ELSEVIER

Expert Systems with Applications 27 (2004) 133–142

Expert Systems with Applications

www.elsevier.com/locate/eswa

Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines

Shieu-Ming Chou[a], Tian-Shyug Lee[b,*], Yuehjen E. Shao[c], I-Fei Chen[b]

[a]*Department of Nursing Chang-Jung Christian University, Tainan, Taiwan, ROC*
[b]*Graduate Institute of Management, Fu-Jen Catholic University, Hsin-Chuang, No 510 Chung-Cheng Rd, Taipei 24205, Taiwan, ROC*
[c]*Department of Statistics and Information Sciences, Fu-Jen Catholic University, Hsin-Chuang, Taipei, Taiwan, ROC*

Abstract

Data mining is a very popular technique and has been widely applied in different areas these days. The artificial neural network has become a very popular alternative in prediction and classification tasks due to its associated memory characteristics and generalization capability. However, the relative importance of potential input variables and the long training process have often been criticized and hence limited its application in handling classification problems. The objective of the proposed study is to explore the performance of data classification by integrating artificial neural networks with the multivariate adaptive regression splines (MARS) approach. The rationale under the analyses is firstly to use MARS in modeling the classification problem, then the obtained significant variables are used as the input variables of the designed neural networks model. To demonstrate the inclusion of the obtained important variables from MARS would improve the classification accuracy of the networks, diagnostic tasks are performed on one fine needle aspiration cytology breast cancer data set. As the results reveal, the proposed integrated approach outperforms the results using discriminant analysis, artificial neural networks and multivariate adaptive regression splines and hence provides an efficient alternative in handling breast cancer diagnostic problems.

Keywords: Data mining; Breast cancer; Classification; Neural networks; Multivariate adaptive regression splines

1. Introduction

Modern medical facilities are equipped with monitoring, collecting and other devices which can provide inexpensive ways to collect and store data in their information systems. Huge amount of data stored in these databases need special techniques for processing, analyzing, and effective use of them before these data can be helpful supports in handling medical related decision-making problems. Data mining (DM), sometimes referred to as knowledge discovery in database (KDD), is a systematic approach to find underlying patterns, trend, and relationships buried in data. According to Curt (1995), the methodologies consist of data visualization, machine learning, statistical techniques, and deductive database. And the related applications using these methodologies can be summarized as classification, prediction, clustering, summarization, dependency modeling, linkage analysis, and sequential analysis (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Data mining has drawn serious attention from both researchers and practitioners due to its applications in decision support, financial forecasting, fraud detection, marketing strategy, process control, medical research and other related fields (Cabena, Hadjinaian, Stadler, Verhees, & Zanasi, 1998; Chen, Han, & Yu, 1996; Fayyad et al., 1996; Lee, Sung, & Chang, 1999; Ngan, Wong, Lam, Leung, & Cheng, 1999; Pendharkar, Rodger, Yaverbaum, Herman, & Benner, 1999).

Breast cancer, a very common and serious cancer for women, affects almost one in every seven women in the United States (Wingo, Tong, & Bolden, 1995). One of the most commonly used methods in detecting breast cancer is mammography. However, literature has reported that radiologists show considerable variation in interpreting a mammography (Elmore et al., 1994). On the other hand, fine needle aspiration cytology (FNAC) is also widely adopted in the diagnosis of breast cancers. But, according to Fentiman (1998), the average correct identification rate of FNAC is only about 90%. It is therefore an absolute necessity to develop better identification tools in recognizing breast cancers. Owing to the above-mentioned needs,

* Corresponding author. Tel.: +886-22903-1111x2905; fax: +886-22901-8475.
E-mail address: badm1004@mails.fju.edu.tw (T.-S. Lee).

0957-4174/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2003.12.013

several researchers have used statistical and artificial intelligence techniques to successfully 'predict' breast cancer (Kovalerchuck, Triantaphyllou, Ruiz, & Clayton, 1997; Pendharkar et al., 1999). Basically, the objective of these identification techniques is to assign patients to either a 'benign' group that does not have breast cancer or a 'malignant' group who has strong evidence of having breast cancer. And hence the breast cancer diagnostic problems are basically in the scope of the more general and widely discussed classification problems (Anderson, 1984; Dillon & Goldstein, 1984; Hand, 1981; Johnson & Wichern, 2002).

Generally, discriminant analysis and logistic regression are two most commonly used data mining techniques to construct classification models. However, linear discriminant analysis (LDA) has often been criticized due to its assumption about the categorical nature of the data and the fact that the covariance matrices of different classes are unlikely to be equal (Reichert, Cho, & Wagner, 1983). In addition to the LDA approach, logistic regression is an alternative to conduct classification tasks. Basically, the logistic regression model was emerged as the technique in predicting dichotomous outcomes. Harrell and Lee (1985) found out that logistic regression is as efficient as LDA. However, it is also being criticized for some strong model assumptions like variation homogeneity and hence limited its application. Theoretically, both LDA and logistic regression are appropriate modeling tools when the relationship among variables is linear. In addition to LDA and logistic regression, artificial neural networks became an efficient alternative in modeling classification problems due to its capability to capture complex nonlinear relationships among variables. Even though neural networks have reported to have better classification capability than LDA and logistic regression (Desai, Crook, & Overstreet, 1996; Jensen, 1992; Lee, Chiu, Lu, & Chen, 2002; Piramuthu, 1999; West, 2000), it is, however, also being criticized for its long training process in designing the optimal network's topology and hard to identify the relative importance of potential input variables, and hence limited its applicability in handling classification problems (Chung & Gray, 1999; Craven & Shavlik, 1997; Lee et al., 2002).

