Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters xxx (2004) xxx–xxx
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Neural vs. statistical classifier in conjunction with genetic algorithm based feature selection

Ping Zhang [a,*], Brijesh Verma [b], Kuldeep Kumar [a]

[a] School of Information Technology, Bond University, Gold Coast 4229, Australia
[b] School of Information Technology, Central Queensland University, Rockhampton 4702, Australia

Received 10 September 2004

Abstract

Digital mammography is one of the most suitable methods for early detection of breast cancer. It uses digital mammograms to find suspicious areas containing benign and malignant microcalcifications. However, it is very difficult to distinguish benign and malignant microcalcifications. This is reflected in the high percentage of unnecessary biopsies that are performed and many deaths caused by late detection or misdiagnosis. A computer based feature selection and classification system can provide a second opinion to the radiologists in assessment of microcalcifications. The research in this paper proposes and investigates a neural-genetic algorithm for feature selection in conjunction with neural and statistical classifiers to classify microcalcification patterns in digital mammograms. The obtained results show that the proposed approach is able to find an appropriate feature subset and neural classifier achieves better results than two statistical models.

Keywords: Microcalcifications pattern classification; Neural networks; Statistical methods; Feature selection; Genetic algorithm

1. Introduction

Breast cancer is a leading cause of cancer death in women between the ages of 40 and 55 [1]. Currently, there is no certain way to prevent breast cancer (Breast Cancer Facts, 2002). This is one reason of why early detection represents a very important factor in its treatment and consequently the survival rate.

Digital mammography is considered to be the most reliable method of early detection, however, in the early stage, the visual clues are subtle and varied in appearance, making diagnosis difficult, challenging even for specialists. In mammography breast abnormalities are divided into exhibiting

* Corresponding author. Tel.: +61755953378; fax: +61755953320.
E-mail addresses: pzhang@staff.bond.edu.au (P. Zhang), b.verma@cqu.edu.au (B. Verma), kkumar@staff.bond.edu.au (K. Kumar).

[1] Y-me Nation breast http://www.y-mechicagoland.org/breast_info.htm

0167-8655/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2004.09.053

microcalcification, circumscribed lesions and spiculated lesions. Microcalcification appears as a small bright spot on the mammogram. Most of the minimal breast cancers are detected by the presence of microcalcifications (Chitre et al., 1993). It is however difficult to distinguish between benign and malignant microcalcifications. To decide whether a suspicious area on a digital mammogram contains benign/malignant microcalcifications, traditionally the tissue has to be removed for examination using breast biopsy techniques. The computer classification system of the microcalcifications can provide a second opinion to the radiologists and reduce the number of unnecessary biopsies. A digital mammogram brought the possibility of using computer-aided diagnosis system.

Current image processing techniques make microcalcification detection easier, however classification of malignant and benign microcalcifications is still very challenging and a difficult problem for researchers. One important factor directly affects the classification result is feature extraction. Researchers spend a lot of time in attempt to find a group of features that will aid them in improving the classification for malignant microcalcifications from benign. In the literature, region-based features (Chitre et al., 1993; Zheng et al., 1994), shape-based features (Shen et al., 1994; Jiang et al., 1996; Shen et al., 1994), image structure features (Chitre et al., 1993; Zokos, 1998; Verma, 1998, 1999; Kevin et al., 1993; Chris and Tina, 1997), texture based features (Maria-luiza et al., 2001; Marcoz and Torres-Torriti, 2001), and position related features (Maria-luiza et al., 2001) are described and used for experiments.

One feature taken alone might not be significant for the classification but might be very significant if combined with other features. The whole set of the features may include the redundant or irrelevant information. Ho (1998), combined and constructed multiple classifiers using randomly selected features which can achieve better performance in classification than using the complete set of features. The only way to guarantee the selection of an optimal feature vector is an exhaustive search of all possible subset of features. However, search spaces to be explored could be very large. For N features, the number of possible subsets is 2^N. Feature subset selection is defined as a process of selecting a subset of features out of the larger set of features, which maximize the classification performance of a given procedure over all possible subsets. The objective of this paper is to propose and investigate a neural-genetic algorithm in conjunction with neural and statistical classifiers to find the most significant features or the sets of features suitable for classifying abnormalities of microcalcifications.

