IOP PUBLISHING PHYSICS IN MEDICINE AND BIOLOGY

Phys. Med. Biol. **53** (2008) 895–908 doi:10.1088/0031-9155/53/4/005

Decision optimization of case-based computer-aided decision systems using genetic algorithms with application to mammography

Maciej A Mazurowski[1], Piotr A Habas[1,3], Jacek M Zurada[1] and Georgia D Tourassi[2]

[1] Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY 40292, USA
[2] Department of Radiology, Duke University Medical Center, Durham, NC 27705, USA

E-mail: maciej.mazurowski@louisville.edu

Received 21 September 2007, in final form 28 November 2007
Published 16 January 2008
Online at stacks.iop.org/PMB/53/895

Abstract

This paper presents an optimization framework for improving case-based computer-aided decision (CB-CAD) systems. The underlying hypothesis of the study is that each example in the knowledge database of a medical decision support system has different importance in the decision making process. A new decision algorithm incorporating an importance weight for each example is proposed to account for these differences. The search for the best set of importance weights is defined as an optimization problem and a genetic algorithm is employed to solve it. The optimization process is tailored to maximize the system's performance according to clinically relevant evaluation criteria. The study was performed using a CAD system developed for the classification of regions of interests (ROIs) in mammograms as depicting masses or normal tissue. The system was constructed and evaluated using a dataset of ROIs extracted from the Digital Database for Screening Mammography (DDSM). Experimental results show that, according to receiver operator characteristic (ROC) analysis, the proposed method significantly improves the overall performance of the CAD system as well as its average specificity for high breast mass detection rates.

1. Introduction

Computer-aided decision (CAD) systems are playing an increasingly important role in medical diagnosis, with a wide variety being proposed within the last decade (Friedman *et al* 1999, van Ginneken *et al* 2001, Kawamoto *et al* 2005, Sampat *et al* 2005, Doi 2006). One of

[3] Present address: Department of Radiology, University of California San Francisco, San Francisco, CA 94143, USA.

the most important tasks of CAD systems is a classification task, where an incoming query case (e.g. a screening mammogram) has to be assigned to one of the prespecified groups (e.g. normal/abnormal). The most popular CAD classifiers are rule-based, where a certain number of previously acquired clinical cases (i.e. examples) is utilized to train the system. During training, decision rules are found. Subsequently, these rules are applied for the proper classification of new, unknown cases. In rule-based systems, all rules about decision making are embedded into the classifier and after the classifier is trained it works independently of the training dataset. The most common techniques for constructing such systems are based on probability principles (Duda *et al* 2000) or involve artificial or computational intelligence (Mitchell 1997). Artificial neural networks (ANNs) (Zurada 1992, Zhang 2000) are one of the most representative examples of the second group.

Recently, case-based CAD (CB-CAD) systems (Chang *et al* 2001, Schmidt *et al* 2001, Tourassi *et al* 2003, El-Naqa *et al* 2004a, 2004b, Tourassi *et al* 2007) have been gaining popularity in the medical domain. CB-CAD systems (also called evidence-based systems) utilize the principles of case-based reasoning (Aha *et al* 1991, Mitchell 1997). When classifying an incoming query, a CB-CAD system relies on a database of stored examples (called case base or knowledge database) which are compared to the query. The outcomes of these comparisons (i.e. similarities/dissimilarities) are used to make a decision regarding the query case. The main advantage of CB-CAD systems over rule-based systems is that CB-CAD systems require virtually no training. Therefore, the knowledge database can be constantly updated without the need for retraining the system.

With growing databases of examples, CB-CAD can be of great use. It must be stressed, however, that in order to obtain effective systems an optimal use of the existing evidence (clinical examples) must be assured. The CB-CAD systems currently reported in the literature are often based on the assumption that the examples stored in the knowledge database are equally important in the decision making process or their importance is dependent on their order of retrieval. The goal of this study is to challenge the above practice and test whether assigning different importance to the examples may improve the CB-CAD performance. The study hypothesis is based on the common sense observation that when a decision is made about a query case, its similarity to certain previous cases may be more important than similarity to others. Furthermore, it may occur that similarity to certain previous cases is simply misleading. The proposed framework accommodates both scenarios.

Following the study hypothesis, a new decision index is proposed that involves a vector of weights, each of which is a measure of importance of the particular image in the knowledge database. Then, a genetic algorithm (GA) (Michalewicz 1999, Eiben and Smith 2003) is utilized to find the optimal weight vector. GAs have been widely used to construct and optimize CAD tools for many clinical applications (Pena-Reyes and Sipper 2000), including breast cancer detection (Campanini and Lanceonelli 2006). These efforts aimed mainly toward optimizing parameters of CAD systems (Bevilacqua *et al* 2001, Gurcan *et al* 2002), optimizing the selection of features for classification (Sahiner *et al* 1996, Anastasio *et al* 1998, Sahiner *et al* 1998, Boroczky *et al* 2006), directly training the classifiers (Fogel *et al* 1998, Pena-Reyes and Sipper 1999), as well as other applications (Peng *et al* 2006).

The framework proposed here is tailored to optimize the performance of a system measured by clinically relevant indices. Receiver operator characteristic (ROC) and two related performance indices are used to optimize and test the system. Such practice will allow a better fit of the CAD system to clinical tasks.