In addition to the above-mentioned techniques, multivariate adaptive regression splines (MARS) is another commonly discussed data mining technique nowadays. MARS is widely accepted by data mining practitioners for the following facts. Firstly, unlike LDA and logistic regression, MARS exhibits the capability of modeling complex relationship among variables without strong model assumptions. Besides, unlike neural networks, MARS can identify 'important' independent variables through the built basis functions (more details will be discussed in Section 2) when many potential variables are considered. Thirdly, MARS does not need long training process and hence can save lots of modeling time when the data set is huge. Finally, one strong advantage of MARS over other classification techniques is the resulting model can be easily interpreted. It not only points out which variables are important in classifying objects/observations, but also indicates a particular object/observation belongs to a specific class when the built rules are satisfied. The final fact has important implications and can help professionals make appropriate decisions.

Aiming at improving the above-mentioned drawbacks of neural networks and increasing the classification accuracies of the existing approaches, the objective of the proposed study is to explore the performance of breast cancer diagnosis using a two-stage hybrid modeling procedure in integrating multivariate adaptive regression splines approach with neural networks technique. The rationale underlying the analyses is firstly to use MARS in modeling the breast cancer diagnostic problems. Then the obtained significant predictor variables are served as the input variables of the designed neural networks model. Please note that it is valuable to use MARS as a supporting tool for designing the topology of neural networks as we can learn more about the inner workings. Besides, as there is no theoretical method in determining the best input variables of a neural network model, MARS can be implemented as a generally accepted method for determining a good subset of input variables when many potential variables are considered in deciding the input vector of the designed neural network model. To demonstrate the feasibility and effectiveness that the inclusion of the obtained predictor variables from MARS would improve the classification accuracy of the neural network model, breast cancer diagnostic tasks are performed on one FNAC dataset. As to the structure of the designed neural network model, sensitivity analysis is firstly employed to solve the issue of finding the appropriate setup of the network's topology. Analytic results demonstrated that the proposed hybrid model provides a better initial solution and hence converges much faster than the conventional neural networks model. Besides, in comparison with the traditional neural network approach, the classification accuracy increases in terms of the proposed hybrid methodology. Moreover, the superior classification capability of the proposed technique can be observed by comparing the results with those using linear discrimintant analysis and solely using MARS approaches.

The rest of the paper is organized as follows. We will give a brief review and related literature of neural networks and multivariate adaptive regression splines in Section 2. The developments as well as the empirical results of breast cancer diagnostic models using linear discriminant analysis, MARS, neural networks, and the hybrid model in integrating MARS and neural networks approaches are presented in Section 3. Finally Section 4 addresses the conclusion and discusses the possible future research areas.

2. Research methodology and literature review

2.1. Artificial neural networks

Neural networks, originally derived from neurobiological models, are massively parallel, computer-intensive, and data-driven algorithmic systems composed of a multitude of highly interconnected nodes, known as neurons as well. Mimicking human neurobiological information-processing activities, each elementary node of a neural network is able to receive an input single from external sources or other nodes and the algorithmic procedure equipped in each node is sequentially activated to locally transforming the corresponding input single into an output single to other nodes or environment. From a holistic point of view on systemic collaboration, a neural network features a number of interconnected nodes serving as signal receivers and senders, the network architecture designed to describe connections between the nodes, and the training algorithm associated with finding values of network parameters (weights) for a particular network (Rumelhart, Hinton, & Williams, 1986). Relying on interactions of linked nodes, an output obtained from one node can serve as an input for others nodes and the conversion of inputs into outputs is activated by virtue of a certain transforming function that is typically monotone, but otherwise arbitrary. Meanwhile, the specified working function has to depend on parameters determined with a training set of inputs and outputs. And the network architecture is the organization of nodes and the types of connections permitted. The nodes are arranged in a series of layers with connections between nodes in different layers, but not between nodes in the same layer. Generally, nodes in the neural network can be divided into three layers: the input layer, the output layer, and one or more hidden layers. The layer receiving the inputs is called the input layer. The final layer provides the target output signal is the output layer. Any layers between the input and output layers are hidden layers. A simple representation of a neural network with one hidden layer can be shown in Fig. 1 (Rumelhart et al., 1986).

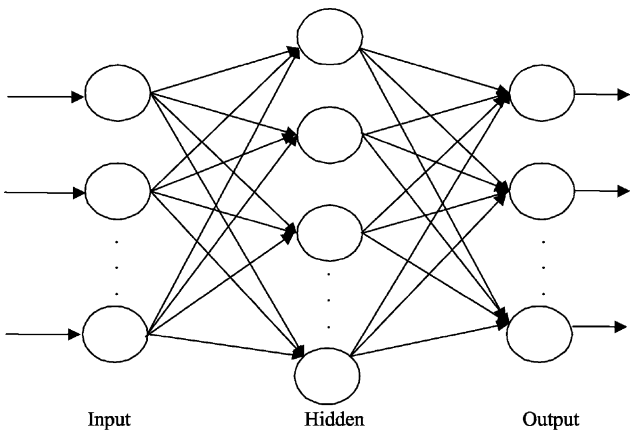

Fig. 1. A three-layer backpropagation neural networks.

Neural networks can be classified into two different categories, feedforward and feedback networks. The feedback networks contain nodes that can be connected to themselves, enabling a node to influence other nodes as well as itself. Kohonen self-organizing network and the Hopfield network are examples of this type of network. On the other hand, the nodes in feedforward networks can just take inputs only from the previous layer and send outputs to the next layer. The ADALINE and backpropagation neural networks (BPN) are two typical examples of this kind of network. BPN is a network essentially using a gradient steepest descent training algorithm and has been the most often utilized paradigm to date. For the gradient descent training algorithm, the step size, called the learning rate, must be specified first. The learning rate is crucial for BPN since smaller learning rates tend to slow down the learning process before convergence while larger ones may cause network oscillation and unable to converge.

