statistical methods such as logistic regression. The neural network models are further shown to be robust to sampling variations. It is demonstrated that for medical diagnosis problems where the data are often highly unbalanced, neural networks can be a promising classification method for practical use.

1. Introduction

The thyroid gland is one of the most important organs in the body as thyroid hormones are responsible for controlling metabolism. As a result, thyroid function impacts on every essential organ in the body. The most common thyroid disorder is an underactive thyroid, known as hypothyroidism, in which the thyroid does not produce enough hormone. Less frequently, the thyroid produces too much hormone which is known as hyperthyroidism. Approximately 2–3% of the general population in the United States suffers from either hypothyroidism or hyperthyroidism [28]. Groups most commonly affected by thyroid dysfunction include women and the elderly where as many as 5–10% of those in these groups may be affected. The seriousness of thyroid disorders should not be underestimated as thyroid storm (an episode of severe hyperthyroidism) and myxedema coma (the end stage of untreated hypothyroidism) may lead to death in a significant number of cases.

The correct diagnosis of thyroid dysfunctions based on clinical and laboratory tests often proves difficult. One reason stems from the nonspecific nature of many thyroid symptoms. This is especially true of hypothyroidism where symptoms such as lethargy, confusion, weight gain, and poor memory are easily confused with other psychiatric and medical conditions. The problem is often exacerbated in older patients whose symptoms are sometimes masked or attributed to other medical conditions [1]. While laboratory tests have become more accurate and are helpful in diagnosing thyroid abnormalities (the positive predictive rates of some tests are recently reported to be over 90%. See [27, p. 339]), the results are still not very satisfactory across all situations. The difficulty in diagnosis comes from the inconsistency in test results across patients and other factors such as pregnancy, drug interactions, nonthyroidal illnesses, and psychiatric problems which are all known to affect the thyroid hormone levels measured in the laboratory tests [7,34].

Thyroid dysfunction diagnosis also presents a challenge to traditional statistical methods because it represents a classification problem with three extremely unbalanced groups. Highly unbalanced groups occur commonly in many fields, particularly in medical diagnosis where a small proportion of the population actually has a specific disease. Statistical and other quantitative methods have long been used as decision-making tools in medical diagnosis including thyroid disease detection. These classification methods include both parametric methods such as discriminant analysis and logistic regression and nonparametric models like k-nearest-neighbor and mathematical programming models, as well as various machine learning methods such as CART and ID3. One major limitation of the traditional statistical models is that they work well only when the underlying assumptions are satisfied. The effectiveness of these methods depends to a large extent on the various assumptions or conditions under which models are developed. Users must have a good knowledge of both data properties and model capabilities before they can successfully apply the model. For example, Press and Wilson [16] show that if the variables have multivariate normal distributions within classes, discriminant analysis is more efficient than logistic regression. However, if the variables do not have within-class normal distributions, logistic regression is preferable.

Research activities over the last decade have shown that artificial neural networks (ANNs) have powerful pattern classification and pattern recognition ability. They have been used extensively in many different problems including thyroid function diagnosis [26]. The success of neural networks may be attributed to their many unique features of pattern recognition and pattern classification. First, neural networks are universal approximators [4,10,11]. They can approximate any nonlinear function with arbitrary accuracy.

© Baltzer Science Publishers BV

Since any classification procedure seeks to establish a functional relationship between the object group membership and the attributes characterizing the object, the accurate estimation of this relationship is essential for the success of the classifier. Second, as opposed to traditional model-based approaches, neural networks are data-driven methods. They learn from examples with few a priori assumptions about the model or the functional form of the relationship for problems under study. Third, neural networks have generalization ability. After learning the data presented to them (a sample), they can often correctly identify the pattern and generalize it to the unseen part of the population. Generalization capability is often the most important criterion for choosing a classification model. Finally, neural networks are able to provide accurate estimates of posterior probabilities on the membership of an object [18]. It is a well-known fact that posterior probabilities play a critical role in the traditional Bayesian classification theory. These advantages of neural networks explain the numerous successful applications of neural networks reported in the recent literature.

In using neural networks, the entire available data set is usually randomly divided into training and test samples. The training sample is used for neural network model building and the test set is used to evaluate the predictive capability of the model. While this practice is adopted in many studies, the random division of a sample into training and test sets may introduce bias in model selection and evaluation in that the characteristics of the test sample may be very different from those of the training sample. Furthermore, different partitions may also have effects on the model building and classification performance. The estimated classification rate can be very different from the true classification rate particularly when small-sized samples are involved. For this reason, it is a major focus of this paper to use a cross-validation scheme to accurately describe predictive performance of neural networks. Cross-validation is a resampling technique which uses multiple random training and test subsamples. The advantage of cross-validation is that all of the observations or patterns in the available sample are used for testing and most of them are also used for training the model. The cross-validation analysis will yield valuable insights into the reliability of the neural networks with respect to sampling variation.

The remainder of the paper is organized as follows. The next section contains an introduction to neural networks. The link between neural networks and the Bayesian classification theory is discussed in section 3. Section 4 is the methodology section which includes the description of the data set, the design of the neural network model employed and the description of cross-validation study. Results are presented in section 5. The final section contains summary remarks and discussions.