The proposed optimization scheme is tested with respect to the featureless information-theoretic CAD (IT-CAD) system presented previously in Tourassi *et al* (2003). The system was developed for the automated detection of breast masses in screening mammograms. The

approach proposed here, however, is not limited to this system and may be applied to virtually any case-based CAD.

The article is organized as follows. Section 2 describes the original IT-CAD system. Section 3 presents the proposed modification of the decision function based on different importance of each example in the knowledge database. Section 4 defines the search for the importance weights of individual examples in the knowledge database as an optimization problem and introduces a genetic algorithm to solve it. Section 5 presents the experimental design including data handling and performance evaluation criteria. Section 6 presents the experimental results of the comparison between the modified and the original IT-CAD system. The study findings and future research directions are discussed in section 7.

2. Information-theoretic computer-aided decision system

2.1. Image database

The study was based on a set of mammograms extracted from the Digital Database for Screening Mammography (DDSM) (Heath *et al* 1998), collected at the University of South Florida. The mammographic films were digitized using the LUMISYS scanner at 50 μm per pixel. From the mammograms depicting true masses, 512×512 pixel regions of interests (ROI) were extracted. The resulting 901 mass ROIs (489 malignant and 412 benign masses) were centered on the physicians annotation. From the normal mammograms and from abnormal mammograms without any annotations for one of the breasts, 919 normal ROIs were extracted randomly. The normal ROIs were also 512×512 pixels in size. The prevalence of mass ROIs in the database was roughly 0.5.

The IT-CAD system described in the following paragraphs was developed to determine the presence or absence of a mass in a specific ROI. It is assumed that the ROI is centered on a location that is suspicious enough to warrant further evaluation. Typically the suspicious location is indicated either by a physician or a computerized, prescreening detection algorithm. The IT-CAD provides an evidence-based second opinion using the knowledge database of reference ROIs.

2.2. IT-CAD system overview

Figure 1 presents an overview of the IT-CAD system. The decision process is composed of two separate steps. First, an incoming query image is compared to all images in the knowledge database. The result of this comparison is a vector of similarity indices between the query image and each of the images in the database. Then, this vector, along with the ground truth for each of the images in the database, is utilized in the second step to make a binary decision, whether a query image depicts a mass or a normal tissue.

The decision process of case-based CAD systems, and IT-CAD in particular, aims at mimicking the radiologist's approach to decision making. Typically, the radiologist recalls the mammograms seen before along with their ground truth to make a proper classification for the examined query case. To implement this process in a CAD system, the CAD designer needs to select an algorithm for evaluating similarities between mammograms and an algorithm to make a decision based on those similarities. The algorithms implemented in IT-CAD are described in detail in the following sections for better understanding of the proposed improvement.

2.3. Image similarity assessment algorithm

To quantify the similarity between images, a so-called similarity measure has to be used. A wide variety of similarity assessment methods are available in the literature. A comprehensive

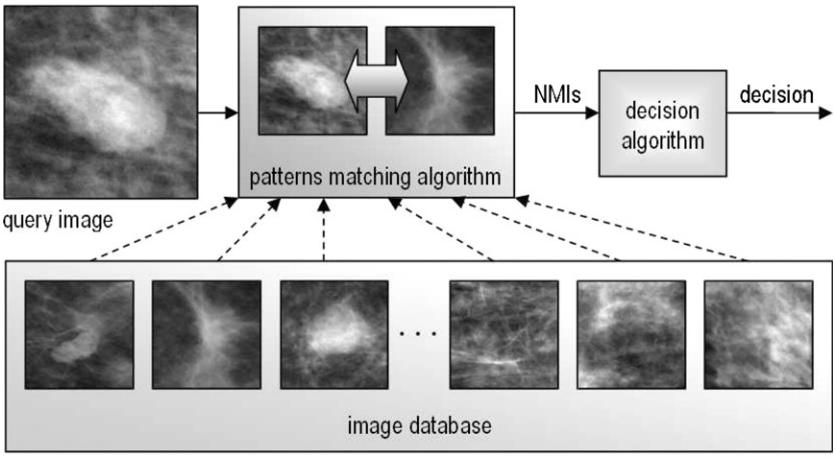

Figure 1. A block diagram of information-theoretic CAD system.

review of the topic can be found in Smeulders *et al* (2000). Typically image similarity is assessed using a set of features previously extracted from images. Here, a featureless approach is utilized. Namely, the classical statistical concept of mutual information (MI) is used (Cover 1991, Maes *et al* 1997). It has been shown to be a very efficient measure of similarity between images (Maes *et al* 1997) and was successfully applied in mammography CAD (Tourassi *et al* 2003, 2007). In information theory, mutual information, also called mutual entropy, describes the statistical dependence between two random variables. For discrete random variables, mutual information $I(X; Y)$ is defined as

$$I(X; Y) = \sum_x \sum_y P_{XY}(x, y) \log_2 \frac{P_{XY}(x, y)}{P_X(x) P_Y(y)}. \tag{1}$$

The mutual information measure can be applied to images by substituting random variables with intensity histograms. To find the value of I between two images X and Y, $P(X)$ and $P(Y)$ in equation (1) are estimated by the intensity histograms of images X and Y, respectively, and $P(X, Y)$ are estimated by the 2D joint histogram of the two images (Maes *et al* 1997). In this study, the normalized mutual information (NMI) index was used, defined as

$$\text{NMI} = \frac{2I(X, Y)}{H(X) + H(Y)} \tag{2}$$

where $H(X)$ and $H(Y)$ are the entropies of the individual images. The normalization results in NMI = 1 for identical images and NMI = 0 for two completely independent (i.e., unrelated) images.