Neural networks are increasingly found to be useful in modeling non-stationary processes due to its associated memory characteristics and outstanding generalization capability (Stern, 1996). More and more computer scientists and statisticians have interests in the computational potentials of neural network algorithms. Haykin (1994) wrote a comprehensive reference on artificial neural networks. Anderson and Rosenfeld (1988) edited a collection of papers that chronicled the major developments in neural network modeling. Cheng and Titterington (1994), Repley (1994), and Stern (1996) provided surveys describing the relevance of neural networks to the statistics community. As to the issue of determining the appropriate network topology: the number of layers, and the number of neurons in each layer, and the appropriate learning rates, please refer to Cybenko (1989), Davies (1994), Hecht-Nielsen (1990), Hornik et al. (1989), Kang (1991), Lippmann (1987), Tang and Fishwick (1993), and Wong (1991) for more details about the above issues.

Neural networks have been widely used in engineering, science, education, social science, medical research, business, forecasting and related fields (Cheng & Titterington, 1994; Chiu, Shao, & Lee, 2003; Lee & Chen, 2002; Lee & Chiu, 2002; Lee et al., 2002; Repley, 1994; Stern, 1996; Vellido, Lisboa, & Vaughan, 1999; Zhang, Patuwo, & Hu, 1998). The majority of the above references have reported that the classification accuracies of neural networks are better than those using discriminant analysis and logistic regression techniques.

2.2. Multivariate adaptive regression splines

MARS is first proposed by Friedman (1991) as a flexible procedure which models relationships that are nearly additive or involve interactions with fewer variables. The modeling procedure is inspired by the recursive partitioning technique governing classification and regression tree (CART, Breiman, Friedman, Olshen, & Stone, 1984) and

generalized additive modeling (Hastie & Tibshirani, 1990), resulting in a model that is continuous with continuous derivatives. It excels at finding optimal variable transformations and interactions, the complex data structure that often hides in high-dimensional data. And hence can effectively uncover important data patterns and relationships that are difficult, if not impossible, for other methods to reveal.

MARS essentially builds flexible models by fitting piecewise linear regressions; that is, the nonlinearity of a model is approximated through the use of separate regression slopes in distinct intervals of the predictor variable space. Therefore the slope of the regression line is allowed to change from one interval to the other as the two 'knot' points are crossed. The variables to use and the end points of the intervals for each variable are found via a fast but intensive search procedure. In addition to searching variables one by one, MARS also searches for interactions between variables, allowing any degree of interaction to be considered.

The general MARS function can be represented using the following equation (Friedman, 1991):

$$\hat{f}(x) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} [s_{km}(x_{v(k,m)} - t_{km})]_+$$

where a_0 and a_m are parameters, M is the number of basis functions, K_m is the number of knots, s_{km} takes on values of either 1 or -1 and indicates the right/left sense of the associated step function, $v(k,m)$ is the label of the independent variable, and t_{km} indicates the knot location.

The optimal MARS model is selected in a two-stage process. Firstly, MARS constructs a very large number of basis functions are selected to overfit the data initially, where variables are allowed to enter as continuous, categorical, or ordinal- the formal mechanism by which variable intervals are defined, and they can interact with each other or be restricted to enter in only as additive components. In the second stage, basis functions are deleted in order of least contribution using the generalized cross-validation (GCV) criterion (Craven & Wahba, 1979). A measure of variable importance can be assessed by observing the decrease in the calculated GCV values when a variable is removed from the model. The GCV can be expressed as follows:

$$\mathrm{LOF}(\hat{f}_M) = \mathrm{GCV}(M) = \frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{f}_M(x_i)]^2 \Big/ \left[1 - \frac{C(M)}{N}\right]^2$$

where there are N observations, and $C(M)$ is the cost-penalty measures of a model containing M basis function (therefore the numerator measures the lack of fit on the M basis function model $f_M(x_i)$ and the denominator denotes the penalty for model complexity $C(M)$).

Missing values can also be handled in MARS by using dummy variables indicating the presence of the missing values. By allowing for any arbitrary shape for the function as well as interactions, and by using the above-mentioned two-stage model building procedure, MARS is capable of reliably tracking the very complex data structures that often hide in high-dimensional data. Please refer to Friedman (1991) for more details regarding the complete model building process.

The interpretation of the resulting MARS model is achieved through individual plots of risk. For variables that enter into the model additively, a risk line plot showing each variable's individual contribution to the risk may be constructed. This is simply a plot of the risk (or log odds) represented by each basis function in the model that involves the variable of interest, for the range of values that the variable takes on in the data. Interactions can also be visualized as risk images showing the combined contribution of the variable's risk in the model. Points are only plotted for the data that are available. High levels of risk are indicated by dark grey areas on the plot and low levels of risk are represented by light grey regions. These types of plots are not only restricted to interactions but can also be used to visualize the contributions of variables that enter into the model additivity and are highly correlated with one another.

MARS has been widely used in handling problems in the areas of forecasting and classifications (De Gooijer, Ray, & Krager, 1998; Friedman & Roosen, 1995; Griffin, Fisher, Friedman, & Ryan, 1997; Kuhnert, Do, & McClure, 2000; Lewis & Stevens, 1991; Nguyen-Cong et al., 1996; Ohmann, Moustakis, Yang, & Lang, 1996). For other detailed list of the referred articles using MARS, the readers can login in to website http://www.salford-systems.com/MARSCITE.PDF provided by Salford Systems for more details.