The remainder of this paper is organized as follows: Section 2 reviews the work has been done in this area. Section 3 describes the proposed research methodology. The experimental results are presented in Section 4. Section 5 discusses the obtained results by the proposed technique. The conclusion and future directions are stated in the final section.

2. Literature review

Researchers put lots of effort to find best feature or best combination of features (i.e. feature vector) that gives highest classification rate using appropriate classifier. Search strategies such as *Hill-climbing* and *Best-first search* have been used by Kohavi and Somerfield (1995) to find subsets of features with high predictive accuracy. Cost and Salzberg (1996) used *feature weighting* technique assigning a real-valued weight to catch feature. The weight associated with a feature, measures its relevance or significance in the classification task. John et al. (1994) examined the use of *heuristic* search for feature subset selection. Most of these techniques assume monotonicity of some measure of classification performance and then use branch and bound search. This monotonicity assumption in some form appears to work reasonably well with linear classifiers. However, they can exhibit poor performance with nonlinear classifiers such as neural networks (Liu and Setiono, 1966).

Racz and Nieniewski (2000), employed most discriminative components analysis and a forward/backward selection strategy to reduce the input size from 189 to 46 for his computer aided diagnosis system based on analysis of microcalcifi-

cations. Some others (Ho, 1998; Guerra-Salcedo and Whitley, 1999; Guerra-Salcedo et al., 1999) have explored *randomized* and population based heuristic search techniques such as genetic algorithms to select feature subsets for use with different classifiers. Genetic algorithms (GAs) offer a particularly attractive approach to multicriteria optimization, which cannot be handled by most of the other methods. In (Marcoz and Torres-Torriti, 2001) genetic algorithm was used for feature selection for texture classifier on synthetic aperture radar airborne imagery. They found a few more effective features than the others of image classification. Guerra-Salcedo and Whitley (1999) and Guerra-Salcedo et al. (1999) involved genetic strategies for feature selection combined CF/RSC (Common Features/Random Sample Climbing) and Decision Tables dealing with large feature spaces showing a good result.

Overall reviewing the literature, neural networks are particularly effective for fine-tuning solutions once promising regions in the search space have been identified. It is currently used for classification by many researchers. Chitre et al. (1993) used a back propagation (BP) neural network for image structure microcalcification classification and compared results with statistical classifiers. Though result is not promising, it is better than the statistical classifiers. Qian and Clarke (1994) and Qian (2001) used the BP algorithm and wavelet transform-based methods with Kalman filtering neural network for mass detection. Verma (1999) employed BP with momentum and DSM (Direct Solution Method) based training algorithms to train a feed forward neural network for classification of microcalcification. He achieved the classification rate of 81.25% for benign and malignant. Verma and Zakos (2001) developed a computer-aided diagnosis system for digital mammograms based on fuzzy-neural and feature extraction techniques. They used a fuzzy technique to detect microcalcification patterns and a neural network to classify it. The microcalcification areas from the Nijmegen digital mammographic database were used for their experimentation. Their research achieved a very commendable result with the classification rate 88.9% for classifying the microclcification as benign or malignant.

Evolutionary algorithms are generally quite effective for rapid global search of large search spaces in multi-modal optimization problems. The use of GAs for training neural network (e.g. in (Metin et al., 2002)) has recently begun to receive a considerable amount of attention. Compare with the gradient-based training like BP, GA is not based on the calculation of the derivative of the error surface, which can be unavailable or sometimes very costly to find. The goal towards which GA training proceeds is determined by the fitness function the user defines. This makes it easier for the ANN for decreasing the overall complexity and generalization for the fitness of its population. Although slow, it is less sensitive to the initial condition.

As far as we know, there has been no technique of NN and GA combination used for extracting the best features from digital mammograms to classify microcalcifications, and make the computer cancer detection more realistic. An overview of the proposed methodology is described below.

3. Research methodology

3.1. Mammographic database

In this research Digital Database for Screening Mammography (DDSM) from University of South Florida (Heath et al., 1998) is used for experiments. It was downloaded from marathon.csee.usf.edu/Mammography/DDSM.

The establishment of DDSM makes the possibility of comparing results from different research groups (Heath et al., 1998). It provides a large set of mammograms in a digital format. In DDSM, the outlines for the suspicious regions are derived from markings made on the film by at least two experienced radiologists. Each boundary for the abnormality is specified as a chain code, which allows easy feature extraction for each of the suspicious areas in the image files.