2.4. Original decision algorithm

After the similarities between the query image and the images in the knowledge database are quantified using the NMI measure, a decision is made whether the query image corresponds to a mass or a normal case. Several approaches to this problem have been presented (Mitchell

1997, Chang *et al* 2001, Tourassi *et al* 2003). Tourassi *et al* (2003) proposed a decision index (DI) defined as

$$\mathrm{DI}_1(Q) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{NMI}(Q, M_i) - \frac{1}{n}\sum_{j=1}^{n}\mathrm{NMI}(Q, N_j) \tag{3}$$

where $\mathrm{NMI}(X, Y)$ is the normalized mutual information between images X and Y, Q is a query image, M_i are images corresponding to positive cases (masses), N_j are images corresponding to negative cases (normal), m and n are the number of positive and negative cases from the database, respectively. The two parts of this index express an average similarity to images depicting masses and to images depicting normal tissue. If the query image is on average more similar to images depicting masses, the average NMI with those images will be higher than the average NMI with images depicting normal tissue which will result in a high DI value. In contrast, when according to NMI, the query image is more similar to the images depicting normal tissue, the value of DI will be small. Finally, a threshold on the decision index has to be applied, such that when the decision index for a given query case is larger than the threshold value, the query is classified as depicting a mass.

The effectiveness of the IT-CAD system was shown in Tourassi *et al* (2003, 2007). This system, however, as many other case-based systems, relies on the assumption that all images in the database are equally important when classifying a query. It is likely, however, that this is not the case. It is easy to imagine that certain previously gathered examples are clinically more useful in the decision making process than others. It may also be the case that some examples in the database are misleading. The proposed optimization framework accommodates these possibilities.

3. Proposed decision algorithm

A more general decision function is introduced below. It offers the advantage that no assumption about equal importance of each image in the database is made *a priori*. It is defined as

$$\mathrm{DI}_2(Q) = \frac{1}{m}\sum_{i=1}^{m} w_{\mathrm{pos},i}\mathrm{NMI}(Q, M_i) - \frac{1}{n}\sum_{j=1}^{n} w_{\mathrm{neg},j}\mathrm{NMI}(Q, N_j) \tag{4}$$

where $w_{\mathrm{pos},i}$ is the importance weight associated with the ith positive example and $w_{\mathrm{neg},j}$ is the importance weight associated with the jth negative example. Low values of $w_{\mathrm{pos},i}$ or $w_{\mathrm{neg},j}$ diminish the contribution of similarity between a query image Q and image M_i or N_j, respectively, to the decision index. When the weight is equal to 0, a corresponding image will not play any role in the classification. Note also that DI_1 can be treated as a special case of DI_2, where the weights are equal to 1 (equal importance of all images in the database). Having a new definition of the decision function, the question arises of how to find the importance of each of the images in the database. Here, the problem of finding optimal weights is formalized as an optimization problem and a genetic algorithm is proposed to find an optimal solution.

4. Genetic algorithm to optimize decision

4.1. Definition of the optimization problem

To approach the presented problem as an optimization task, a suitable objective function (also called a cost function) needs to be defined. Many classification performance measures can be applied to a resulting CAD system as the objective function. In medical diagnosis, however,

the receiver operator characteristic curve (Bradley 1997, Obuchowski 2003, Fawcett 2006) is of particular interest and has been widely used to assess CAD classifiers. The ROC curve illustrates the relation between true positive fraction (TPF) and false positive fraction (FPF) for the full range of possible decision thresholds. For a given decision threshold, TPF and FPF are defined as

$$\mathrm{TPF} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \mathrm{FPF} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}$$

where TP is the number of true positive, FN is the number of false negative, FP is the number of false positive and TN is the number of true negative decisions. TPF is also known as sensitivity while (1 − FPF) is known as specificity.

A desired classifier has high values of TPF for all possible values of FPF. The overall performance of a classifier can be measured by the area under the ROC curve (AUC) or partial AUC ($_p$AUC). A larger AUC value indicates better overall performance of the classifier (the maximum possible value is 1). $_p$AUC index is of particular interest in medical decision problems, as it reflects the performance of a system when operating at the high sensitivity range. $_p$AUC denotes the area under the ROC curve that is above the TPF = p line.

In the proposed framework, either AUC or $_{0.9}$AUC index is utilized for optimization to show that the system can be optimized according to a selected, clinically relevant figure of merit. In this study, the non-parametric Wilcoxon approach (Bradley 1997) of estimating AUC and $_p$AUC was used. Note that, for a given set of examples in the database, the performance of the system measured by AUC and $_{0.9}$AUC depends solely on the vector $\mathbf{w}$ (see (4)) where $\mathbf{w}$ denotes the vector of importance weights. Therefore, the objective function is a function of $\mathbf{w}$ and will be denoted simply as $f^{\mathrm{AUC}}(\mathbf{w})$ or $f^{0.9\mathrm{AUC}}(\mathbf{w})$. Since there is one weight assigned to each example ($w_{\mathrm{pos},i}$ to positive examples and $w_{\mathrm{neg},i}$ to negative examples), the length of the vector $\mathbf{w}$ is equal to the number of cases ($n + m$) in the database of the system.