3. Empirical study

In order to verify the feasibility and effectiveness of the proposed two-stage hybrid modeling procedure, one FNAC dataset provided by department of surgery, human oncology and computer sciences, University of Wisconsin at Madison is used in this study (Bennett & Mangasarian, 1992; Mangasarian, Setiono, & Wolberg, 1990). The data set consists of 569 patients' records. Among them, 212 are reported to have breast cancers while the remaining 357 are not. The diagnostic results of each patient consist of 30 predictor variables and are summarized in Table 1. And the response variable of the classification model is the diagnostic status of the patient-with or without breast cancers (the readers can refer to the website http://www.cs.wisc.edu/~olvi/uwmp/cancer.html for more details and descriptions about this data set). With the 569 patients used in this study, 398 patients (70% of the total patients) with respect to the ratio of having and not having breast cancers were randomly selected as the model building set (training sample) while the remaining 171 (30% of the total

Table 1
Predictor variables used in classifying breast cancer patterns

Mean radius	Standard error radius	Worst radius
Mean texture	Standard error texture	Worst texture
Mean perimeter	Standard error perimeter	Worst perimeter
Mean area	Standard error area	Worst area
Mean smoothness	Standard error smoothness	Worst smoothness
Mean compactness	Standard error compactness	Worst compactness
Mean concavity	Standard error concavity	Worst concavity
Mean concave points	Standard error concave points	Worst concave points
Mean symmetry	Standard error symmetry	Worst symmetry
Mean fractal dimension	Standard error fractal dimension	Worst fractal dimension

Table 2
Diagnostic results using discriminant analysis

Actual class	Classified class	
	1 (without breast cancer)	2 (with breast cancer)
1(without breast cancer)	100(93.46%)	7(6.54%)
2(with breast cancer)	0(0.00%)	64(100.00%)
	Average correct classification rate: 95.91%	

patients) will be retained as the validation set (testing sample).

The neural network simulator Qnet97, developed by Vesta Services Inc (1998), was utilized to develop the neural networks as well as the two-stage hybrid diagnostic models. It is a C based simulator that provides a system for developing backpropagation neural network configurations using the generalized delta learning algorithm. The discriminant analysis models will be implemented using the popular SPSS 1997 (1998) software. And MARS 2.0 (2001) provided by Salford Systems is used in building the MARS diagnostic models. All the modeling tasks are implemented on a PC with Intel Pentium II 750 MHz CPU processor. The detailed classification results using the above-mentioned four modeling techniques can be summarized as follows.

3.1. Discriminant analysis model

Among the variable selection procedures which can be used in this study, the stepwise discriminant analysis approach (Johnson & Wichern, 2002; Neter, Kutner, Nachtsheim, & Wasserman, 1996) is adopted in building the discriminant analysis diagnostic model.[1] Ten significant predictor variables are selected in the final discriminant function, namely worst concave points, worst radius, worst texture, worst area, standard error smoothness, mean perimeter, standard error radius, standard error compactness, worst cancavity and standard error cancavity. The diagnostic results using the obtained discriminant function are summarized in Table 2. From the results revealed in Table 2, we can observe that the average correct classification rate is 95.91% with 7 class 1 patients misclassified as class 2 patients (Here a class 1 patient is a patient whose status is without breast cancer while a class 2 patient is a patient with breast cancer. The latter discussion will be used the same terminology accordingly).

[1] If the covariance matrices of the given populations are not equal, the quadratic discriminant analysis (QDA) should be applied because the separation surface is a quadratic function. Despite the fact that LDA is a special case of QDA with stronger assumptions which should restrict its applications, in fact LDA has reported to be a more robust method when the theoretical presumptions are violated (Sanchez & Sarabia, 1995; Sharma, 1996). And hence the LDA approach will be used in building the diagnostic model in this paper.

3.2. Multivariate adaptive regression splines model

The variable selection results using MARS diagnostic model can be summarized in Table 3. It is observed that worst area, mean radius, mean texture, mean concave points, worst concave points, worst symmetry, standard error concavity and standard error compactness do play crucial roles in deciding the MARS diagnostic models. The diagnostic results using the obtained MARS model can be summarized in Table 4. From the results in Table 4, we can observe that the average correct classification rate is 97.66% with 1 (3) class 1 (2) customers misclassified as class 2 (1) customers.

3.3. Neural networks model

Since Vellido et al. (1999) pointed out that more than 75% of applications using neural networks will use the BPN training algorithms, this study will also use the popular BPN

Table 3
Basis functions and important predictor variables using MARS

Fun	Std. dev.	GCV	NO.BF	Variable	Relative importance (%)
1	0.504	0.055	2	Worst area	100.000
2	0.334	0.048	1	Mean radius	70.064
3	0.076	0.045	1	Mean texture	54.408
4	0.182	0.045	1	Mean concave points	54.152
5	0.110	0.044	2	Worst concave points	49.757
6	0.040	0.042	1	Worst symmetry	27.571
7	0.042	0.041	1	Standard error concavity	23.841
8	0.037	0.041	1	Standard error compactness	15.533

Table 4
Diagnostic results using MARS

Actual class	Classified class	
	1(without breast cancer)	2(with breast cancer)
1(without breast cancer)	106(99.07%)	1(0.93%)
2(with breast cancer)	3(4.69%)	61(95.31%)
	Average correct classification rate: 97.66%	

