Knowledge-discovery incorporated evolutionary search for microcalcification detection in breast cancer diagnosis

Yonghong Peng a,*, Bin Yao b, Jianmin Jiang b

a Department of Computing, University of Bradford, West Yorkshire BD7 1DP, UK
b Department of Electronic Imaging and Media Communications, University of Bradford, West Yorkshire BD7 1DP, UK

Received 6 February 2005; received in revised form 13 September 2005; accepted 29 September 2005

Summary

Objectives: The presence of microcalcifications (MCs), clusters of tiny calcium deposits that appear as small bright spots in a mammogram, has been considered as a very important indicator for breast cancer diagnosis. Much research has been performed for developing computer-aided systems for the accurate identification of MCs, however, the computer-based automatic detection of MCs has been shown difficult because of the complicated nature of surrounding of breast tissue, the variation of MCs in shape, orientation, brightness and size.

Methods and materials: This paper presents a new approach for the effective detection of MCs by incorporating a knowledge-discovery mechanism in the genetic algorithm (GA). In the proposed approach, called knowledge-discovery incorporated genetic algorithm (KD-GA), the genetic algorithm is used to search for the bright spots in mammogram and a knowledge-discovery mechanism is integrated to improve the performance of the GA. The function of the knowledge-discovery mechanism includes evaluating the possibility of a bright spot being a true MC, and adaptively adjusting the associated fitness values. The adjustment of fitness is to indirectly guide the GA to extract the true MCs and eliminate the false MCs (FMCs) accordingly.

Results and conclusions: The experimental results demonstrate that the incorporation of knowledge-discovery mechanism into the genetic algorithm is able to eliminate the FMCs and produce improved performance comparing with the conventional GA methods. Furthermore, the experimental results show that the proposed KD-GA
1. Introduction

Breast cancer is the most common form of cancer in women in the world, especially in the western countries. It affects more than 8% of women during their lives and has become a leading cause of mortality in women in each year [1]. The early detection of breast cancer is vital for the success of treatment, so as to increase the chance of survival for patients. Of all the diagnostic methods for breast cancer, the most effective method is the use of mammography [2,3]. As a result of the implementation of nationwide screening programs in many western countries, the number of mammograms to be analyzed is enormous. Careful manual reading to search for signs of disease (which are frequently small or subtle) is labour intensive, very time-consuming, and demands great concentration. For improving the accuracy, double reading of mammograms (consecutive reading by two radiologists) is necessary and widely applied in many countries. This increases significantly the costs and manpower. These problems motivate the development of computer-aided diagnosis (CAD) systems to assist medical staff to achieve high efficiency and effectiveness in the early detection of breast cancer [1].

One of the important early symptoms of breast cancer in the mammograms is the appearance of clustered MCs [4,5]. Microcalcifications are tiny calcium deposits that appear as small bright spots in a mammogram, and have a higher X-ray attenuation than the normal breast tissue [6]. The MC detection in the early stage of breast cancer has become one important task of developing computer-aided diagnosis systems [7], however, the MCs are usually difficult to be detected automatically because of the surrounding breast tissue, especially the associated large variation of MCs in shape, orientation, brightness and size during the early stage of the breast cancer. Extensive research has been performed, during the past decades, to develop computer-aided systems for improving the accuracy of the diagnosis, reducing false positives and decreasing the workload for the radiologists and doctors. However, the performance of the existing techniques (especially the false positives) still needs to be improved significantly.

Many image processing algorithms and pattern recognition methods have been proposed for the processing of mammograms and identification of MCs. A higher-order statistics-based method was developed in [8] and methods based on weighted difference of Gaussian filtering were used in [9]. Wavelet-based approaches were studied in [10], and a multi-scale analysis-based detection method was proposed in [11]. Wang and Karayiannis [12] designed their own wavelets to detect MCs based on the observation that the MCs consist mainly of high-frequency components in the spectrum of mammograms. Furthermore, fuzzy logic approaches and neural network approaches have been investigated in [13–17] respectively. Since huge numbers of medical images has been accumulated, the application of data mining and knowledge-discovery techniques have been considered as an attractive area in medical image domain [18,19]. Several classical data mining techniques, such as the Support Vector Machine (SVM) [20,21] and association rules classification [22] have been recently applied for mammograms classification and have improved the understanding of breast cancer diagnosis.