3.2. Feature extraction

To find the best feature or combination of features and get the high classification rate for micro-

calcification classification is one of the main aims of the proposed research. The feature extraction technique consists of three parts: (1) area extraction from the marked mammograms; (2) feature extraction from the extracted areas (3) feature selection for the classification.

3.2.1. Area extraction

Area extraction deals with extracting the grey values from all the suspicious areas in the mammograms marked by the expert radiologists. It accomplished by three steps: (1) According to the chain codes described in the ".OVERLAY" files of the database, extract the boundary of the suspicious areas. (2) Resize the boundary. (3) Extract all the grey values in the area and in the boundary area.

3.2.2. Feature extraction from extracted areas

A set of 14 features is calculated for each suspicious area in this research. The 10 features are commonly used existing features in the literatures and 4 are modified features by us (Verma and Zakos) in our previous research, which achieved higher classification rates than its traditional counter part.

All 14 features are: (1) number of pixels, (2) average histogram (AHg), (3) average grey level (AG), (4) modified energy (MEgy), (5) modified entropy (Metp), (6) modified standard deviation (MSD), (7) modified skew (MSk), (8) average boundary grey level (BAG), (9) difference (Df), (10) contrast (Ctr), (11) energy (Egy), (12) entropy (Etp), (13) standard deviation (SD), (14) skew (Sk).

The formulae for every feature are described below: For each of the formulae: T is the total number of pixels, g is an index value of image I, K is the total number of grey levels (i.e. 4096), j is the grey level value (i.e. 0–4095), $I(g)$ is the grey level value of pixel g in image I, $N(j)$ is the number of pixels with grey level j in image I, $P(I(g))$ is the probability of grey level value $I(g)$ occurring in image I, $P(g) = N(I(g))/T$, $P(j)$ is the probability of grey level value j occurring in image I, $P(j) = N(j)/T$.

Number of pixels is the count of the pixels in the extracted area.

$$\mathrm{AG} = \frac{1}{T}\sum_{g=0}^{T-1} I(g) \tag{1}$$

$$\mathrm{AHg} = \frac{1}{k}\sum_{j=0}^{T-1} N(j) \tag{2}$$

$$\mathrm{Egy} = \sum_{g=0}^{T-1} [P(I(g))]^2 \tag{3}$$

$$\mathrm{MEgy} = \sum_{g=0}^{T-1} [P(1(g))]^2 \tag{4}$$

$$\mathrm{Etp} = -\sum_{j=0}^{k-1} P(j)\log_2[P(j)] \tag{5}$$

$$\mathrm{MEtp} = \sum_{g=0}^{T-1} P(g)\log_2[P(I(g))] \tag{6}$$

$$\mathrm{SD}(\sigma) = \sqrt{\sum_{g=0}^{T-1} (j - \mathrm{AG})^2 P(j)} \tag{7}$$

$$\mathrm{MSD}(\sigma_m) = \sqrt{\sum_{g=0}^{T-1} (I(g) - \mathrm{AG})^2 P(I(g))} \tag{8}$$

$$\mathrm{Sk} = \frac{1}{\sigma j^3}\sum_{j=0}^{k-1} (j - \mathrm{AG})^3 P(j) \tag{9}$$

$$\mathrm{MSk} = \frac{1}{\sigma^3}\sum_{g=0}^{T-1} (I(g) - \mathrm{AG})^3 P(I(g)) \tag{10}$$

$$\mathrm{Dffe} = \mathrm{AG} - \mathrm{BAG} \tag{11}$$

$$\mathrm{Ctr} = \frac{\mathrm{Dff}}{\mathrm{AG} + \mathrm{BAG}} \tag{12}$$

3.2.3. Feature selection algorithm

In this research, a neural-genetic algorithm is developed for feature selection based on the neural network pattern classifiers. Each individual in the population represents a candidate solution to the feature subset selection problem. Here, there are 2^{14} possible feature subsets.