Table 5
BPN model prediction results

Number of hidden nodes	Learning rates	Training RMSE	Testing RMSE
58	0.006	0.127845	0.137439
	0.008	0.127363	0.137026
	0.010	0.126763	0.136250
59	0.006	0.127793	0.137362
	0.008	0.127285	0.136796
	0.010	0.126879	0.136637
60	0.006	0.127777	0.137408
	0.008	0.127097	0.136581
	0.010	0.126278	0.135929
61	0.006	0.127900	0.137158
	0.008	0.127264	0.136811
	0.010	0.126781	0.136669
62	0.006	0.128131	0.137849
	0.008	0.127501	0.136933
	0.010	0.128987	0.139142

in building the neural network diagnostic model. As recommended by Cybenko (1989) and Hornik et al. (1989) that one-hidden-layer network is sufficient to model any complex system with any desired accuracy, the designed network model will have only one hidden layer. And since there are 30 input nodes in the input layer (refer to Table 1 for details), the initial number of hidden nodes to be tested was chosen to be 58, 59, 60, 61, and 62 (other possible number of hidden nodes have also been tested and no better results can be obtained). And the network has only one output node, status of the patient-with or without breast cancers. As Rumelhart et al. (1986) concluded that lower learning rates tended to give better network results and the networks were unable to converge when the learning rate is greater then 0.010, learning rates 0.006, 0.008, and 0.010 are tested during the training process. The convergence criteria used for training are a root mean squared error (RMSE) less than or equal to 0.0001 or a maximum of 3000 iterations. The network topology with the minimum testing RMSE is considered as the optimal network topology.

The prediction results of the BPN networks with combinations of different hidden nodes and learning rates are summarized in Table 5. From Table 5, the {30-60-1} topology with a learning rate of 0.010 gives the best result (minimum testing RMSE). Here {ni-nh-no} stands for the number of neurons in the input layer, number of neurons in the hidden layer and number of neurons in the output layer, respectively. To examine the convergence characteristics of the proposed neural networks model, the RMSE during the training process for the {30-60-1} network with the learning rate of 0.010 are depicted in Fig. 2. The excellent convergence characteristics of the constructed {30-60-1} networks can easily be observed.

The diagnostic results using the designed BPN model can be summarized in Table 6. From the results in Table 6, we can observe that the average correct classification rate is 98.25% with only three class 1 patients misclassified as class 2 patients. By comparing the results of Tables 2–6, it can be observed that BPN has the highest average correct classification rate in comparison with discriminant analysis and MARS approaches.

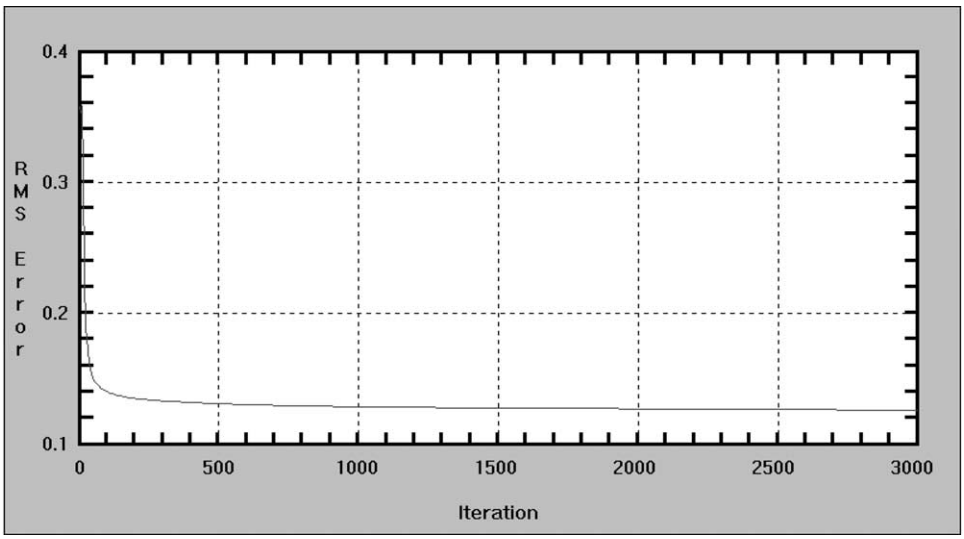

Fig. 2. The RMSE history of the {30-60-1} network during the training process.

Table 6
Diagnostic results using BPN

Actual class	Classified class	
	1(without breast cancer)	2(with breast cancer)
1(without breast cancer)	104(97.20%)	3(2.80%)
2(with breast cancer)	0(0.00%)	64(100.00%)
	Average correct classification rate: 98.25%	

3.4. Hybrid model

The single-layer BPN model will again be used in building the hybrid diagnostic model by integrating MARS and BPN. The input layer of the hybrid model contains eight input nodes (refer to Table 3 for more details), as the hybrid model will use the significant predictor variables of the obtained MARS diagnostic model as the input nodes. As there are eight input nodes in the input layer, the initial number of hidden nodes to be tested was set to be 12, 13, 14, 15, 16, 17, 18, 19 and 20 (again other possible number of hidden nodes have also been tested and no better results can be reported). And the network has only one output node, the status of the patient—with or without breast cancers. As the networks were unable to converge when the learning rate is greater then 0.005, learning rates 0.001, 0.003, and 0.005 are tested during the training process. The convergence criteria used for training are a root mean squared error (RMSE) less than or equal to 0.0001 or a maximum of 4000 iterations. Again the network topology with the minimum testing RMSE is considered as the optimal network topology.

The prediction results of the hybrid model are summarized in Table 7. From Table 7, the {8-13-1} topology with a learning rate of 0.005 gives the best result. The RMSE during the training process for the {8-13-1} network are depicted in Fig. 3. Again the excellence convergence characteristics of RMSE of the proposed hybrid model can easily be observed.