The presented problem can now be defined as the problem of finding a vector $\mathbf{w}^*$ such that $\forall_{\mathbf{w}} f^{\mathrm{AUC}}(\mathbf{w}) \leqslant f^{\mathrm{AUC}}(\mathbf{w}^*)$ and formally finding

$$\mathbf{w}^* = \arg\max_{\mathbf{w}}[f^{\mathrm{AUC}}(\mathbf{w})]. \tag{5}$$

When optimizing $_{0.9}$AUC, the corresponding problem is

$$\mathbf{w}^* = \arg\max_{\mathbf{w}}[f^{0.9\mathrm{AUC}}(\mathbf{w})]. \tag{6}$$

A large variety of optimization methods have been proposed for such optimization problems. Selection of the optimization technique depends on the properties of the objective function. Here, the objective function is not given by an arithmetic expression and can be only sampled for each value of $\mathbf{w}$. Furthermore, the objective function is not continuous and thus not differentiable. For these reasons the traditional optimization techniques such as the gradient descent (Bertsekas 1999) are not applicable. As a consequence, a genetic algorithm was used to find a solution.

4.2. Genetic algorithm to find an optimal solution

Genetic algorithm is an optimization technique loosely inspired by the theory of natural selection and genetics (Michalewicz 1999, Eiben and Smith 2003, Campanini and Lanceonelli 2006). The basic concept of the genetic algorithm is to treat different solutions to a problem as competing individuals. Solutions are represented in the chromosomes (vectors of numbers) of these individuals. The best solution evolves among the individuals by means of recombination, mutation and selection. Each individual is evaluated according to its fitness (the better the fitness is, the better the chances for survival and becoming a parent are), which is simply

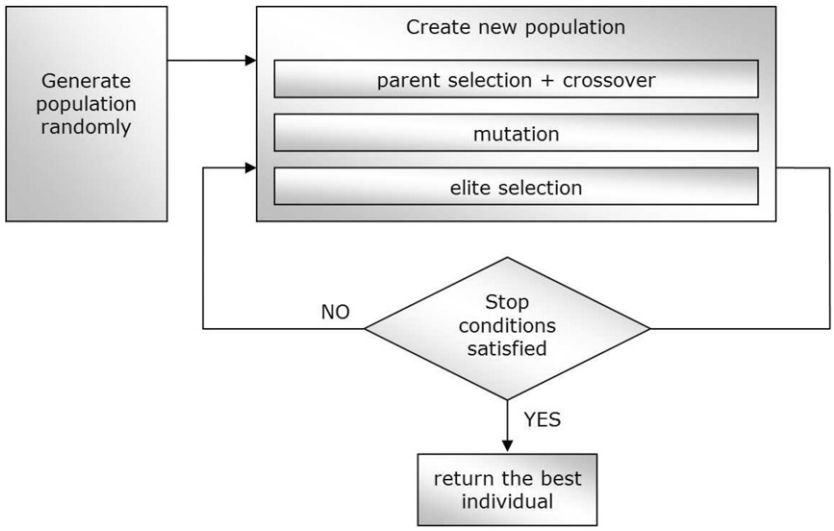

Figure 2. A block diagram of genetic algorithm.

a value of an objective function (i.e. $f^{\mathrm{AUC}}(\mathbf{w})$ or $f^{0.9\mathrm{AUC}}(\mathbf{w})$) in the point represented by its chromosome (i.e. $\mathbf{w}$). A diagram of a simple genetic algorithm is presented in figure 2. Different variations of GAs can be found in the literature.

When using a genetic algorithm, one must make a decision regarding the following properties: representation of the solutions in chromosome, population initialization, selection algorithm, recombination technique and mutation algorithm. The following is a short description of all these steps in the algorithm used in the study (for a more comprehensive description of these steps, see Michalewicz (1999) and Eiben and Smith (2003)).

- *Representation of solutions in chromosomes.* Since a candidate solution is a vector of weights, each chromosome has to represent a vector of real numbers. Therefore, a real-valued representation is used such that there is a real number on each position of the chromosome (each gene).
- *Population initialization.* Each position of each chromosome was initialized with a random value in a given, problem specific, range with uniform probability distribution.
- *Selection.* This operation selects the parents for the next generation. Here, the stochastic universal sampling algorithm (Gen and Cheng 1997) is used for this purpose. In this algorithm all individuals are assigned intervals proportional to their fitness. The intervals are placed on a line in random order. Then, starting from the beginning of the line, parents are picked by moving along the line in equal steps. This way each individual will be selected the expected number of times.
- *Recombination.* Recombination is used to generate new individuals by combining chromosomes of selected parents. Here, the uniform (also called random or scattered) crossover is used. It means that each gene of a child comes unaltered from one of the parents with equal probability for both parents, for each gene.
- *Mutation.* Mutation operation alters selected genes in the population to introduce additional random factor into the search. In this study, uniform mutation is used to replace the randomly selected genes with random values uniformly distributed in the range of the particular position.

Additionally, in the GA used in this paper an elite selection technique was applied. This technique assured that the best individuals always survive to a next generation.

5. Experimental design and assessment tools

To assess the effectiveness of the proposed optimization framework, the CAD system optimized by the GA (GA-IT-CAD) was compared to the original IT-CAD system with all the patterns in the database having equal importance.

To evaluate system performance, a ten-fold crossvalidation scheme was used. First, the entire dataset was divided into ten equal-size mutually exclusive parts (folds). Then, each fold was used once as a testing set leaving the remaining nine folds for CAD system development. This data handling scheme resulted in ten splits with 1638 examples in the development dataset and 182 examples in the testing dataset for each split.