2. An introduction to neural networks

A neural network is a massively parallel system of interconnected computing elements called nodes. Information is processed via the interaction between a large number of nodes where knowledge is not stored in the individual nodes, but rather it is represented by the weights of the connections between the nodes. Figure 1 contains a simple three-layer feedforward network representative of those used in this research. The first or lowest layer is called the input layer where external information enters the network while the last or top layer is called the output layer where the network produces the model solution. The middle layer(s) or hidden layer(s) provide the connections necessary for the ANN to identify complex patterns in the data. All nodes in adjacent layers are connected by arcs from the input layer to the hidden layer to the output layer.

Arc weights are the parameters in a neural network model. As in any statistical model, these parameters need to be estimated before the network can be adopted for further use. Neural network training is a process in which these weights are determined, and hence is the way the network learns. Network training for classification problems is performed via supervised learning in which known outputs and their associated inputs are presented to the net-

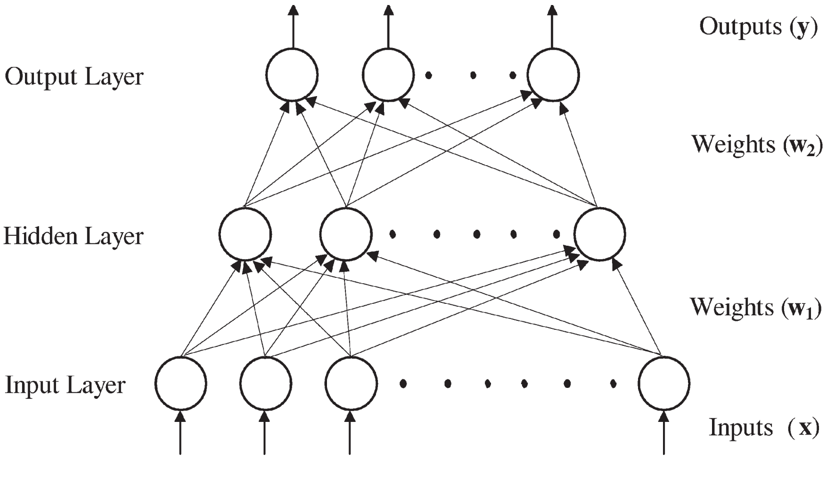

Figure 1. Multi-layer feedforward neural network.

work. The input node activation values are weighted and summed at each hidden layer node. The weighted sum is then transmitted by an appropriate transfer function into the hidden node's activation value, which becomes the input to the output layer nodes. The same computation process is repeated at the output nodes. The network output values are then compared to the known actual values for the purpose of minimizing the differences between network output values and the known target values for all training patterns.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ be a d-vector of attribute values, $\mathbf{y} = (y_1, y_2, \ldots, y_M)$ be the M-dimensional output vector from the network, and $\mathbf{w}_1$ and $\mathbf{w}_2$ be the matrices of linking weights from input to hidden layer and from hidden to output layer, respectively. Then a three-layer neural network is in fact a nonlinear model of the form

$$\mathbf{y} = f_2\big(\mathbf{w}_2 f_1(\mathbf{w}_1\mathbf{x})\big), \qquad (1)$$

where f_1 and f_2 are the transfer functions for the hidden nodes and output nodes, respectively. The most popular choice for f_1 and f_2 is the sigmoid (logistic) function. That is,

$$f_1(x) = f_2(x) = (1 + \exp(-x))^{-1}.$$

Theoretically speaking, any differentiable function can qualify as a transfer function. The reason for using the logistic function is that it is simple, has a number of good characteristics (bounded and monotonically increasing), and bears a better resemblance to real neurons [9]. Klimasauskas [13] suggests that the logistic function should be used for classification problems that involve learning about average behavior.

The purpose of network training is to estimate the weight matrices in (1) so that an overall error measure such as the mean squared errors (MSE) or sum of squared errors (SSE) is minimized. MSE can be defined as

$$\mathrm{MSE} = \frac{1}{M}\frac{1}{N}\sum_{m=1}^{M}\sum_{j=1}^{N}(a_{mj} - y_{mj})^2, \qquad (2)$$

where a_{mj} and y_{mj} represent the target value and network output (activation) at the mth node for the jth training pattern respectively, M is the number of output nodes, and N is the number of training patterns. Network training can therefore be seen as an unconstrained nonlinear minimization problem.

A number of researchers have illustrated the connection of neural networks to traditional statistical methods. Cheng and Titterington [2] make a detailed analysis and comparison of various neural network models with traditional statistical methods. In their paper, perceptrons like the feedforward neural networks are shown to have strong associations with discriminant analysis and regression, and unsupervised networks such as self-organizing neural networks with cluster analysis. Sarle [23] translates neural network jargon into statistical terminology and illustrates the relationship between neural networks and statistical models such as generalized linear models, projection pursuit and cluster analysis. Warner and Misra [30] contrast neural networks to regression models. Schumacher et al. [25] and Vach et al. [29] present a thorough comparison between feedforward neural networks and the logistic regression. The conceptual similarities and discrepancies between the two methods are analyzed. Gallinari et al. [6] study the relations between discriminant analysis and multilayer perceptrons for classification problems. Richard and Lippmann [18] show that neural networks can provide estimates of Bayesian posterior probabilities. Ripley [20,21] discusses the statistical aspects of neural networks and classifies neural networks as one of a class of flexible nonlinear regression methods. White [31] provides statistical properties and asymptotic results for neural network learning. Ciampi and Lechevallier [3] show that the statistical modeling approach provides a good starting point for constructing a neural network model and neural networks, on the other hand, provide a powerful expansion to classical statistical model families.