In this step, a binary vector of dimension 14 represents the individual in the population. In other words, the chromosome defined contains 14 genes, one gene for each feature, which can take 2 values. A value of 0 indicates that the corresponding feature is not selected, and a value 1 means that the feature is selected. An initial population of chromosomes is randomly generated. 1-point binary crossover and binary mutation are performed. The roulette wheel selection strategy is also used in the algorithm for feature selection. The relevant parameter settings are: Population size: 30; Number of generation: 200; Probability of crossover: 0.8; Probability of mutation: 0.2.

The fitness of the chromosome is calculated according to the classification rate of the evolved subset of features, as it is shown in Fig. 1.

3.3. Classification

3.3.1. Neural classifier

The selected features are the inputs of the Neural Networks, which are used for classification. The number of the inputs are decided by the automatically selection of GA processing. The values of the inputs are the normalized features that are between 0 and 1. One hidden layer is used in the NN. The nodes of hidden layer were adjusted in an attempt to achieve optimum classification rates. One output of NN is used in the proposed research. The value is also set to be between 0 and 1. The desired output was specified as 0 for benign and 1 for malignant. An output value of an actual NN less than threshold [2] is classified to be benign. That means the relevant input features belong to a benign microcalcification. An output value of more than threshold means that the neural net has classified the input features as belonging to a malignant microcalcification.

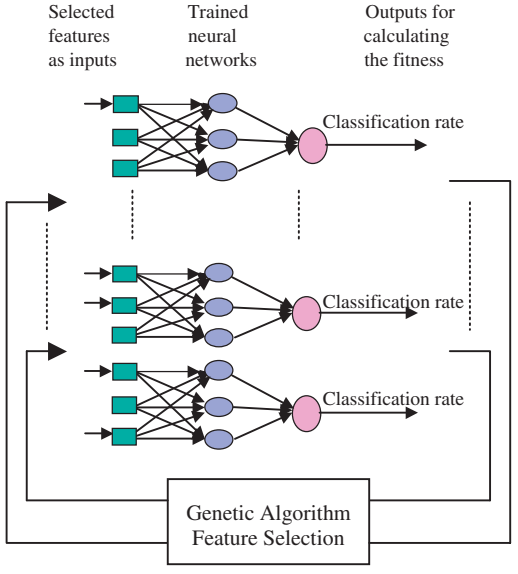

Fig. 1. Feature selection based on NN classification.

The NNs for classification with different selected inputs are trained separately by another genetic algorithm. In the genetic algorithm for feature selection involves many generations. In each generation, evaluation of an individual (a feature subset) involves training neural networks.

A standard genetic algorithm with a roulette wheel selection strategy is used in this research. In the process of NN training, the genes of every individual in the population represent the weights between input and hidden layer and the weights between hidden layer and output of the NN. The results are based on random initialization to the weights of every individual in the population with the following parameters: Population size: 40; Number of generation: 100; Probability of crossover: 0.8; Probability of mutation: 0.2.

Here crossover is performed by 2 points real value crossover. Two points (point1 and point2) are selected randomly, where point1 < point2, and point1 > 1, point2 < n, n is the number of genes (here are weights) of one individual NN. For mutation, a small random value between 0.1 and 0.2 is added to every weight of selected member that would perform mutation. After NN is trained the best weights of the NN and the classification rates are saved for the further features selection.

All the programs are implemented using C language on UNIX platforms.

[2] Threshold was initially set to 0.5 as it is the middle value between 0 and 1. However it is not necessarily the most ideal value to achieve the best possible classification rate. In the later experiments it is changed to find the best classification rate.

3.3.2. Discriminant analysis

Discriminant analysis as a whole is concerned with the relationship between a categorical variable and a set of inter-related variables. More precisely, suppose there is a finite number say k of distinct populations, categories, classes or groups. In discriminant analysis the existence of the groups is known a priori, for example, in the case study considered here we know there are two ($k = 2$) types of mammograms namely malignant and benign.

In order to classify a particular individual as to which of the two groups it belongs, the procedure would be to compare the distance (Mahalanobis distance) of the particular individual of unknown origin from the mean of the respective groups. That is, after measuring the appropriate variables (these are 14 features extracted from mammograms in our case study) compute the Mahalanobis distance of it from each group mean and classify the mammogram as follows:

Rule: Allocate a mammogram with the observation vector X to GP1 (malignant) if

$$(x-\bar{x}_1)'S^{-1}(x-\bar{x}_1) < (x-\bar{x}_2)'S^{-1}(x-\bar{x}_2)$$

that is: $a_1'x + c_1 > a_2'x + c_2$.