The GA is an evolutionary search method that search for the optimal solutions by means of a set of evolutionary operations, including crossover, mutation and elitism of chromosomes. GA methods have been successfully applied for image processing. For example, in [23] a GA method is used to cluster the image pixels into a number of groups for image segmentation. Due to the complexity of search space, the conventional GA methods tend to slow down the search process and may produce many local optimal solutions. A GA-based MC detection approach has been developed recently in our research [29]. It has been shown that, by many experimental results obtained in our research (e.g. [29]), the conventional GA methods are capable of detecting the MCs (appear as bright spots) in mammogram images, of which, however, many have been confirmed being FMC spots, namely many MCs detected are actually not true MCs. The reduction of false positives is important in real applications as they lead to the unnecessary re-inspection and treatment, which drain resources, and cause unnecessary concern. It thus is necessary to improve the accuracy by reducing the false positives. On the other hand, as a false negative leads to postponed treatment, failure to treat or even death, the reduction of false positives must be performed carefully so as not to increase the false negatives.

This paper presents a new method, called knowledge-discovery incorporated genetic algorithm
(KD-GA), to effectively reduce the false positives by means of providing knowledge-based guide for the system in searching for the true MCs from mammograms. Incorporation of external knowledge, such as expert knowledge and meta-heuristics, into the evolutionary searching process has gained increasing interests in recent years[24]. Different from the conventional GA methods, which search for objects according to a pre-defined fitness function, the proposed approach integrates a knowledge-based evaluation system to adaptive update the fitness value of the associated chromosomes (clusters of image pixels). The fitness value measures ‘how good the searched object is’. The knowledge to be applied for the object evaluation should reflect the visual perception about the interesting object (the MCs in this particular case). In order to extract the knowledge for the fitness evaluation, a knowledge-discovery mechanism is incorporated, which extracts the knowledge characterizing the true MCs in terms of shape and texture. The knowledge-based intelligent adjustment of fitness function is intended to guide the GA to extract true MCs and eliminate the FMCs.

The contribution of this research can be summarized as following: (1) it provides a new approach for the incorporation of knowledge-discovery mechanism in the evolutionary search process; (2) the proposed approach presented is highly generic that can be well applied in medical diagnosis and image segmentation.

This paper is organized as follows. A three-step framework for MC detection, based on the evolutionary search, is introduced in Section 2. Section 3 discusses the knowledge representation for MC identification, while Section 4 details the method and algorithm for the proposed method of knowledge-discovery incorporation in the evolutionary search process for the MC detection. Section 5 presents the experimental results, and conclusions are given in Section 6.

2. Microcalcification detection framework based on evolutionary search

The main motivation behind this research is to develop a highly generic method with high performances for the automatic detection of MCs and other medical applications. This research is built based on the incorporation of a knowledge-discovery mechanism to extract the knowledge for enhancing in the evolutionary search process. The schematic approach proposed is illustrated in Fig. 1, which consists of three main steps:

(a) the extraction of bright spots (including possibly MCs and FMCs);
(b) the identification of MCs;
(c) the knowledge-based feedback to guide the GA to effectively search for the true MCs, and eliminate the FMCs.

The extraction of bright spots is performed by means of a pixel clustering approach using GA, which searches for the pixels that are similar in terms of the characteristics of their surroundings. As among the bright spots some are true MCs while some are false MCs, the membership of the bright spots to MCs is thus evaluated in the second step. The evaluation results are applied, in the third step, to provide guidance for the GA to search further for the true MCs. These three steps form a close-loop process of searching and evolution, by which the MCs are extracted while the FMCs are eliminated.