3. Neural networks and Bayesian classifiers

Bayesian decision theory is the traditional statistical approach to pattern classification that defines the problem in probabilistic terms and assumes that all of the relevant probabilities are known [5]. In applying the theory to an M-group classification problem, consider that each object is associated with a d-vector $\mathbf{x}$ of attributes. Assume that $X \subseteq R^d$ is the sample space which is divided into M subspaces. Letting ω_j represent that an object is a member of group j, the following notations will be used:

$P(\omega_j)$ = the prior probability that a randomly selected object belongs to group j;

$f(\mathbf{x} \mid \omega_j)$ = the probability density function for $\mathbf{x}$ given that its membership is ω_j. This probability function is also known as the likelihood function.

According to Bayes' rule, if we obtain an observation $\mathbf{x}$, the prior probability will be modified into the posterior probability, $P(\omega_j \mid \mathbf{x})$, that object $\mathbf{x}$ belongs to group j by incorporating all this information, that is,

$$P(\omega_j \mid \mathbf{x}) = \frac{f(\mathbf{x}, \omega_j)}{f(\mathbf{x})} = \frac{f(\mathbf{x} \mid \omega_j)P(\omega_j)}{\sum_{j=1}^{M} f(\mathbf{x} \mid \omega_j)P(\omega_j)}, \quad j = 1, 2, \ldots, M.$$

The above Bayes rule shows how observing the value of $\mathbf{x}$ changes the prior probability $P(\omega_j)$ to the posterior probability $P(\omega_j \mid \mathbf{x})$ upon which the classification decision is based.

Suppose that a particular $\mathbf{x}$ is observed and is to be assigned to a group. Let $\lambda_{ij}(\mathbf{x})$ be the cost of misclassifying $\mathbf{x}$ to group i when it actually belongs to group j. Since $P(\omega_j \mid \mathbf{x})$ is the probability that the object belongs to group j given **x**, the expected loss associated with assigning **x** to group i is

$$L_i(\mathbf{x}) = \sum_{j=1}^{M} \lambda_{ij}(\mathbf{x}) P(\omega_j \mid \mathbf{x}), \quad i = 1, 2, \ldots, M. \tag{3}$$

$L_i(\mathbf{x})$ is also known as the conditional risk function. Since **x** will be assigned to only one group, let $L(\mathbf{x})$ be the resultant loss. The objective is to minimize the total expected loss,

$$L = \int_{\mathbf{x} \in X} L(\mathbf{x}) f(\mathbf{x})\, \mathrm{d}x.$$

Function L is minimized when each term $L(\mathbf{x})$ is minimized. This is accomplished by following what is known as the Bayesian decision rule in classification:

$$\text{Decide } \omega_k \text{ for } \mathbf{x} \quad \text{if } L_k(\mathbf{x}) = \min_{i=1,2,\ldots,M} L_i(\mathbf{x}).$$

A loss function of special interest in the literature is known as the symmetrical or zero-one loss function. The zero-one loss function is specified as $\lambda_{ij}(\mathbf{x}) = 0$ for $i = j$, and 1 otherwise. In this case, the conditional risk function of (3) can be simplified to

$$L_i(\mathbf{x}) = \sum_{j \neq i} P(\omega_j \mid \mathbf{x}) = 1 - P(\omega_i \mid \mathbf{x}).$$

Note that $P(\omega_i \mid \mathbf{x})$ is the conditional probability that the correct classification is group i given the feature vector **x**. Therefore, to minimize the average probability of error, we should select the i that maximizes the posterior probability $P(\omega_i \mid \mathbf{x})$. As a result, the Bayesian decision rule becomes

$$\text{Decide } \omega_k \text{ for } \mathbf{x} \quad \text{if } P(\omega_k \mid \mathbf{x}) = \max_{i=1,2,\ldots,M} P(\omega_i \mid \mathbf{x}). \tag{4}$$

The above discussion clearly shows the important role of posterior probabilities in the Bayesian classification decision.

To see the relationship between neural networks and Bayesian classifiers, we need the following theorem [14]:

Theorem 1. Consider the problem of predicting **y** from **x**, where **x** is a d-vector random variable and **y** is an M-vector random variable. The function mapping $F : \mathbf{x} \to \mathbf{y}$ which minimizes the squared expected error

$$E[\mathbf{y} - F(\mathbf{x})]^2 \tag{5}$$

is the conditional expectation of **y** given **x**,

$$F(\mathbf{x}) = E[\mathbf{y} \mid \mathbf{x}].$$

The result stated in the above theorem is the well-known least squares estimation theory in statistics.