The diagnostic results using the hybrid model are summarized in Table 8. Table 8 reveals that the average correct classification rate is 98.25% with only three class 1 patients misclassified as class 2 patients. It can also be observed that both BPN and the hybrid diagnostic models have the same classification accuracies. However, we believe that the hybrid model should be a better alternative since it will identify important predictor variables which may provide valuable information for further diagnostic purposes.

Finally, in order to evaluate the classification capabilities of the above four constructed diagnostic models, the summarized results can be shown in Table 9. From the results revealed in Table 9, we can conclude that both BPN and the hybrid model have the best diagnostic capability in terms of the average classification rate in comparison with those using discriminant analysis and MARS models.

Table 7
Integrated hybrid model prediction results

Number of hidden nodes	Learning rates	Training RMSE	Testing RMSE
12	0.001	0.141755	0.158061
	0.003	0.140051	0.154882
	0.005	0.137766	0.152775
13	0.001	0.142014	0.157117
	0.003	0.140196	0.154977
	0.005	0.136845	0.151705
14	0.001	0.142645	0.158239
	0.003	0.140092	0.154937
	0.005	0.138805	0.153336
15	0.001	0.142114	0.157261
	0.003	0.140168	0.155278
	0.005	0.139143	0.153285
16	0.001	0.142610	0.157923
	0.003	0.139616	0.154429
	0.005	0.138192	0.152531
17	0.001	0.141526	0.156722
	0.003	0.139952	0.154866
	0.005	0.138979	0.152830
18	0.001	0.142437	0.157684
	0.003	0.140516	0.155227
	0.005	0.139081	0.153253
19	0.001	0.142394	0.157615
	0.003	0.140613	0.155258
	0.005	0.139028	0.143059
20	0.001	0.142353	0.157918
	0.003	0.139502	0.154227
	0.005	0.139086	0.152993

3.5. Type I, type II errors and the CPU times of the constructed models

It is well known that, in order to evaluate the overall classification capability of the designed diagnostic models, the misclassification costs also have to be taken into account (Johnson & Wichern, 2002; West, 2000). It is apparent that the costs associated with Type I error (a patient without breast cancer is misclassified as a patient with breast cancer) and Type II error (a patient with breast cancer is misclassified as a patient without breast cancer) are significantly different. In general, the misclassification costs associated with Type II errors are much higher than those associated with Type I errors. Hence, special attention should pay to Type II errors in order to evaluate the overall diagnostic capability. Table 10 summarizes the Type I and Type II errors of the four models being discussed. As the results revealed in Table 10, both BPN and the hybrid model have the lowest Type II error in comparison with the other two approaches. Therefore we can conclude that both BPN and the hybrid model not only have better average classification rate, but also has lower Type II errors and hence can successfully reduce the possible risks

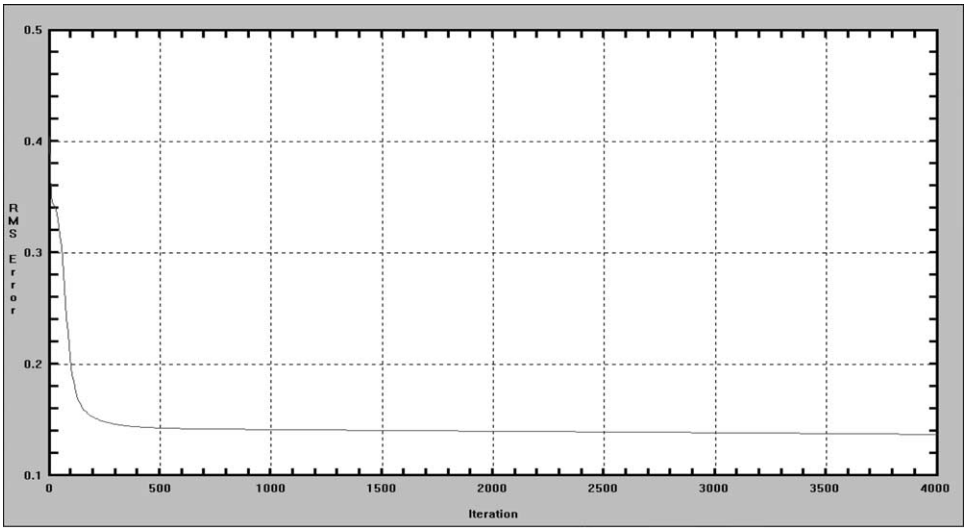

Fig. 3. The RMSE history of the {8-13-1} hybrid model during the training process.

due to the high misclassification costs associated with Type II errors.

Finally, the running time in implementing the classification tasks on the computer is another factor needs to be evaluated. Table 11 compares the CPU time in implementing the BPN and hybrid diagnostic models (for one combination of hidden node and learning rate). From the results revealed in Table 11, we can see that the CPU time for the hybrid model is only about one third of that when using BPN. Therefore the hybrid model should be a better alternative than BPN since it can save lots of modeling time with the same classification accuracy.