Otherwise as benign (GP2); where $\bar{x}_i$ $(i = 1, 2)$ are the group mean vector of the known two groups and S is the common covariance matrix and $a_i = S^{-1}\bar{x}_i, c_i = \bar{x}_i'S^{-1}\bar{x}_i; i = 1, 2$.

Here we make a very strong assumption that the two groups have equal variance covariance matrix. This aspect was examined and found reasonably satisfied using Box-M test statistic under the assumption of normality. The above rule will lead to a linear discriminant function (ldf) which is easy to handle. However, we also examined the quadratic discriminant function (qdf) by relaxing the above strong assumption of equal variance-covariance and the results are reported and compared. For a detailed account of discriminant analysis the readers are referred to McLachlan (1992).

3.3.3. Logistic regression

The logistic regression model is widely used in survival analysis, where the response y is typically measured as 0 or 1, depending on whether the experimental unit, for example, a patient survives or not. The same concept can be used in classification problem in classifying whether the person has benign or malignant tumor based on certain features. Logistic regression model for a binary dependent variable can be written as

$$E(y) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}$$

where $y = 1$, if the patient has malignant tumor; $y = 0$, if the patient has benign tumor.

$E(y) = P(\text{patient is malignant}) = \pi$

$X_1, X_2, \ldots, X_k$ are quantitative or qualitative independent variables.

Estimates of β parameters in the logistic model can be obtained by using maximum likelihood estimation technique. This method which is used by SPSS package has certain desirable properties as compared to ordinary least square method which can not be used if errors are not normally distributed.

4. Experimental results

A total of 67 microcalcification areas were extracted from the digital mammograms taken from a Benchmark database for the experiments. The experiments presented here were run using 47 microcalcification areas (24 benign, 23 malignant) for training and 20 microcalcification areas (11 benign, 9 cancer) were used for testing. All the programs were implemented in C language.

Many experiments using different parameters were run to find the feature or combination of features that best classifies a microcalcification area into benign and malignant. It was also performed to determine the ideal neural network parameter settings for microcalcification classification with the selected feature set. The experiments were conducted by the classification rate of testing set to calculate the fitness for reproduction of Genetic feature selection. The number of hidden units and output threshold were adjusted in the experiments to find the combination of the features and NN structures, which can achieve the best classification rate. The results of the experiments are described as following.

In all the tables below, the column 'Features' was described by using the values 0s and 1s. The 0 means the responded feature is not selected and the 1 means the feature is selected. "B-E" (Benign-Error) is used to represent the number of classification errors for benign microcalcifications of the whole training set or testing set. "M-E" (Malignant-Error) represents the number of classification errors for malignant microcacifications of the whole training set or testing set. "T-E" (Total-Error) refers to the number of classification errors for all the microcalcifications of the whole training set or testing set. "T-rate" is the abbreviation of Total-classification Rate. It is calculated by (TN-"T-E")/TN, where TN is the total number of experiment samples.

4.1. Experiments using threshold 0.5 and hidden units 2–18

When analyzing the results of the experiments from 2 hidden units to 18 hidden units, it is noticed that there are a few features which are more frequently selected than the others are. In every experiment, most of the feature sets that get the highest classification rate are selected many times or are the most frequently selected in all the generations.

Table 1 gathered all the feature combinations, which get the highest classification rate using different number of hidden units. We can see that in every selection, feature number 6 and feature number 7 are selected constantly. The 8th feature is almost selected by all of them, too. Fig. 2 indicates the frequency of every feature occurred in the feature sets which gave the high classification rate above 80%.

4.2. Experiments using threshold 0.4 and threshold 0.3

All previous experiments were carried out using 0.5 as the output threshold for classification, because it is the middle value between 0 and 1. It was found from the results that in most of the cases the benign classification rates are higher than malignant classification rates. Here change the threshold lower to do the further experiments and find out if it can achieve better results. Actually, the threshold 0.6 was used to do the experiment too. The result is the same as predicted. It is not better than using threshold 0.5. That is probably because the benign classification rate is higher than malignant, but not opposite.