For each split the development dataset was used to construct the GA-IT-CAD system while the test set was left for the final testing. Construction of the system consisted of two separate tasks. The first task was including images from the development dataset in the knowledge database of the GA-IT-CAD system for further use in the decision making process. The second task was to find the importance weights for the images in the database.

To complete the above tasks, the development dataset can be utilized in various ways. One way is to include only some of the images from the training dataset in the knowledge database of the GA-IT-CAD system. In such a case the remaining images can be utilized as queries to calculate the objective function ($f^{\mathrm{AUC}}(\mathbf{w})$) in a GA run in order to find an optimal set of weights for the images stored in the knowledge database. This approach, however, has significant disadvantages. Namely, not all images available in the development dataset are included in the knowledge database of the CAD and since the excluded images may be important for the problem at hand the performance of the resulting system could be significantly compromised. This approach should be used only for very large development datasets.

Since data availability is often limited in medical applications, more efficient data handling schemes should be applied to capitalize on the available clinical cases. Consistent with many CAD-related studies published before, the leave-one-out data handling scheme was used in the development process of this study (Efron and Tibshirani 1993).

Initially, each of the N images from the training dataset is included in the knowledge database of the GA-IT-CAD. Then, in the optimization process, the objective function for each individual ($f^{\mathrm{AUC}}(\mathbf{w})$ *or* $f^{0.9\mathrm{AUC}}(\mathbf{w})$) is computed in the following way: first, k decision indices are obtained, each of them by using one of the images from the knowledge database as a query. When Q is provided as a query to the CAD, it is temporarily excluded from the knowledge database and the decision index is calculated according to (4). This way the self-similarity (NMI(Q, Q)) of the image does not affect the resulting decision index. Such calculation is repeated for all images from the database, which results in k decision indices. The resulting decision indices along with the ground truth of the corresponding images are used to calculate $f^{\mathrm{AUC}}(\mathbf{w})$ or $f^{0.9\mathrm{AUC}}(\mathbf{w})$ using the Wilcoxon approach.

After optimization for each of the data splits, the 'trained' GA-IT-CAD system was tested on the testing set (i.e. the remaining part of the available database).

To compare the optimized CAD system with the original CAD system, a paired student t-test was used without an assumption about equal variances of the two populations to assure conservativeness of the results. A very similar scheme for statistical comparison of two models based on ten-fold crossvalidation is presented in Bradley (1997). Such scheme allows

Table 1. Comparison of AUC and $_{0.9}$AUC values for the original and optimized IT-CAD. The row 'Range' indicates the value range of the vector **w** elements.

Objective	AUC		$_{0.9}$AUC		Original
Range	[0, 1]	[−1, 1]	[0, 1]	[−1, 1]	IT-CAD
AUC	$\mathbf{0.907 \pm 0.024}$	0.905 ± 0.026	0.905 ± 0.022	0.894 ± 0.028	0.865 ± 0.030
$_{0.9}$AUC	0.533 ± 0.104	0.532 ± 0.100	$\mathbf{0.540 \pm 0.094}$	0.515 ± 0.076	0.466 ± 0.091

for incorporating in the analysis the uncertainty of performance estimation due to a finite test set as well as the random factor due to the stochastic nature of the GA.

6. Experimental results

The GA optimization was performed using the MATLAB programming environment and the MATLAB Genetic Algorithm and Direct Search Toolbox. The following standard parameters of the GA algorithm were used for the optimization. The number of individuals in the population (100) and the number of generations (100) were selected to provide extensive search without severely compromising the time complexity of the algorithm. Some experimentation was performed to select the optimal mutation ratio (0.01). The crossover fraction of 0.8 was a default value of this parameter as set in MATLAB.

Table 1 shows the average test performance of the original and optimized CAD systems across all ten splits. The optimization was performed with two different objective functions: $f^{\mathrm{AUC}}(\mathbf{w})$ and $f^{0.9\mathrm{AUC}}(\mathbf{w})$ and two different ranges of **w**: [−1, 1] and [0, 1] resulting in four separately analyzed scenarios (each combination of an objective function and the range of **w**). The **w** range of [0, 1] corresponds to a situation where the minimum importance of a pattern mean that the pattern is not considered in the decision process. When **w** is in the range of [−1, 1] a situation where some patterns are misleading (weights less than 0) is also taken into account. Thus, it is possible through a negative weight that an example will serve as an example of the opposite class in the decision process.

It is apparent from table 1 that the proposed optimization framework improved the overall performance of the CAD system from AUC $= 0.865 \pm 0.030$ to AUC $= 0.907 \pm 0.024$. The improvement was statistically significant (two-sided p-value less than 0.0005). The best performance was obtained when the CAD system was optimized using $f^{\mathrm{AUC}}(\mathbf{w})$ with **w** varying in [0, 1]. The overall performance with other optimization parameters (objective function and range of **w**) was also statistically significantly better than the performance of the original CAD system. The obtained results depended on the choice of the **w** range and consistently better performance was obtained for the range [0, 1].

It can also be seen that the proposed optimization framework allowed for a large and consistent improvement of the system in terms of highly clinically relevant partial area under the ROC curve. The improvement was obtained in nine out of ten splits. The mean $_{0.9}$AUC index improved from 0.466 ± 0.091 for the original CAD system to 0.540 ± 0.094 for the CAD system optimized with $f^{0.9\mathrm{AUC}}(\mathbf{w})$ and [0, 1] range for **w**. The improvement was statistically significant (two-sided p-value less than 0.001). A statistically significant improvement was observed for all objective functions and weight ranges investigated in this study (p-value was less than 0.05).