In the M-group classification context, if **x** is the observed attribute vector and **y** is the true membership vector, that is

$$\begin{array}{ll} \mathbf{y} = (1,0,0,\ldots,0,0)^t & \text{if } \mathbf{x} \in \text{group}\, 1, \\ \mathbf{y} = (0,1,0,\ldots,0,0)^t & \text{if } \mathbf{x} \in \text{group}\, 2, \\ \vdots & \\ \mathbf{y} = (0,\ldots,0,1,0,\ldots,0)^t & \text{if } \mathbf{x} \in \text{group}\, m, \\ \vdots & \\ \mathbf{y} = (0,0,0,\ldots,0,1)^t & \text{if } \mathbf{x} \in \text{group}\, M, \end{array}$$

where t represents the transpose of a vector, then for $\mathbf{x} \in \text{group}\, m$, there is a unit probability with $\mathbf{y} = (0,\ldots,0,1,0,\ldots,0)^t$ and a zero probability with other **y**'s. Hence, $F(\mathbf{x})$ becomes

$$\begin{aligned} &F(\mathbf{x}) \\ &\quad = E[\mathbf{y} \mid \mathbf{x}] \\ &\quad = (0,\ldots,0,1,0,\ldots)^t\, P(\mathbf{y} = (0,\ldots,0,1,0,\ldots,0)^t \mid \mathbf{x}) \\ &\quad = (0,\ldots,0,P(\mathbf{y} = (0,\ldots,0,1,0,\ldots,0)^t \mid \mathbf{x}),0,\ldots,0)^t \\ &\quad = (0,\ldots,0,P(\omega_m \mid \mathbf{x}),0,\ldots,0)^t. \end{aligned} \tag{6}$$

Equation (6) shows that the least squares estimate for the mapping function in a classification problem is exactly the posterior probability.

As noted earlier, neural networks are universal function approximators. A neural network in a classification problem can be viewed as a mapping function, $F : R^d \to R^M$, where d-dimensional input **x** is submitted to the network and an M-dimensional network output **y** is obtained to make the classification decision. If all the data in the entire population are available for training and the global optimal solution can be found in neural network training, then (2) and (5) are equivalent and neural networks produce exact posterior probabilities in theory. In practice, however, training data is almost always a sample from an unknown population and the global optimal solution can not be guaranteed. Thus it is clear that the network output is actually the estimate of the posterior probability. Hung et al. [12] show that neural networks are able to provide accurate estimates of posterior probabilities in practical applications.

4. Research design and methodology

The purpose of this research is to study how robust the neural network performance is in predicting thyroid disease in terms of sampling variability. Specifically, we are interested in the impact of sampling variability on classification of thyroid patients based on neural network posterior probability estimates. A four-fold cross-validation approach is employed in this study. This section first describes the data set. Then a detailed description of the issues in neural network model building is given. Finally, we illustrate the cross-validation methodology.

4.1. Data set

The data set employed in this study comes from Quinlan [17], which contains information related to thyroid dysfunction. The problem is to determine whether a patient has a normally functioning thyroid, an under-functioning thyroid (hypothyroid), or an over-active thyroid (hyperthyroid). There are 7200 cases in the data set with 3772 from the year 1985 and 3428 from 1986. The hyperthyroid class represents 2.3% (166 cases) of the data points, the hypothyroid class accounts for 5.1% (368 cases) of the observations, while the normal group makes up the remaining 92.6% (6666 cases). This highly unbalanced data set is a notoriously difficult problem for traditional classification methods. For each of the 7200 cases, there are 21 attributes with 15 binary and 6 continuous variables used to determine in which of the three classes the patient belongs. These attributes represent information on patients such as age, sex, health condition, and the results of various medical tests [17]. Appendix A lists these 21 input variables and their descriptions.

4.2. Neural network design

Several factors are important in designing a feedforward neural network for classification problems. These include the input, hidden and output layer configurations as well as the training methodology used. While there are some guidelines for the determination of some factors, no uniformly applicable principles exist to guide neural network design. As a result, much of the neural network architecture is determined by experimentation in practice.

Artificial neural networks are characterized by their architectures. Network architecture refers to the number of layers, nodes in each layer, and the number of connection arcs. It has been shown by Cybenko [4], Hornik [10], and Patuwo et al. [15] that neural networks with one hidden layer are generally sufficient for most problems. All the networks investigated in this study use one hidden layer. The number of input nodes is 21 corresponding to the input attributes in the original data set. Three binary output nodes are employed, corresponding to the three classes of normal, hypothyroid, and hyperthyroid. The target values for each node are either zero or one depending on the desired output class. For example, a target output of 0-0-1 corresponds to a hypothyroid case, 0-1-0 to a hyperthyroid case, and 1-0-0 to a normal patient. The logistic activation function is specified for both hidden and output nodes while the layers are fully connected from input to hidden to output.

The number of hidden nodes is not easy to determine a priori. There are several rules of thumb proposed for determining the number of hidden nodes, but none of them works well for all situations. Hence different numbers of hidden nodes from 5 to 50 were tested. Using the training sample of 3772 cases and 3428 test cases, we found that 10 hidden nodes provided the best test set results in terms of MSE and the test set classification error rate. Therefore, neural networks with 10 hidden nodes are utilized for experimentation in the cross-validation study.