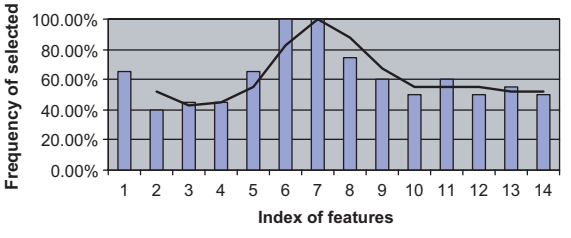

Fig. 2. The selection rate of every feature in the experiments with output threshold 0.5.

Table 1
The highest classification rate from experiments of different hidden units

Features (1-selected)	Hidden units	Training set				Testing set			
		B-E	M-E	T-E	T-rate (%)	B-E	M-E	T-E	T-rate (%)
01000111110100	2	2	11	13	72.3	1	3	4	80.0
10101111000100	8	6	10	16	66.0	1	3	4	80.0
01010111111010	10	10	6	16	66.0	1	3	4	80.0
10001111010001	12	2	12	14	70.2	1	3	4	80.0
01110111101111	12	5	9	14	70.2	1	3	4	80.0
00011111000011	14	8	11	19	59.6	3	1	4	80.0
10000111100000	16	3	11	14	70.2	1	3	4	80.0
01000110101000	16	3	15	18	61.7	0	4	4	80.0
11100110011100	18	6	9	15	61.7	1	3	4	80.0

Table 2
The feature selection reached classification rate >80% using threshold 0.4

Features (1-selected)	Hidden units	Training set				Testing set			
		B-E	M-E	T-E	T-rate (%)	B-E	M-E	T-E	T-rate (%)
00000110000001	4	8	6	14	70.2	3	1	4	80.0
11001111000111	6	2	15	17	63.8	1	3	4	80.0
01000111000011	8	6	10	16	66.0	0	3	3	85.0
10011011100011	8	11	6	17	63.8	4	0	4	80.0
01100011010111	8	7	9	16	66.0	0	4	4	80.0
01100110110000	12	6	11	17	63.8	1	3	4	80.0
00000110101100	14	3	16	19	59.6	0	4	4	80.0
11000110110110	16	2	13	15	68.1	1	3	4	80.0
00101110110010	18	2	12	14	70.2	1	3	4	80.0

Table 2 shows the results of the experiments by using output threshold 0.4. It presents the feature subsets what achieved the classification rate of testing set not less than 80%. Obviously, in every feature set include feature number 7. Except two of them, all include number 6. It is interesting that these two selections are the two, which did not reach the highest classification rate in the experiments using 8 hidden units.

This result is mainly consistent with that of the experiments using threshold 0.5. Fig. 3 shows the selected frequency of every feature in the high classification rate cases of the experiments using threshold 0.4. The trend line shows the consistent result as it is shown in Fig. 2.

Although the feature selection results are consistent in using threshold 0.4 and 0.5, in the experiments using threshold 0.4 achieved the higher classification rate. Further experiments using 0.3 as the output threshold were run in order to get more evidence for the consistent feature selection result, and expected to reach higher classification rate.

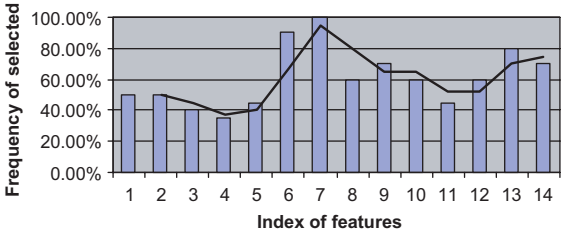

Fig. 3. The selection rate of every feature in the experiments with output threshold 0.4.

It is not surprising that the feature selection shows the similar result as it appeared in the previous experiments. Number 7, feature 'modified skew' is still most frequently selected, and traditional skew here is selected more than it is in the previous experiments. A few feature subsets have achieved the highest classification rate 85.0%. More feature combinations classified the testing set with correct rate 80%. Table 3 shows feature combinations, which reached 85.0% classification rate so far.

A new character of the results is that all the feature subsets here include traditional standard deviation and skew, beside the most popular feature modified standard skew. Another feature, boundary average grey level is also selected by every one of them. More experiments using 0.2 as threshold were conducted, which achieved the highest classification rate 80.0% with 4 and 8 hidden units. The overall results were not better than in previous experiments.