The ROC curves for the original CAD system and the optimized CAD systems were reconstructed by averaging the slope and intercept parameters of ROC curves across all splits. Parametric estimations of ROC curves were used for that purpose with the ROCKIT software developed at the University of Chicago (Metz *et al* 1998a, 1998b). The reconstructed curves

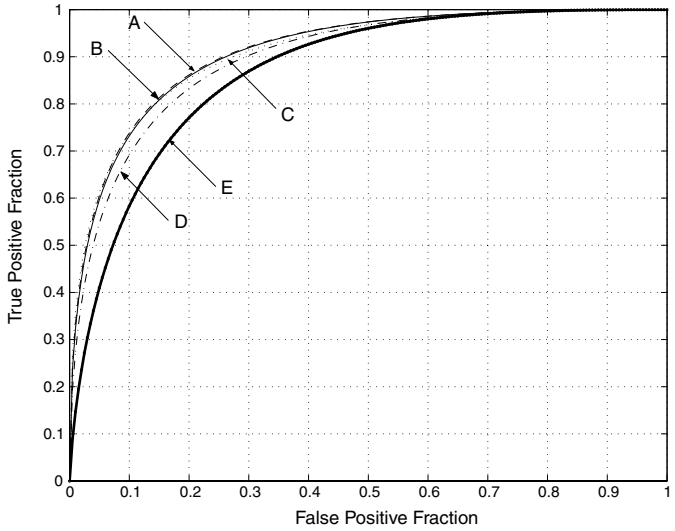

Figure 3. Averaged ROC curves: (A) GA-IT-CAD optimized with AUC and [0, 1] weight range (- - - -), (B) GA-IT-CAD optimized with $_{0.9}$AUC and [0, 1] weight range (——), (C) GA-IT-CAD optimized with AUC and [−1, 1] weight range (· · · · · ·), (D) GA-IT-CAD optimized with $_{0.9}$AUC and [−1, 1] weight range (— · —) and (E) original IT-CAD system (═══).

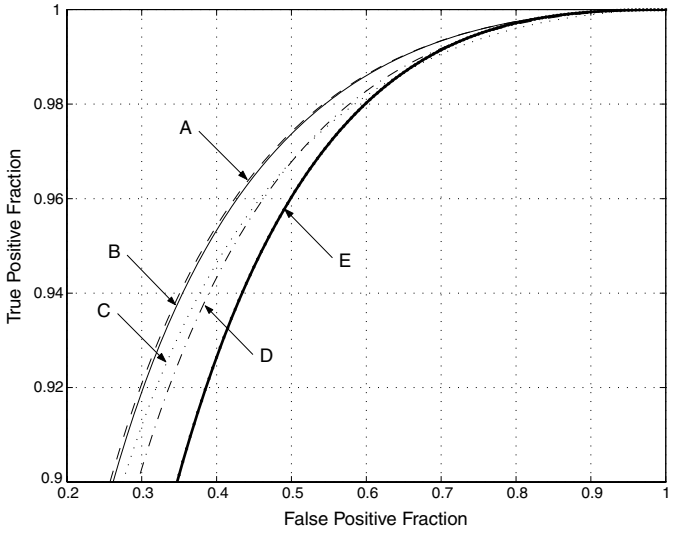

Figure 4. Averaged ROC curves for high sensitivities: (A) GA-IT-CAD optimized with AUC and [0, 1] weight range (- - - -), (B) GA-IT-CAD optimized with $_{0.9}$AUC and [0, 1] weight range (——), (C) GA-IT-CAD optimized with AUC and [−1, 1] weight range (· · · · · ·), (D) GA-IT-CAD optimized with $_{0.9}$AUC and [−1, 1] weight range (— · —) and (E) original IT-CAD system (═══).

are shown in figures 3 and 4. It can be seen that the proposed optimization framework improved the performance of the system in the regions of high specificity (for FPF = 0.1, TPF improved from 0.583 to 0.738) as well as in the regions of high sensitivity (for TPF = 0.9, specificity

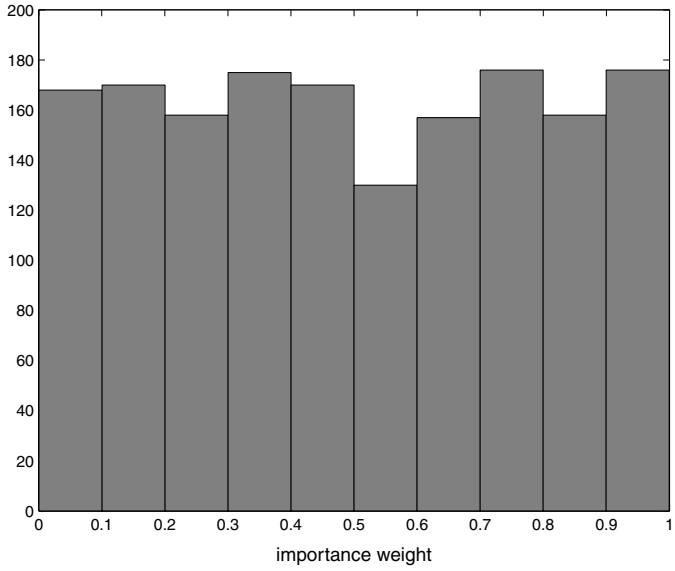

Figure 5. A typical histogram of weights assigned to images when system is optimized with AUC and [0, 1] weight range.

improved from 0.653 to 0.738 and for TPF = 0.95, specificity improved from 0.536 to 0.612). The values of TPF and FPF were extracted directly from the reconstructed ROC curves.