Table 8
Diagnostic results using the hybrid model

Actual class	Classified class	
	1(without breast cancer)	2(with breast cancer)
1(without breast cancer)	104(97.20%)	3(2.80%)
2(with breast cancer)	0(0.00%)	64(100.00%)
	Average correct classification rate: 98.25%	

Table 9
Classification results of the four constructed models

Diagnostic models	Classification results		
	{1–1} (%)	{2–2} (%)	Average correct classification rate (%)
Discriminant analysis	93.46	100.00	95.91
MARS	99.07	95.31	97.66
BPN	97.20	100.00	98.25
Hybrid model	97.20	100.00	98.25

4. Conclusions and areas of future research

Breast cancer is a very common and serious cancer for women through out the world. The commonly used diagnostic techniques, like mammography and FNAC, are reported to lack of high diagnostic capability. Therefore, there is an absolute necessity in developing better diagnostic techniques. Basically, the objective of these identification techniques is to assign patients to either a 'benign' group that does not have breast cancer or a 'malignant' group who has strong evidence of having breast cancer. And hence the breast cancer diagnostic problems are in the scope of the more general and widely discussed classification problems. Discriminant analysis is the most commonly used statistical classification technique, but often being criticized due to its strong model assumptions and lack of classification accuracy. On the other hand, the artificial neural networks has become a very popular alternative in the classification

Table 10
Type I and Type II errors of the four models

Diagnostic models	Performance assessment	
	Type I error (%)	Type II error (%)
Discriminant analysis	6.54	0.00
MARS	0.93	4.69
BPN	2.80	0.00
Hybrid model	2.80	0.00

Table 11
CPU time for BPN and hybrid models (one combination of parameters)

Diagnostic models	Number of input nodes	CPU time (s)
BPN	30	360
Hybrid model	8	120

tasks due to its associated memory characteristic and outstanding generalization capability. However, it is also being criticized for its long training process in designing the optimal network's topology and hard to identify the relative importance of potential input variables.

The purpose of this research is to propose a hybrid breast cancer diagnostic model by integrating artificial neural networks and multivariate adaptive regression splines (MRAS). The rationale underlying the analyses is firstly to use MARS in modeling the breast cancer diagnostic problems. Then the obtained significant predictor variables are served as the input nodes of the designed neural networks model. To demonstrate the feasibility and effectiveness that the inclusion of the obtained predictor variables from MARS would improve the classification accuracy of the neural networks model, breast cancer diagnostic tasks are performed on one FNAC dataset. Analytic results demonstrated that, both BPN and the proposed hybrid model have better classification accuracy and lower Type II errors associated with high misclassification costs, in comparison with discriminant analysis and MARS approaches. However, the hybrid model should be a better alternative since it exhibits the capability in identifying important predictor variables which may provide valuable information for further diagnostic purposes. Besides, the hybrid model can save lots of implementation time on the computer and therefore shorten the time for on time decisions.

Future researches may aim at collecting more important variables that will increase the classification accuracies. Using other data mining techniques, like CART, in evaluating their diagnostic capabilities is also recommended. Integrating other artificial intelligence techniques, like fuzzy discriminant analysis, genetic algorithms and grey theory, with neural networks in further refining the network structure and improving the classification accuracies may also being discussed.

References

Anderson, T. W. (1984). An introduction to multivariate statistical analysis. New York: Wiley.

Anderson, J. A., & Rosenfeld, E. (1988). Neurocomputing: foundations of research. Cambridge, MA: MIT Press.

Bennett, K. P., & Mangasarian, O. L. (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, *1*, 23–34.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Pacific Grove, CA: Wadsworth.

Cabena, P., Hadjinaian, P. O., Stadler, J., Verhees, J., & Zanasi, A. (1998). Discovering data mining from concept to implementation. Upper Saddle River, NJ: Prentice-Hall.

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, *8*(6), 866–883.

Cheng, B., & Titterington, D. M. (1994). Neural Network: a review from a statistical perspective (with discussion). *Statistical Science*, *9*, 2–54.

Chiu, C. C., Shao, Y. J., & Lee, T. S. (2003). Identification of process disturbance using SPC/EPC and neural networks. *Journal of Intelligent Manufacturing*, *14*(3), 379–388.

Chung, H. M., Gray, P., & Guest Editors, (1999). Special section: data mining. *Journal of Management Information Systems*, *16*, 11–16.

Craven, M. W., & Shavlik, J. W. (1997). Using neural networks for data mining. *Future Generation Computer Systems*, *13*, 221–229.

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numberische Mathematik*, *31*, 317–403.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal Function. *Mathematical Control Signal Systems*, *2*, 303–314.

Curt, H. (1995). The devil's in the detail: techniques: Tools, and applications for database mining and knowledge discovery-Part. *Intelligent Software Strategies*, 1–15.

Davies, P. C. (1994). Design issues in neural network development. *Neurovest*, *5*, 21–25.

De Gooijer, J. G., Ray, B. K., & Krager, H. (1998). Forecasting exchange rates using TSMARS1. *Journal of International Money and Finance*, *17*(3), 513–534.

Desai, V. S., Crook, J. N., & Overstreet, G. A., Jr. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, *95*, 24–37.

Dillon, W. R., & Goldstein, M. (1984). Multivariate analysis methods and applications. New York: Wiley.

Elmore, J., Wells, M., Carol, M., Lee, H., Howard, D., & Feinstein, A. (1994). Variability in radiologists interpretation of mamograms. *New England Journal of Medicine*, *331*(22), 1493–1499.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*, 27–34.

Fentiman, I. S. (1998). Detection and treatment of breast cancer (2nd ed.). London: Martin Dunitz.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, *19*, 1–141.

Friedman, J. H., & Roosen, C. B. (1995). An introduction to multivariate adaptive regression splines. *Statistical Methods in Medical Research*, *4*, 197–217.

Griffin, W. L., Fisher, N. I., Friedman, J. H., & Ryan, C. G. (1997). Statistical techniques for the classification of chromites in diamond exploration samples. *Journal of Geochemical Exploration*, *59*, 233–249.

Hand, D. J. (1981). Discrimination and classification. New York: Wiley.

Harrell, F. E., & Lee, K. L. (1985). A comparison of the discrimination of discriminant analysis and logistic regression. In P. K. Se (Ed.), *Biostatistics: statistics in biomedical, public health, and environmental sciences*. Amsterdam: North-Holland.