3. Knowledge representation for microcalcification identification

Recent research from the evolutionary computing shows that an effective way to improve the effi-
ciency in complex searching space is to incorporate external knowledge [24], by means of the population representation, initialization, genetic operations, and fitness evaluation. Appropriate representation of the searching objects is particularly crucial for a successful knowledge-incorporation scheme. This section discusses the representation of pixels and the bright spots.

3.1. Representation of pixels

In this paper, one pixel located at \((i, j)\) is characterized by the surrounding pixels within a certain size window. Let the grey level of pixel located at \((k, l)\) be denoted as \(p(k, l)\), and suppose an \(m \times n\) window is used (\(m, n\) are odd integers, and \(m = n = 9\) is used in this study), the average (mean) and the standard deviation (std) of the grey levels of surrounding pixels within the window are used to characterize the associated pixels:

\[
\text{mean}(i, j) = \frac{\sum_{l=-\lfloor n/2 \rfloor+1}^{\lfloor n/2 \rfloor} \sum_{k=-\lfloor m/2 \rfloor+1}^{\lfloor m/2 \rfloor} p(k, l)}{m \times n}
\]

\[
\text{std}(i, j) = \sqrt{\frac{\sum_{l=-\lfloor n/2 \rfloor+1}^{\lfloor n/2 \rfloor} \sum_{k=-\lfloor m/2 \rfloor+1}^{\lfloor m/2 \rfloor} (p(k, l) - \text{mean}(i, j))^2}{m \times n}}
\]

where \((i, j)\) is the location of the associated pixel.

It has been seen in our experiments that for a pixel of MC spot its mean and std are relatively larger than that of a FMC pixel [29], as illustrated in Fig. 2. This observation suggests the mean and std are applicable for characterizing the pixel in order to distinguish the pixels of a MC from the pixels of a FMC.

3.2. Representation of bright spots

A MC appears as a small bright spot in the mammogram image and the size and shape of MCs may be varied significantly at different stages of breast cancer. In this study, a total of nine features are designed to characterize a bright spot in terms of its shape and texture.

3.2.1. Shape descriptor

A careful visual investigation of hundred mammogram images suggests that the shape of a bright spot can be characterized by an ellipse with flexible orientation [15,29]. A total of five parameters are used to describe the shape of a bright spot: (1) a major axis and (2) a minor axis, denoted as major and minor, and (3) a parameter called long measuring the longest distance between two pixels within one spot, as shown in Fig. 3. Furthermore, (4) the associated perimeter \((L_B)\) and (5) the area \((A_m)\) of the associated bright spot are also calculated. Here the \(L_B\) denotes the number of pixels located at the spot boundary and the \(A_m\) is the number of pixels within the spot.

Based on these five parameters, four features are designed to characterize the shape of spots:

\[
x_1 = A_m
\]

\[
x_2 = \frac{\text{major}}{\text{minor}}
\]

\[
x_3 = \frac{\text{major}}{\text{long}}
\]

\[
x_4 = \frac{L_B}{A_m}
\]

3.2.2. Texture descriptor

The following features are designed to characterize the texture of a bright spot

\[
x_5 = \max_{(i, j) \in B} \{g_x(i, j)\}
\]

\[
x_6 = \max_{(i, j) \in B} \{g_y(i, j)\}
\]

![Figure 2](image1.png) Distribution of (mean, std).

![Figure 3](image2.png) Shape features for a MC spot.
where $B$ is the set of pixels located on the boundary of the bright spot under investigation, $g_x(i, j)$ and $g_y(i, j)$ are the gradient projected on the x- and y-axis respectively at pixel $(i, j)$, and $m_x = (1/L) \sum_{(i,j) \in B} |g_x(i, j)|$ and $m_y = (1/L) \sum_{(i,j) \in B} |g_y(i, j)|$.

### 3.3. Knowledge representation for MC identification

Among the bright spots extracted in GA searching process, some are FMCs (bright spots located in the normal areas of mammogram) while some are true MCs (bright spots that have been identified as true MCs by the doctors). For improve the performance in detecting true MCs, the extracted bright spots are necessarily evaluated if they are true MCs or FMCs. Please note that the MC evaluation is one sub-step of the whole MC search process, namely it does not produce the final results of MC detection but intends to recursively gain information for updating the associated fitness values and guide the GA to eliminate FMCs in-process.