The training methodology is another important factor in designing feedforward neural networks. As noted earlier, the purpose of neural network training is to estimate the node connection weights in order to minimize an error measure such as MSE. The most commonly used training algorithm is the backpropagation [22] which is simply a steepest descent method with constant step size. In this study, we use a more efficient, faster converging training method than standard backpropagation called Rprop for **R**esilient back**prop**agation [19]. The basic principle of the Rprop algorithm is to eliminate the harmful effect of the partial derivative magnitude on the step size of each arc weight and hence give a more efficient search for the solution.

The improvement in the training convergence of the Rprop algorithm compared to that using the standard backpropagation is significant. We find that the Rprop algorithm converges in 100 to 500 epochs with training completed in a matter of seconds while 6000 to 70000 epochs and several hours are required using a backpropagation algorithm.

4.3. Cross validation

Cross-validation methods are used in examining the robustness of classifiers. The simplest of these methods is the single training and testing scheme that is often employed in the medical literature. The original data set is split into two groups and one is used for designing the classifier while the hold-out sample is used for testing purposes. The classification error rate on this test set is then reported as the estimate of the classifier's true error rate. There are several problems related to this method. First, since the number of cases in the test sample is often relatively small in practice, the estimate of the true classification capability of a classifier is often not satisfactory. Second, with the single training and testing method, the training set is much smaller than the whole data set available. Hence the resulting model is unlikely to be the one that would be obtained using all observations. Third, the single training and testing partition may be uncharacteristic of the true underlying population, resulting in large sampling errors or biases.

Resampling techniques such as random subsampling, leaving-one-out, and *k*-fold cross validation can reduce the bias problem by averaging the results over several randomly generated training-and-testing partitions. While the leaving-one-out technique is preferred for small samples, it is computationally difficult for large data sets. We elect to use a four-fold cross validation method in this study. Two cross-validation schemes using mutually exclusive random subsets of the data are used. Both schemes utilize the same training samples while the test samples vary to measure different perspectives of the classifier performance.

The original 7200 cases in the data set are randomly divided into four mutually exclusive partitions of approximately equal size. Stratification of the original cases is employed to ensure that the percentage of each class found in the overall sample is approximately the same in each partition. Four partitions are used here to ensure that each partition has enough cases in the smallest class (hyperthyroid) because if there are too few observations in one group, then almost all classifiers will fail to perform satisfactorily. Training is performed on three of the partitions while the fourth is used for the testing purpose. The process is repeated until each partition has served as the test data. The training sample is used for model fitting and/or parameter estimation while the predictive effectiveness of the fitted model is evaluated in two ways. First, each fitted model is applied to the unseen portion of the data (the hold-out partition). The average classification error of four test partitions (we call them small test samples) is a good indicator of the out-of-sample performance of the classifier. Second, to get a better picture of the true overall capability of the classifier for the unknown population, we test each of the four fitted models using the whole data set. The idea behind this scheme is that the total sample should be more representative of the true underlying population than a small test set. Additionally, when the whole data set is employed as the test sample, sampling variation in the testing environment is completely eliminated as the same sample is tested four different times. Therefore, the variability across the four large test set results reflects only the effect of training sample variation.

The results of neural networks are compared to those of logistic regression. Logistic regression is chosen as the traditional statistical method of comparison because it is often preferred over discriminant analysis in practice [8,16]. Additionally, the statistical property of logistic regression is well understood. We are interested in knowing which method gives more accurate estimates of the posterior probabilities and hence leads to better classification results. Since logistic regression is a special case of the neural network without hidden nodes, it is expected in theory that ANNs will produce more accurate estimates than logistic regression because of their flexible nonlinear modeling capabilities. Logistic regression is implemented using the SAS procedure LOGISTIC (SAS Institute Inc., 1990).

5. Results

Table 1 presents the results for both neural network and logistic regression models when the single partition strategy is used. The 3772 observations from 1985 are used as the training set and the 3428 cases from 1986 are naturally used as the test set. The failure of the logistic regression model is clearly seen from the table. It simply places all cases into the largest group of euthyroid (normal) which represents 92.58% of the total observations. This is a typical phenomenon for many classical statistical classifiers when the data are highly unbalanced. They fail to correctly classify any smaller group member which in reality is usually more important. The use of neural networks improves over the logistic regression significantly not only in the overall classification result but also in the classification performance within each group. In the training set, the overall classification rate increases from 95.70% to 99.42%. The correct classification rates for two smaller groups of hyperthyroid and hypothyroid groups are 95.70% and 96.86%, respectively. From test sample results, neural networks also significantly outperform logistic regression models. The classification rates for individual groups of hyper-, hypo-, and euthyroid are 82.19%, 94.92% and 98.71%, respectively. The relatively low classification rate for the hyperthyroid group may result from the small number of cases in that group (only 73 cases). Overall, we have a 98.16% correct prediction rate for the 1986 data set using neural networks.

Table 1
Single partition results.