The standard deviations shown in table 1 represent the variability introduced by the stochastic nature of the GA as well as the uncertainly of the ROC estimation associated with different data splits. To evaluate the variability introduced by the GA alone, for selected splits the GA optimization experiments were repeated 20 times per split. The observed ROC performance per split was extremely robust showing a standard deviation less than 0.002 for AUC and less than 0.02 for $_{0.9}$AUC.

Figure 5 shows a typical histogram of the importance weights assigned to examples when the system is optimized with AUC as the fitness function and weights vary in the range [0, 1]. The weight distribution is nearly uniform which means almost equal number of images with high, moderate and low clinical importance. The distributions were similar when the system was optimized with $_{0.9}$AUC as a fitness function.

As described in the previous section, the optimization process was repeated once for each of the ten splits of the data. Such data handling allowed for analysis of consistency in weights assigned to each particular image. Note that each image appeared in nine out of ten development datasets. It was observed that the weights assigned to a particular image can vary significantly including cases when a particular example has high importance (a weight close to 1) for one split and marginal importance (a weight close to 0) for another. This finding can be potentially explained by the fact that there are groups of similar examples. If a particular example from such a group is assigned a large weight (i.e., high importance), the other examples from the same group may be assigned a small weight since they have nothing new to contribute. When the experiment is repeated with another data split, it is possible that a different example from the same group will emerge as the representative one assigned a large weight. Thus, it is possible to derive different representations in each optimization trial. In terms of the optimization problem at hand, this finding suggests that there are multiple equally

good solutions to the problem often corresponding to very different weights. Finally, GA is not guaranteed to find an optimal solution for a given problem. Thus, for certain algorithm runs, non-optimal weights are likely to occur, further increasing variability of weights for particular images among runs.

7. Discussion

In this study, a new optimization framework for case-based CAD systems was presented. The study hypothesis was that each image stored in the knowledge database of CB-CAD system has different importance in the diagnostic process. A new decision index was proposed that incorporates different importance weights for each stored example. Then, a genetic algorithm approach was employed to find an optimal set of weights according to clinically relevant objectives. It was shown experimentally that the modified CAD system performs significantly better as evaluated by the area under the ROC curve and partial area under the ROC curve indices. As a result, the optimized CAD system is characterized by a higher mass detection rate for a given specificity as well as a reduced false positive rate for a given sensitivity.

The technique presented in this paper has some similarities to the importance evaluation based on the order of retrieval (Tourassi *et al* 2003). A significant difference, however, must be stressed. In the approach based on the order of retrieval, the importance weights of patterns in the database are query dependent, while in the approach presented here they are universal for all incoming queries. Since in the presented approach the vector of importance weights is not query dependent, the technique could also serve as the foundation for selecting which cases should be stored in the knowledge database and which could be eliminated. Experiments are underway to confirm the potential as a technique of selecting patterns for building knowledge databases.

It should be noted that there is one limitation to the proposed optimization technique. Each time the knowledge database is modified (e.g. by adding new examples), the same optimizations process needs to be repeated. The optimization process, however, is run offline and does not affect the response time of the constructed system.

Although the proposed GA approach was demonstrated with respect to CAD system for a breast cancer detection, it is certainly applicable to any type of case-based CAD system (feature-based and featureless). As CB-CAD systems become more popular for clinical applications, sophisticated techniques are expected to emerge for more effective construction and use of knowledge databases. The proposed technique is a step in this direction.

Acknowledgments

The work of Georgia D Tourassi was supported by grant R01 CA101911 from the National Cancer Institute. The authors would like to thank Dr Mehmet K Muezzinoglu for his helpful comments, Katie Todd for her help in preparation of this paper and two anonymous reviewers for their valuable suggestions.

References

Aha D W, Kibler D and Albert M K 1991 Instance-based learning algorithms *Mach. Learn.* **6** 37–66

Anastasio M A, Yoshida H, Nagel R, Nishikawa R M and Doi K 1998 A genetic algorithm-based method for optimizing the performance of a computer-aided diagnosis scheme for detection of clustered microcalcifications in mammograms *Med. Phys.* **25** 1613–20