In this paper, the bright spots are classified into two groups (MCs and FMCs), and the knowledge for spot evaluation is represented by a set of classification rules represented in the format of "if $x_1 \otimes a_1$ and ... $x_m \otimes a_m$ then the probability of this spot being a MC is $\delta \%$", in which $x_i$ denotes one of the features defined in Section 3.2, and the operator $\otimes$ could be 'bigger, $>$' or 'smaller, $<$'. One example of the rule to characterize an MC is "if $x_8 < 0.85$ and $x_9 > 0.7$, then the probability of this spot being an MC is $95\%$". By means of these rules, a spot extracted by GA can then be recognized as being an MC or not, and the associated probability.

### 4. Knowledge incorporated evolutionary searching for MC detection

In the proposed approach, as shown in Fig. 4, a GA clustering mechanism extracts recursively the bright spots, and the knowledge produced by the knowledge-discovery is used to guide the searching process to focus on extracting the MCs by means of two sub-routes: (1) to determine whether the bright spots extracted are MCs or FMCs, and (2) to adaptively update the fitness values for the associated spots.

Different from a conventional GA method, in which the fitness values are updated by a pre-defined fitness function, the proposed approach evaluates the fitness values of chromosomes based on the probability of the associated spot being a true MC or FMC. A spot being recognized as true MCs is assigned with higher fitness values, while the spots of FMCs will be given a lower fitness values. Through the evolutionary operations of GA, this adjustment of the fitness value effectively guides the GA automatically to converge on true MCs rather than FMCs.

#### 4.1. GA-based evolutionary searching for bright spots

The proposed method for spot extraction is illustrated in Fig. 5, in which the original image is first transformed into the feature space, and then they are clustered into $N$ groups in terms of their similarity in the feature space using GA method.

During the evolutionary searching, the GA generates a set of chromosomes in each searching step. A chromosome contains a totally of $N$ genes,
clusters of pixels, represented by \( \text{Ch}_i(i) = (\text{gene}_{1i}(i), \text{gene}_{2i}(i), \ldots \text{gene}_{Ni}(i)) \). Here each gene \((\text{gene}_{lk}(i) = (\text{mean}_{lk}(i), \text{std}_{lk}(i)))\), denotes the centre of the \( i \)-th cluster (group of pixels) of the \( k \)-th chromosome at the \( l \)-th generation. The selection of the value of \( N \) is based on the following considerations that: (i) the structure of a chromosome should provide sufficient variety to ensure that all important elements are included for obtaining a globally optimal solution; (ii) the balance between the optimisation performance and the computing cost. An empirical study was carried out for \( N = 3, 5, 7, 9, \) and \( 11 \). The results reveal that there were little difference among \( N = 5, 7, 9, \) or \( 11 \), and \( N = 3 \) is inferior. Therefore, we selected \( N = 5 \) in this study.

When the cluster centres have been generated, the memberships of a pixel \((i, j)\) to them are determined in terms of the distance between the pixel and the associated cluster centres:

\[
\mu_l(i, j) = \begin{cases} 
1 & \text{if } d_l(i, j) \leq d_{k \neq l}(i, j) \\
0 & \text{otherwise}
\end{cases}
\]  

(12)

where \( \mu_l(i, j) \) denotes membership of pixel \((i, j)\) to the cluster-\( l \), and \( d_l(i, j) \) is the distance between pixel \((i, j)\) to the centre of cluster-\( l \), which is calculated by:

\[
d_l(i, j) = \sqrt{(\text{mean}(i) - \text{mean}(i, j))^2 + (\text{std}(i) - \text{std}(i, j))^2}
\]

(13)

In the implementation of GA, tournament selection and two points crossover is used to generate offspring. For the gene (mean, std) selected for mutation, the following operation will be performed:

\[
\text{mean}_{l+1} = \text{mean}_l \times (1 \pm \delta)
\]

(14)

\[
\text{std}_{l+1} = \text{std}_l \times (1 \pm \delta)
\]

(15)

where \( \delta \) is a random number between 0 and 1. The probabilities for crossover and mutation are set to be 0.8 and 0.01 respectively in our implementation [29].