Class	Measure	Neural network		Logistic regression	
		Training	Test	Training	Test
Hyperthyroid	% Correct	95.70%	82.19%	0.00%	0.00%
	Correct #	89	60	0	0
Hypothyroid	% Correct	96.86%	94.92%	0.00%	0.00%
	Correct #	185	168	0	0
Euthyroid	% Correct	99.66%	98.71%	100.00%	100.00%
	Correct #	3476	3137	3488	3178
Overall	% Correct	99.42%	98.16%	92.47%	92.71%
	Correct #	3750	3365	3488	3178

Cross-validation results for the training set, the small test set, and the large test set are given in tables 2(a)–2(c), respectively. Again, logistic regression models do not provide good discrimination particularly for the important smaller groups of hyper- and hypo-thyroid at each cross-validation subsample. Neural networks, however, exhibit consistently greater classification capabilities not only in the overall classification result but also in the individual group identification. It can be seen from these tables that both neural networks and logistic regression models show high robustness in the overall classification performance. But neural networks provide better classification results not only in the training sample but also in the test samples.

For the training sample, table 2(a) shows that neural networks can achieve high classification rates for all individual groups. Across different training samples, the results are quite stable. The variability in the classification results for the smallest group of hyperthyroid is slightly higher than that for the other two groups. This can be explained because the hyperthyroid group has far fewer observations than the other two groups. The large number of cases in the euthyroid group explains the highly robust classification results for this group as well as the overall classification rates among different training samples.

Tables 2(b) and 2(c) contain classification results for small test samples and large test samples. As mentioned earlier, the small test sets are used to evaluate the neural network predictive capability since the observations in each test set are not used in the model building process while the large test sets are employed to measure the variability among different training sets. Overall, we see the effectiveness of neural networks in classifying unseen objects.

Table 2

(a) Cross-validation performance results of the training subsamples.

Method	Measure	Subsample 1				Subsample 2				Subsample 3				Subsample 4			
		Hyper	Hypo	Euth	Overall	Hyper	Hypo	Euth	Overall	Hyper	Hypo	Euth	Overall	Hyper	Hypo	Euth	Overall
Neural network	% Correct	91.94	97.10	99.50	99.20	94.35	94.93	99.52	99.19	88.80	96.38	99.48	99.07	93.60	96.74	99.48	99.20
	Correct #	114	268	4974	5356	117	262	4975	5354	111	266	4974	5351	117	267	4974	5358
Logistic regression	% Correct	0.00	0.00	100.00	92.59	0.00	0.00	100.00	92.59	0.00	0.00	100.00	92.59	0.00	0.00	100.00	92.59
	Correct #	0	0	4999	4999	0	0	4999	4999	0	0	5000	5000	0	0	5000	5000

(b) Cross-validation performance results of the predictive performance for small test sets.

Method	Measure	Subsample 1				Subsample 2				Subsample 3				Subsample 4			
		Hyper	Hypo	Euth	Overall	Hyper	Hypo	Euth	Overall	Hyper	Hypo	Euth	Overall	Hyper	Hypo	Euth	Overall
Neural network	% Correct	73.81	91.30	99.46	98.45	83.33	92.39	99.16	98.45	75.61	92.39	99.58	98.67	92.68	97.83	98.80	98.61
	Correct #	31	84	1658	1773	35	85	1653	1773	31	85	1659	1775	38	90	1646	1774
Logistic regression	% Correct	0.00	0.00	100.00	92.56	0.00	0.00	100.00	92.56	0.00	0.00	100.00	93.86	0.00	0.00	100.00	93.86
	Correct #	0	0	1667	1667	0	0	1667	1667	0	0	1666	1666	0	0	1666	1666

(c) Cross-validation performance results of the estimation of true classification rates for large test sets.

Method	Measure	Subsample 1				Subsample 2				Subsample 3				Subsample 4			
		Hyper	Hypo	Euth	Overall	Hyper	Hypo	Euth	Overall	Hyper	Hypo	Euth	Overall	Hyper	Hypo	Euth	Overall
Neural network	% Correct	87.35	95.65	99.49	99.01	92.16	94.29	99.44	98.99	85.54	95.38	99.50	98.97	93.37	97.01	99.31	99.06
	Correct #	145	352	6632	7129	152	347	6628	7127	142	351	6633	7126	155	357	6620	7132
Logistic regression	% Correct	0.00	0.00	100.00	92.58	0.00	0.00	100.00	92.58	0.00	0.00	100.00	92.58	0.00	0.00	100.00	92.58
	Correct #	0	0	6666	6666	0	0	6666	6666	0	0	6666	6666	0	0	6666	6666

The effects of sampling variation on the classification performance of hypothyroid and euthyroid groups as well as the overall performance are very small. The range of the overall classification rates is only 0.22% for the small test samples and 0.09% for the large test samples. In the largest euthyroid group, the classification rate ranges from 98.80% to 99.58% for the small test set case and from 99.31% to 99.50% for the large test samples. The variability in classification rates for the hypothyroid group is slightly higher, ranging from 91.30% to 97.83% for the small test samples. In both test cases, there is relatively high variability in the classification rate across different samples for the smallest hyperthyroid group. For example, using subsample 1, we have only 73.81% classification rate for the correct identification of hyperthyroid patients. Using subsample 4, however, a high classification rate of 92.68% is achieved.