Hastie, T., & Tibshirani, R. (1990). Generalized additive models. London: Chapman & Hall.

Haykin, S. S. (1994). Neural networks: a comprehensive foundation. New York: Macmillan.

Hecht-Nielsen, R. (1990). Neurocomputing. Menlo Park, CA: Addison-Wesley.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximations. *Neural Networks*, *2*, 336–359.

Jensen, H. L. (1992). Using neural networks for credit scoring. *Managerial Finance*, *18*, 15–26.

Johnson, R. A., & Wichern, D. W. (2002). Applied multivariate statistical analysis (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Kang, S (1991). *An investigation of the use of feedforward neural networks for forecasting*, PhD thesis, Kent State University.

Kovalerchuck, B., Triantaphyllou, E., Ruiz, J. F., & Clayton, J. (1997). Fuzzy logic in computer-aided breast-cancer diagnosis: analysis of lobulation. *Artificial Intelligence in Medicine*, *11*, 75–85.

Kuhnert, P. M., Do, K.-A., & McClure, R. (2000). Combining non-parametric models with logistic regression: an application to motor

vehicle injury data. *Computational Statistics and Data Analysis*, *34*, 371–386.

Lee, G., Sung, T. K., & Chang, N. (1999). Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction. *Journal of Management Information Systems*, *16*, 63–85.

Lee, T. S., & Chen, N. J. (2002). Investigating the information content of non-cash-trading index futures using neural networks. *Expert Systems with Applications*, *22*(3), 225–234.

Lee, T. S., & Chiu, C. C. (2002). Neural network forecasting of an opening cash price index. *International Journal of Systems Science*, *33*(3), 229–237.

Lee, T. S., Chiu, C. C., Lu, C. J., & Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, *23*(3), 245–254.

Lewis, P. A. W., & Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). *Journal of American Statistical Association*, *86*, 864–877.

Lippmann, R. P. (1987). An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4–22.

Mangasarian, L., & Wolberg, W. H. (1990). Pattern recognition via linear programming: theory and application to medical diagnosis. In T. F. Coleman, & Y. Li (Eds.), *Large-scale numerical optimization* (pp. 2–30). Philadelphia: SIAM.

MARS 2.0—for windows 95/98/NT, Salford Systems, San Diego, CA, 2001.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied linear statistical models. Chicago, IL: Irwin.

Ngan, P. S., Wong, M. L., Lam, W., Leung, K. S., & Cheng, J. C. Y. (1999). Medical data mining using evolutionary computation. *Artificial Intelligence in Medicine*, *16*, 73–96.

Nguyen-Cong, V., Van D, G., & Rode, B. M. (1996). Using multivariate adaptive regression splines to QSAR studies of dihydroartemisinin derivatives. *European Journal of Medical Chemistry*, *31*, 797–803.

Ohmann, C., Moustakis, V., Yang, Q., & Lang, K. (1996). Evaluation of automatic knowledge acquisition techniques in the diagnosis of acute abdominal pain. *Artificial Intelligence in Medicine*, *13*, 23–36.

Pendharkar, P. C., Rodger, J. A., Yaverbaum, G. J., Herman, N., & Benner, M. (1999). Associations statistical, mathematical and neural approaches for mining breast cancer patterns. *Expert Systems with Applications*, *17*, 223–232.

Piramuthu, S. (1999). Financial credit-risk evaluation with neural and neurofuzzy systems. *European Journal of Operational Research*, *112*, 310–321.

Qnet 97 (1998). *Neural Network Modeling for Windows 95/98/NT*, Vesta Services, Winnetka, IL.

Reichert, A. K., Cho, C. C., & Wagner, G. M. (1983). An examination of the conceptual issues involved in developing credit-scoring models. *Journal of Business and Economic Statistics*, *1*, 101–114.

Repley, B. (1994). Neural networks and related methods for classification (with discussion). *Journal of the Royal Statistical Society, Series B*, *56*, 409–456.

Rumelhart, D. E., Hinton, D. E., & Williams, R. J. (1986) (*Vol. 1*). *Learning internal representations by error propagation in parallel distributed processing*, Cambridge, MA: MIT Press, pp. 318–362.

Sanchez, M. S., & Sarabia, L. A. (1995). Efficiency of multi-layered feedforward neural networks on classification in relation to linear discriminant analysis. Quadratic discriminant analysis and regularized discriminant analysis. *Chemometrics and Intelligent Laboratory Systems*, *28*, 287–303.

Sharma, S. (1996). Applied multivariate techniques. New York: Wiley.

SPSS 1997 (1998). *Statistic Modeling for Windows 95/98/NT*, SPSS Inc.

Stern, H. S. (1996). Neural networks in applied statistics. *Technometrics*, *38*(3), 205–216.

Tang, Z., & Fishwick, P. A. (1993). Feedforward neural nets as models for time series forecasting. *ORSA Journal on Computing*, *5*, 374–385.

Vellido, A., Lisboa, P. J. G., & Vaughan, J. (1999). Neural networks in business: a survey of applications (1992–1998). *Expert Systems with Applications*, *17*, 51–70.

West, D. (2000). Neural network credit scoring models. *Computers and Operations Research*, *27*, 1131–1152.

Wingo, P. A., Tong, T., & Bolden, S. (1995). Cancer statistics. *Ca—A Cancer Journal for Clinicians*, *45*(1), 8–30.

Wong, F. S. (1991). Time series forecasting using backpropagation neural networks. *Neurocomputing*, *2*, 147–159.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: the state of the art. *International Journal of Forecasting*, *14*, 35–62.