Table 3
The feature selections reached classification rate 85% using different number of hidden units and thresholds experiments of threshold 0.4

Features (1-selected)	Hidden units	Threshold	Training T-rate (%)	Testing T-rate (%)
11100111011111	2	0.3	63.8	85.0
11111011111011	6	0.3	68.1	85.0
00011011010011	8	0.3	63.8	85.0
01000111000011	8	0.4	66.0	85.0

Table 4
The selected features used by discriminant classifier

Features (1-selected)	Training set				Testing set			
	B-E	M-E	T-E	T-rate (%)	B-E	M-E	T-E	T-rate (%)
11001111011111	7	6	13	72.3	1	5	6	70.0
11111011111011	5	6	11	76.6	3	7	10	50.0
00110011010011	9	5	14	70.2	5	5	10	50.0
10000111000011	10	7	17	63.8	2	6	8	60.0

Table 5
The selected features used by logistic regression technique for classification using cutoff 0.4

Features (1-selected)	Training set				Testing set			
	B-E	M-E	T-E	T-rate (%)	B-E	M-E	T-E	T-rate (%)
11001111011111	10	5	15	68.1	3	3	6	70.0
11111011111011	7	3	10	78.7	4	5	9	55.0
00110011010011	9	3	12	74.5	5	5	10	50.0
10000111000011	12	2	14	70.2	5	1	6	70.0

The experiments were also conducted using discriminant and logistic regression techniques and the results are displayed in Tables 4 and 5. As it can be seen, the results are much worse than using neural classifier.

5. Discussion and analysis

As we can see from the tables in previous section, the highest classification rate 85% is achieved by NN with a few sets of selected features. Both statistical models used the selected sets of features, and both get the highest classification rate 70%. However the lowest classification rate reached by the statistical methods with the selected feature set is only 50%.

This is not hard to explain. Firstly, the feature selection is based on NN and conducted by the classification rate of testing sets. Secondly, NN results are affected by many factors: the initial parameters, hidden units, even different training methods etc. The main factor that limited the classification rate can be the database and feature extraction limitation. This can be explained by comparing the number of training set used for training. When using 37 samples for training the classification rate of training set reached 86.5%. When using 47 samples for training the highest training classification rate is only 80.9%. This indicates that it is hard to find a proper NN for the features extracted from the database for classification. When we used 47 samples for making the models by using statistical methods, we found 4 of the cases were always misclassified.

Table 6
Classifications using 3 features with different classifiers

Classifier	Training set classification rate (%)	Testing set classification rate (%)
NN	61.7	80.0
Discriminant analysis	72.3	70.0
Logistic regression	72.3	60.0

We also used the random selection of feature sets and did the experiments using neural network, discriminant analysis and logistic regression methods respectively. We proved that the feature selection of GA is effective. Observing the results, we found that all the selections with the highest classification rate include number 7, 13 and 14. So we tried to use only these 3 features and did further experiments. The results are shown in Table 6.

6. Conclusions and further research

We have proposed and investigated a GA based feature selection method and three classifiers. A few feature subsets are selected as the best for microcalcification classification. The highest classification rate of 85.0% for testing set is achieved using the proposed feature selection method with a NN classifier. By using the statistical techniques we validated that GA for features selection is very effective. We can say that the 3 features such as modified skew, standard deviation and skew are most significant and effective among our 14 features. Besides these, histogram, modified standard deviation and boundary average grey level should also be considered as more significant features than others. A further research needs to be conducted by adding more features in the whole set for further selection on larger database.

References

Breast Cancer Facts, 2002. http://www.breastcancerfund.org/disease_facts.htm.

Chitre, Y., Dhawan, A.P., Moskowitz, M., 1993. Artificial neural network based classification of mammographic microcalcifications using image structure features, Internat. J. Pattern Recognition Artificial Intell., World Scientific Publishing, Vol. 7. No. 6, pp. 1377–1401..

Chris, C., Tina, Y., 1997. Performance Enhanced Genetic Programming, Evolutionary Programming VI, Lecture Notes in Computer Science 1213, pp. 87–100.

Cost, S., Salzberg, S., 1996. A weighted nearest neighbor algorithm for learning with symbolic features. Machine Learning, 57–58.

Guerra-Salcedo, C., Whitley, D., 1999. Genetic Approach to Feature Selection for Ensemble Creation, GECCO-99, http://www.cs.colostate.edu/~genitor/Pubs.html.