Bertsekas D P 1999 *Nonlinear Programming* 2nd edn (Cambridge, MA: Athena Scientific)
Bevilacqua A, Campanini R and Lanconelli N 2001 A distributed genetic algorithm for parameters optimization to detect microcalcifications in digital mammograms *Lect. Notes Comput. Sci.* **2037** 278–87
Bradley A P 1997 The use of the area under the ROC curve in the evaluation of machine learning algorithms *Pattern Recognit.* **30** 1145–59
Boroczky L, Zhao L and Lee K P 2006 Feature subset selection for improving the performance of false positive reduction in lung nodule CAD *IEEE Trans. Inform. Technol. Biomed.* **10** 504–11
Campanini R and Lanceonelli N 2006 Genetic algorithms in CAD mammography *Recent Advances in Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer* pp 129–57
Chang Y-H, Hardesty L A, Hakim C M, Chang T S, Zheng B, Good W F and Gur D 2001 Knowledge-based computer-aided detection of masses on digitized mammograms: preliminary assessment *Med. Phys.* **28** 455–61
Cover T 1991 *Elements of Information Theory* (New York: Wiley)
Doi K 2006 Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology *Phys. Med. Biol.* **51** R5–27
Duda R O, Hart P E and Stork D G 2000 *Pattern Classification* (New York: Wiley-Interscience)
Efron B and Tibshirani R J 1993 *An Introduction to the Bootstrap* (London: Chapman and Hall)
Eiben A E and Smith J E 2003 *Introduction to Evolutionary Computing* (Berlin: Springer)
El-Naqa I, Yang Y, Galatsanos N P and Wernick M N 2004a Mammogram retrieval based on incremental learning *IEEE/NIH International Symposium on Biomedical Imaging* pp 1163–6
El-Naqa I, Yang Y, Nishikawa R N and Wernick M N 2004b A similarity learning approach to content based image retrieval: application to digital mammography *IEEE Trans. Med. Imaging* **23** 1233–44
Fawcett T 2006 An introduction to ROC analysis *Pattern Recognit. Lett.* **27** 861–74
Fogel D B, Wasson E C III, Boughton E M and Porto V W 1998 Evolving artificial neural networks for screening features from mammograms *Artif. Intell. Med.* **14** 317–26
Friedman C P, Elstein A S, Wolf F M, Murphy G C, Franz T M, Heckerling P S, Fine P L, Miller T M and Abraham V 1999 Enhancement of clinicians diagnostic reasoning by computer-based consultation: a multisite study of 2 systems *JAMA* **282** 1851–6
Gen M and Cheng R 1997 *Genetic Algorithms and Engineering Design* (New York/Piscataway, NJ: Wiley/IEEE)
Gurcan M N, Chan H-P, Sahiner B, Hadjiiski L, Petrick N and Helvie M A 2002 Optimal neural network architecture selection: improvement in computerized detection of microcalcifications *Accad. Radiol.* **9** 420–9
Heath M *et al* 1998 Current status of the digital database for screening mammography *Digital Mammography* (Dordrecht/New York: Kluwer/Academic)
Kawamoto K, Houlihan C A, Balas E A and Lobach D F 2005 Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success *Br. Med. J.* **330** 765–72
Maes F, Collignon A, Vandermeulen D, Marchal G and Suetens P 1997 Multimodality image registration by maximization of mutual information *IEEE Trans. Med. Imaging* **16** 187–98
Metz C E, Herman B A and Roe C A 1998a Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets *Med. Decis. Making* **18** 110–21
Metz C E, Herman B A and Shen J-H 1998b Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data *Stat. Med.* **17** 1033–53
Michalewicz Z 1999 *Genetic Algorithms + Data Structures = Evolutionary Programs* (Berlin: Springer)
Mitchell T 1997 *Machine Learning* (New York: McGraw-Hill)
Obuchowski N A 2003 Receiver operating characteristic curves and their use in radiology *Radiology* **229** 3–8
Pena-Reyes C A and Sipper M 1999 A fuzzy-genetic approach to breast cancer diagnosis *Artif. Intell. Med.* **17** 131–55
Pena-Reyes C A and Sipper M 2000 Evolutionary computation in medicine: an overview *Artif. Intell. Med.* **19** 1–23
Peng Y, Yao B and Jiang J 2006 Knowledge-discovery incorporated evolutionary search for microcalcification detection in breast cancer diagnosis *Artif. Intell. Med.* **37** 43–53
Sahiner B, Chan H-P, Petrick N, Helvie M A and Goodsitt M M 1998 Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis *Phys. Med. Biol.* **43** 2853–71
Sahiner B, Chan H-P, Wei D, Petrick N, Helvie M A, Adler D D and Goodsitt M M 1996 Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue *Med. Phys.* **23** 1671–84
Sampat M P, Markey M K and Bovik A C 2005 Computer-aided detection and diagnosis in mammography *Handbook of Image and Video Processing* pp 1195–217
Schmidt R, Montani S, Bellazzi R, Portinale L and Gierl L 2001 Cased-based reasoning for medical knowledge-based systems *Int. J. Med. Inform.* **64** 355–67
Smeulders A W M, Worring M, Santini S, Gupta A and Jain R 2000 Content-based image retrieval at the end of the early years *IEEE Trans. Patterns Anal. Mach. Intell.* **22** 1349–80

Tourassi G D, Frederick E D, Markey M K and Floyd C E Jr 2001 Application of the mutual information criterion for feature selection in computer-aided diagnosis *Med. Phys.* **28** 2394–402

Tourassi G D, Haarawood B, Singh S, Lo J Y and Floyd C E 2007 Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms *Med. Phys.* **34** 140–50

Tourassi G D, Vargas-Voracek R and Floyd C E Jr 2003 Content-based image retrieval as a computer aid for the detection of mammographic masses *Proc. SPIE* **5032** 590–7

Tourassi G D, Vargas-Voracek R, Catarious D M Jr and Floyd C E Jr 2003 Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information *Med. Phys.* **30** 2123–30

van Ginneken B, ter Haar Romeny B M and Viergever M A 2001 Computer-aided diagnosis in chest radiography: a survey *IEEE Trans. Med. Imaging* **20** 1228–41

Zhang G P 2000 Neural networks for classification: a survey *IEEE Trans. Syst. Man. Cybern.* C **30** 451–62

Zurada J M 1992 *Introduction to Artificial Neural Systems* (New York: West Publishing Company)