Comparing the results for small test sets in table 2(b) and those for large test sets in table 2(c), we make the following observations. First, the variability in results across the four large test samples is smaller than that of the small test set. As pointed out earlier, this is to be expected because the large test set is the same for each of the four different training sets and the variability in the test results reflects only the difference in the training set. Second, the neural network performance improves from small test sets to large test sets. The explanation lies in the fact that neural networks have much better classification rates in the training samples.

Table 3 compares classification performance of the neural network model using the simple partition test set and the small cross-validation test sets. The classification rate for the cross-validation set is the average across four test subsamples. Cross-validation achieves higher overall classification rate and euthyroid classification rate than the simple partition method. However, for the smaller groups, the single partition has higher classification rate although the difference is not very significant.

Table 3

Comparison of neural network test results.

Class	Single partition	Cross-validation
Hyperthyroid	82.19%	81.36%
Hypothyroid	94.92%	93.48%
Euthyroid	98.71%	99.25%
Overall	98.16%	98.55%

6. Discussion

In this paper we have investigated the potential of neural networks in thyroid dysfunction diagnosis. Thyroid disease identification is an important yet difficult task from both clinical diagnosis and statistical classification points of view. The large number of interrelated patient attributes as well as extremely unbalanced groups in the thyroid diagnosis problem complicate the relationship between these attributes and the patient true group membership, which causes poor performance for traditional model-based statistical methods. Artificial neural networks, being a flexible modeling technique for complex function mapping, show promise in the thyroid disease diagnosis.

Many successful studies have been reported in the literature using neural networks for classification including medical diagnosis. However, the connection between neural networks and traditional statistical methods are often not fully understood. In this paper we present the basic framework of understanding the role of neural networks in classification problems. We show that the neural network outputs are estimates of the posterior probabilities – which play an important role in traditional Bayesian classification theory. The results in this research and previous studies clearly show the superiority of neural networks over traditional statistical methods in the estimation of the posterior probabilities and hence in classification performance.

In this paper, we have examined the robustness of neural networks in thyroid diagnosis with respect to sampling variations. Model robustness has important managerial implications particularly when the model is used for prediction purposes. A useful model is the one which is robust across different samples or time periods. The cross-validation technique described in this study provides decision makers with a method for examining predictive validity and hence the usefulness of the classification method. From the point of view of medical diagnosis, employing a classifier with high robustness and reliability in different sampling situations is a very critical issue. Our results show that neural networks are both robust and accurate methods for the task of diagnosing thyroid dysfunction. Not only can they provide excellent overall classification rate, they are also able to identify the more important, harder-to-classify smaller group members. It is also encouraging to note that overall the variation across samples in training and test classification rates are reasonably small. The classification results for individual groups, however, are quite sensitive to the number of observations in each group. Our results for the three thyroid function groups clearly show that for larger groups with relatively more observations, better classification results are often obtained. Therefore, increasing the number of cases in small groups will improve the performance of neural network classifiers. The above discussion also suggests that the traditional single partition method is valid for neural network model building and evaluation when the overall classification rate is of primary concern. However, when robust identification of the small group member is more critical, the cross-validation approach should be used.

Future research will focus on the following issues. First, there is a need to develop a variable selection method via neural networks to choose, from a large number of attributes, the best subset of variables. The regression-based method to select variables may not be appropriate for many complex medical diagnosis problems due to highly nonlinear relationships in the data. Second, there is a possibility to further improve classification by combining neural networks and traditional statistical classification methods. Statistical methods may provide a basis for neural network model selection. Finally, the development of cost-based neural network decision model should be more useful and appropriate for medical diagnosis. The impact of unequal misclassification costs on the classification performance should be investigated.

Appendix A: Input variables used for thyroid diagnosis

	Input	Description
1	Age	Patient age in years
2	Gender	Patient gender
3	Illness indicator	Patient reports malaise
4	Pregnancy indicator	Patient is pregnant
5	Thyroid surgery	Patient has history of thyroid surgery
6	Iodine 131	Patient is currently receiving iodine 131 treatment
7	Hypothyroid indicator	Patient responses indicate likelihood of hypothyroidism
8	Hyperthyroid indicator	Patient responses indicate likelihood of hyperthyroidism
9	Lithium treatment	Patient is on lithium treatment
10	Thyroxine indicator 1	Patient on thyroxine treatment
11	Thyroxine indicator 2	Patient thyroxine treatment status unknown or unreported
12	Goitre	Patient has goitre
13	Antithyroid indicator	Patient is on antithyroid medication
14	Tumor	Patient has thyroid tumor
15	Hypopituitary	Patient is hypopituitary
16	Psycological indicator	Patient has psychological symptoms
17	TSH results	TSH test results
18	T3 results	T3 test results
19	TT4 results	TT4 test results
20	T4U results	T4U test results
21	FTI value	FTI calculated from TT4 and T4U values

Acknowledgements

An earlier version of this paper has been presented at the *1997 Midwest Decision Sciences Institute Annual Meeting* in Indianapolis, Indiana. We would like to thank the participants for the helpful discussion as well as three anonymous referees for their constructive comments and suggestions.

References

[1] M. Bahemuka and H.M. Hodkinson, Screening for hypothyroidism in elderly inpatients, British Medical Journal 2 (1975) 601–603.