At each step of the evolutionary searching, each pixel is transformed from its original grey level to dichotomous white/black scale in that a group of pixels having relatively higher values of mean and std will be marked as a white pixel, while other pixels are marked as black. This transformation results in a distinctive white/black image, from which a bright spots will be easily identified, as illustrated in Fig. 5.

4.2. Knowledge-discovery incorporated evolutionary searching for MC detection

Due to the complexity of mammogram images, among these bright spots, some are true MCs and some are not. The spots generated by GA clustering are in-process evaluated based on the classification of bright spots into either MCs or FMCs. The result of MC identification is then used to control the GA search process by means of adaptively adjusting the fitness value of the associated chromosome.
When the spots have been confirmed as MCs/FMCs, then the associated fitness value will be increased/decreased accordingly. This proposed algorithm is shown in Fig. 6.

4.2.1. Knowledge-discovery for MC evaluation

In this study, a typical inductive learning method (decision tree induction) is employed for the knowledge-discovery for the evaluation of spots and distinguishing the true MCs from FMCs. Decision tree induction [25,26], such as the ID3 family (C4.5 and C5.0), has been widely used in the generation of the classification rules due to its good robustness and learning efficiency.

A decision tree is constructed through a recursively selection of appropriate features to partition the training dataset until the whole decision space has been completely partitioned into mutually exclusive subspaces. In this study, the spots are classified into MCs and FMCs, in terms of the shape and texture of the spot using the features presented in Section 3.2. These features are all continuous-valued data; they thus need to be discretized during the tree induction process. The method applied in our research is to select an appropriate threshold $T$ which partitions the domain of the associated feature $A$ into two intervals $A_1 = [\min(A), T]$ and $A_2 = (T, \max(A))$ [27]. The induction of decision trees from continuous-valued data is to repeat selecting an appropriate feature and the associated threshold $T$ that can produce the best classification power.

A typical decision tree and the associated rules for MC classification are shown in Fig. 7, in which each branch from root to a leaf forms a classification rule. Every leaf of the tree is associated with a parameter ($\alpha/\beta$), where the value of $\alpha$ is the number of cases in the training dataset that are correctly classified by the leaf, and $\beta$ is the number of them that are classified incorrectly by the leaf. For instance, the shortest branch from root to the left-most leaf shown in Fig. 7 forms a rule "If $x_9 > 0.825$, then the associated spot is FMC (23/1)" in which $\alpha = 23$ and $\beta = 1$ indicate that there are 23 cases are correctly classified into FMCs while one case is classified as FMCs incorrectly, and the feature $x_9$ is defined by Eq. (11). Based on the values of $\alpha$ and $\beta$, the probability of a spot being a MC (or a FMC) can be estimated by $p = \alpha / (\alpha + \beta)$.

4.2.2. Adjustment of fitness function

Following the identification of MCs and FMCs, the second step of evaluation is to adjust the fitness value for each bright spot. The method used is to award the associate chromosomes that have detected MCs successfully, and give a penalty to the chromosomes that are associated with FMCs. Two general rules have been applied in the presented algorithm for the fitness adjustment: (a) if a extracted spot is a true MC then the associated fitness is increased by $\text{fit}^{\text{new}} = \text{fit}^{\text{old}} + \Delta_1$; (b) if an extracted spot is in fact a FMC then the associated fitness is decreased by $\text{fit}^{\text{new}} = \text{fit}^{\text{old}} - \Delta_2$, where $\Delta_1 = \delta_1 A_{\text{max}}$ and $\Delta_2 = \delta_2 A_{\text{max}}$, and $\delta_1$ and $\delta_2$ are parameters to control the adjustment of fitness. The $\delta_1 = \delta_2 = p = \alpha / (\alpha + \beta)$ is used in this study.