Guerra-Salcedo, C., Chen, S., Whitley, D., Stmith, S., 1999. Fast and Accurate Feature Selection Using Hybrid Genetic Strategies, GECCO-99, 1999, http://www.cs.colostate.edu/~genitor/Pubs.html.

Heath, M., Bowyer, K.W., Kopans, D. et al., 1998. Current status of the digital database for screening mammography. In: Digital Mammography. Kluwer Academic Publishers, pp. 457–460.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Machine Intell. 20 (8), 832–844.

Jiang, Y., Nishikawa, R.M., Wolverton, D.E., Metz, C.E., Giger, M.L., Schmidt, R.A., Vyborny, C.J., Doi, K., 1996. Malignant and benign clustered microcalcifications: automated feature analysis and classification. Radiology 198, 671–678.

John, G., Kohavi, R., Pfleger, K., 1994. Irrelevant Features and the Subset Selection Problem. In: Proceeding of 11th International Conference Machine Learning. Morgan Kaufmann, San Francisco, pp. 121–129.

Kevin, S., Christopher, C., Kevin, W., 1993. Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. Internat. J. Pattern Recognition Artificial Intell. 7 (6), 1417–1436.

Kohavi, R., Somerfield, D., 1995. Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology, The First International Conference on Knowledge Discovery and Data Mining, http://citeseer.nj.nec.com/kohavi95feature.html.

Liu, H., Setiono, R., 1966. Aprobabilistic Approach to Feature Selection—A filter Solution. In: Proceeding of 13th International Conference Machine Learning. Morgan Kaufmann, pp. 319–327.

Maria-luiza, A., Osmar, Z., Alexandru, C., 2001. Application of Data Mining Techniques for Medical Image Classification, In: Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, pp. 94–101.

Marcoz, Y., and Torres-Torriti, M., 2001. Feature Selection using Genetic Algorithms in SAR airborne imagery, http://www.ifrance.com/silkswift/ga/.

McLachlan, G.J., 1992. Discriminant Analysis and Statistical Pattern Recognition. John Wiley.

Metin, N., Heang-Ping, C., Berkman, S., 2002. Optimal neural network architecture selection: improvement in computerized detection of microcalcifications. Acad Radiol. 9, 420–429.

Qian, W., Clarke, L.P., 1994. Adaptive multistage nonlinear filtering and wavelet for medical image enhancement. ICIP 3, 711–715.

Qian, W., Sun, X.J., Song, D.S., Clark, R.A., 2001. Digital mammography: wavelet transform and kalman-filtering neural network in mass segmentation and detection. Acad Radiol. 8, 1074–1082.

Racz, J., and Nieniewski, M., 2000. Computer Aided Diagnosis based on Analysis of Microcalcifications, IDWM2000, Canada, http://www.sunnybrook.on.ca/~iwdm.

Shen, L., Rangayyan, R.M., Desautels, J.E.L., 1994. Detection and classification mammography clacifications. International Journal of Pattern Recognition and Artificial Intelligence, World Scientific Publishing, pp. 1403–1416.

Shen, L., Rangayyan, R.M., Desautels, J.E.L., 1994. Application of shape analysis to mammographic calcifications. IEEE Trans. Medical Imaging 13, 263–274.

Verma, B. K., 1998. A neural network based technique to locate and classify microcalcifications in digital mammograms, IEEE World Congress on Comput. Intell., WCCI'98, Anchorage, USA, pp. 2163–2168.

Verma, B.K., 1999. Comparative evaluation of two neural network based techniques for classification of microcalcifi-

cations in digital mammograms. Knowledge and Information System 1 (1), 107–117.

Verma, B., Zakos, J., 2001. A computer-aided diagnosis system for digital mammograms based on fuzzy-neural and feature extraction techniques. IEEE Trans. Information Technol. Biomedicine 5 (1), 46–54.

Zheng, B., Qian, W., Clarke, L., 1994. Multistage neural network for pattern recognition in mammogram screening. IEEE Internat. Conf. Neural Networks ICNN, 3437–3447.

Zokos, J., 1998. Computer-aided Diagnosis of Digital Mammograms Using Computational Intelligence Techniques, Honours Thesis, Griffith University.