[2] B. Cheng and D.M. Titterington, Neural networks: A review from a statistical perspective, Statistical Sciences 9 (1994) 2–54.

[3] A. Ciampi and Y. Lechevallier, Statistical models as building blocks of neural networks, Communications in Statistics: Theory and Methods 26 (1997) 991–1009.

[4] G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematical Control Signals Systems 2 (1989) 303–314.

[5] R.O. Duda and P. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).

[6] P. Gallinari, S. Thiria, F. Badran and F. Fogelman-Soulie, On the relations between discriminant analysis and multilayer perceptrons, Neural Networks 4 (1991) 349–360.

[7] L.A. Gavin, The diagnostic dilemmas of hyperthyroxinemia and hypothyroxinemia, Advances in Internal Medicine 33 (1988) 185–204.

[8] F.E. Harreli and K.L. Lee, A comparison of the discriminant analysis and logistic regression under multivariate normality, in: *Biostatistics: Statistics in Biomedical, Public Health, and Environmental Sciences*, ed. P.K. Sen (North-Holland, Amsterdam, 1985).

[9] G.E. Hinton, How neural networks learn from experience, Scientific American (September, 1992) 145–151.

[10] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Networks 4 (1991) 251–257.

[11] K. Hornik, M. Stinchcombe and H. White, Multilayer feedforward networks are universal approximators, Neural Networks 2 (1989) 359–366.

[12] M.S. Hung, M.Y. Hu, B.E. Patuwo and M. Shanker, Estimating posterior probabilities in classification problems with neural networks, International Journal of Computational Intelligence and Organizations 1 (1996) 49–60.

[13] C.C. Klimasauskas, Applying neural networks. Part 3: Training a neural network, PC-AI (May/June, 1991) 20–24.

[14] A. Papoulix, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York, 1965).

[15] E. Patuwo, M.Y. Hu and M.S. Hung, Two-group classification using neural networks, Decision Science 24 (1993) 825–845.

[16] S.J. Press and S. Wilson, Choosing between logistic regression and discriminant analysis, Journal of American Statistical Association 73 (1978) 699–705.

[17] J. Quinlan, Simplifying decision trees, International Journal of Man-Machine Studies 27 (1987) 221–234.

[18] M.D. Richard and R.P. Lippmann, Neural network classifiers estimate Basyesian a posteriori probabilities, Neural Computation 3 (1991) 461–483.

[19] M. Riedmiller and H. Braun, A direct adaptive method for faster backpropagation learning: The Rprop algorithm, in: *Proceedings of the IEEE International Conference on Neural Networks*, 1993.

[20] B.D. Ripley, Statistical aspects of neural networks, in: *Networks and Chaos—Statistical and Probabilistic Aspects*, eds. O.E. Barndorfe-Nielsen, J.L. Jensen and W.S. Kendall (Chapman and Hall, London, 1993) pp. 40–111.

[21] B.D. Ripley, Neural networks and related methods for classification (with discussion), Journal of Royal Statistical Society, Series B 56 (1994) 409–456.

[22] D.E. Rumelhart, G.E. Hinton and R.J. Williams, Learning internal representations by error propagation, in: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, eds. D.E. Rumelhart and J.L. Williams (MIT Press, Cambridge, 1986).

[23] W.S. Sarle, Neural networks and statistical models, in: *Proceedings of the 19th Annual SAS Users Group International Conference*, 1994.

[24] SAS Institute Inc., *SAS/STAT User's Guide, Version 6*, 4th ed., Vol. 2 (SAS Institute Inc., Cary, NC, 1990).

[25] M. Schumacher, R. Robner and W. Vach, Neural networks and logistic regression: Part I, Computational Statistics & Data Analysis 21 (1996) 661–682.

[26] P.K. Sharpe, H.E. Solberg, K. Rootwelt and M. Yearworth, Artificial neural networks in diagnosis of thyroid function from vitro laboratory tests, Clinical Chemistry 39 (1993) 2248–2253.

[27] H.C. Sox, M.A. Blatt, M.C. Higgins and K.I. Marton, *Medical Decision Making* (Butterworths, Boston, 1988).

[28] W.M. Tunbridge et al., The spectrum of thyroid disease in a community: the Eidkham survey, Clinical Endocrinology 7 (1977) 481–493.

[29] W. Vach, R. Robner and M. Schumacher, Neural networks and logistic regression: Part II, Computational Statistics & Data Analysis 21 (1996) 683–701.

[30] B. Warner and M. Misra, Understanding neural networks as statistical tools, The American Statistician 50 (1996) 284–293.

[31] H. White, Learning in artificial neural networks: A statistical perspective, Neural Computation 1 (1989) 425–464.

[32] H. White, Some asymptotic results for learning in single hidden layer feedforward networks, Journal of American Statistical Association 84 (1989) 1008–1013.

[33] T.J. Wilke, Estimation of free thyroid hormone concentrations in the clinical laboratory, Clinical Chemistry 32 (1986) 585–592.

[34] E.T. Wong and M.W. Steffes, A fundamental approach to the diagnosis of diseases of the thyroid gland, Clinical Laboratory Medicine 4 (1984) 655–670.