5. Experiments and results

The proposed approach has been verified by mammogram samples selected from the digital mammography database (DDSM) at university of south Florida (http://marathon.csee.usf.edu/Mammography/Database.html, last accessed: 1 June 2005). In line with medical imaging specifications, all the mammograms selected are of pixel depth of 12 bits. In our experiments, the original mammogram were first segmented into small pieces (called regions of interest, ROIs) with sizes from 100 to 200. This is intended to evaluate the results in more statistical sense and reducing the computation time.

A total of 213 samples obtained from diverse background have been selected from the original mammograms for training purpose. Among the 213 training samples, 162 samples have been identified to include MCs and 51 samples are associated with

49

| $x_9 <= 0.825$ | $x_0 = 2.94$ |
| $x_9 <= 92$: FMC (19/2) | $x_9 > 92$: MC (156/2) |
| $x_9 > 2.94$ | $x_9 <= 23$: FMC (7) |
| $x_9 > 23$: MC (3) | $x_9 > 0.825$: FMC (23/1) |

(a) Decision tree
(b) Evaluation rules

Figure 7 Classification rules for the MC evaluation.
no MCs. To test the proposed approach, another separate dataset containing 288 samples are taken for testing. Among these 288 testing samples, 188 samples are associated with MCs and 100 samples containing no MCs. After the detection process, the binary image version of ROIs (white suspicious MCs with black background) are then presented for comparison and performance evaluation.

5.1. MC detection results

Fig. 8 illustrates results of MC detection obtained by the proposed method through three generations. In the original image of mammogram, as shown in Fig. 8(a), the MCs are really hardly seen. After careful inspection by the doctors, three MC spots have been identified in this image. The results of the first generation are shown in Fig. 8(b), in which the average fitness is 0.614. It is obvious that no MC has been extracted from the first generation. The evaluation of MCs is thus performed to reduce the fitness values for these spots that have been recognized as not true MCs, and increases the fitness for the spots that have been recognized as true MCs. This adjustment results in a new searching result, as shown in Fig. 8(c), in which the average fitness is 3.103. It shows that the MCs have roughly been revealed but some FMC spots are included. By performing the MC evaluation again, the fitness value has been increased significantly as 28.719 in the third generation, as shown in Fig. 8(d), from which all three MCs have been clearly revealed whilst the FMCs have been successfully eliminated. These results demonstrate the effectiveness of the interaction between GA and knowledge-discovery mechanism to extract the true MC spots and eliminate the FMC spots.

5.2. Performance evaluation

To demonstrate the effectiveness of the proposed method, a conventional GA clustering method (without incorporating external knowledge, denoted as GA), are compared with the results of our KD-GA method. In the conventional GA method, the fitness function is defined based on the Davise–Bouldin index, which is computed by the ratio of the sum of within-cluster scatter to the sum of between-cluster separation [28].

Let \( \{ C_i \} \) \((i = 1, \ldots, m)\) define the clusters. The scatter within \( C_i \) is computed in Eq. (16) and the distance between \( C_i \) and \( C_j \) is defined in Eq. (17):

\[
S_i,q = \left( \frac{1}{|C_i|} \sum_{x \in C_i} \| x - z_i \|_2^q \right)^{1/q} \quad (16)
\]
\[
d_{ij} = \| z_i - z_j \| \quad (17)
\]

where \( z_i \) is the centroid of cluster \( C_i \), and the \( |C_i| \) denotes the number of points in cluster \( C_i \). In this study, \( q = 2 \) is used. The DB index is computed according to these two elements:

\[
DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left\{ \frac{S_{ij} + S_{ji}}{d_{ij}} \right\} \quad (18)
\]

The fitness function for chromosome \( j \) is defined as \( 1/DB_j \), where \( DB_j \) is the DB index computed for the \( j \)th chromosome. This fitness function attempts to search for optimal clusters that have smallest scatter within each cluster and greatest distance between clusters.

Both the GA method and the KD-GA method have been tested by the same testing mammogram samples. Fig. 9 shows two samples of results obtained by these two methods respectively. In these two samples, four and three MCs have been recognized respectively by the doctors based on careful inspection. For the first sample, the KD-GA method has detected all the four MCs while the GA method missed two of them. For the second sample, the GA method extracted many FMCs and was not able to detect the true MCs, while the KD-GA had successfully revealed the true MCs by eliminating the FMCs and noise.

Experimental results for the 288 testing data confirm consistently that the KD-GA out-performs the conventional GA method. In line with the common practice in the community of digital mammography, a free-response receiver operating characteristic...
(FROC) curve, i.e. the true positive rate (TPR) versus the false positive rate (FPR), is adopted to evaluate the performance of the proposed approach. Both TPR and FPR are calculated based on the MC identification results for the 288 testing samples: $TPR = TP/(TP + FN)$ and $FPR = FP/(TN + FP)$ respectively. The TP (true positive) is the number of samples that contain true MCs and have also been recognized correctly as having MCs, and the FN (false negative) is the number of samples that contain MCs but have been recognized incorrectly as having no MCs. The TN (true negatives) is the number of samples containing no MCs and having been correctly classified as normal samples; and the FP (false positives) is the number of samples containing no MCs but having been incorrectly classified as having MC. For plotting the FROC curve, a set of varied TPR and FPR results are needed, which are calculated by changing the threshold of the fitness function of the GA.

The FROC curve for proposed KD-GA method is shown in Fig. 10, which shows that the proposed KD-GA algorithm achieves 96.8% and 98.9% accuracy at a cost of 20% and 40% false positive respectively. The FROC of conventional GA method is also presented in Fig. 10, which clearly the KD-GA consistently performs better than the GA method (the FROC curve located closer to the upper-left-hand corner indicates that the associated method has better overall performance). For example, for the same cost of 20% and 40% false positives, the GA method achieves 85.1% and 95.2% accuracy respectively, which indicates that the KD-GA method has improved the conventional method with 11.7% and 3.7% TPR respectively at the cost of 20% and 40% FPR.

6. Conclusions

This paper presents a novel approach towards effective MC detection in digital mammograms by incorporating a knowledge-discovery mechanism into the evolutionary searching process. In the proposed method, the GA method is applied to search for the bright spots that contain the possible MCs, and the knowledge-discovery method is used to evaluate if or not the extracted spots are MCs. These evaluation results are then applied to adjust adaptively the fitness values of the associated chromosome, which indirectly eliminates the FMC spots. Based on the incorporation of KD and the adaptive adjustment of fitness, it has been shown that the GA is able to focus on searching for the true MCs. Experimental results demonstrate in dramatic fashion the effectiveness of the proposed approach.
In summary, the main contributions of this research can be summarized as follows:

1. It presents a new method for the implementation of knowledge incorporation-based evolutionary search. The experimental results demonstrate that the proposed method achieves high performance in terms of the success rate in MC detections measured by both true positive rate and false positive rate.

2. The GA method designed in this study is efficient in searching for the bright spots through the transformed (mean and deviation) space. The experimental results illustrate that the transformation of original data into this new search space is effective in that it is no longer necessary to set a threshold for the extraction of spots. The GA is able to adaptively group the pixels in terms of their similarity in the (mean and deviation) space.

3. This research provides an effective method to integrate the GA process and knowledge-discovery mechanism. The experimental results demonstrate that the proposed method is able to converge effectively on the true MCs and to eliminate the FMC spots.

The knowledge-discovery incorporated evolutionary searching provides a new and effective platform for the MC detection and medical image inspection. The knowledge needed in GA searches can be self-updated, when new successful MC samples become available. This facilitates the system be self-enhanced and upgraded, and provides an effective approach for the design Computer-Aided Diagnosis systems for breast cancer.

Acknowledgement

Bin Yao is supported by the funding from the Yorkshire (UK) Cancer Research.

References


