Fissures Segmentation Using Surface Features:

Content-based Retrieval for Mammographic Mass Using Ensemble Classifier

Hong Liu, PhD, Yihua Lan, PhD, Xiangyang Xu, PhD, Enmin Song, PhD, Chih-Cheng Hung, PhD

Rationale and Objectives: Accurate classification is critical in mammography computer-aided diagnosis using content-based image retrieval approaches (CBIR CAD). The objectives of this study were to: 1) develop an accurate ensemble classifier based on domain knowledge and a robust feature selection method for CBIR CAD; 2) propose three new features; and 3) assess the performance of the proposed method and new features by using a relatively large imaging data set.

Materials and Methods: The data set used in this study consisted of 2114 regions of interest (ROI) extracted from a publicly available image database. The proposed ensemble classifier method we called E-DGA-KNN included four steps. In the first step, 804 ROIs depict masses were divided into five classes according to their boundary types. Then, each class of ROI with an equal number of negative ROIs were put together to create a sub-database. Second, a dual-stage genetic algorithm, which was called DGA, was applied on those five sub-databases for feature selection and weights determination respectively. In the third step, five base K-nearest neighbor (KNN) classifiers were created by using the results of the second step on 2114 ROIs, and five detection scores for a given queried ROI were obtained. Finally, these classifiers are combined to yield a final classification. The performances of the proposed methods were evaluated by using receiver operating characteristic (ROC) analysis. A comparison with eight different methods on the data set was provided which include the stepwise linear discriminative analysis algorithm (SLDA) and particle swarm optimization (PSO) algorithm with KNN classifier.

Results: When four hybrid feature selection methods were applied with single KNN classifier (ie, DGA-KNN, SLDA-WGA-KNN, SLDA-PSO-KNN, GA-PSO-KNN) and the proposed E-DGA-KNN method to the data set, the computed areas under the ROC curve (Az) were 0.8782 ± 0.0080, 0.8675 ± 0.0081, 0.8623 ± 0.0083, 0.8725 ± 0.0079, and 0.8927 ± 0.0073, respectively. If all features and single KNN classifier were used, the Az value was 0.8478 ± 0.0088. Az values were 0.8592 ± 0.0083 and 0.8632 ± 0.0081 when SLDA or GA algorithm used alone.

Conclusions: In this study, an ensemble classifier based on domain knowledge and a dual-stage feature selection method was proposed. Evaluation results indicated that the proposed method achieved largest value of ROC compared to other algorithms. The proposed method shows better performance and has the potential to improve the performance of CBIR CAD in interpreting and analyzing mammograms.

Key Words: Computer-aided detection and diagnosis; mammography; content-based image retrieval; feature selection; stepwise linear discriminative analysis; genetic algorithm; domain knowledge; K-nearest neighbor.

Breast cancer is still one of the most common classes of cancer among women all over the world (1,2). Early detection of breast cancer plays a major role in reducing mortality. Mammography is widely regarded as the most effective tool for breast cancer early detection and diagnosis available today (3). However, mammogram interpretation is a difficult and error-prone task. To aid radiologists in reading mammograms, many researchers have developed computer-aided diagnosis (CAD) systems (4,5). At present, researches are focusing on developing mammography CAD scheme using content-based image retrieval approaches (CBIR CAD) (6–8).

The goal of CBIR CAD scheme is to detect whether and what degree the queried region of interest (ROI) similar to breast masses. In other words, the CBIR CAD scheme will retrieve K reference ROIs most similar to the queried ROI. It gives visual aids and a number of relative information to radiologists. For the time being, there are two typical types of CBIR CAD that have been developed in mammography: multifeature-based K-nearest neighbor (KNN) methods (7,8) and template matching–based methods (9,10). In this study, we focus on the former, which yields significantly higher performance than the latter that was demonstrated by Wang et al (11).

Acad Radiol 2011; 18:1475–1484

School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, P.R. China (H.L., Y.L., X.X., E.S.); Key Laboratory of Education Ministry for Image Processing and Intelligent Control, Wuhan, Hubei, P.R. China (H.L., Y.L., X.X., E.S.); School of Computer Engineering, Huaihai Institute of Technology, Lianyungang, Jiangsu, P.R. China (Y.L.); School of Computing and Software Engineering, Southern Polytechnic State University, Marietta, GA (C.-C.H.). Received June 10, 2011; accepted August 23, 2011. Supported in part by the China International Science and Technology Cooperation Project (Grant No. 2009DFA12290), and Huaihai Institute of Technology Natural Science Foundation (Grant No. Z2009013). **Address correspondence to:** X.X. e-mail: xuxy@mail.hust.edu.cn

doi:10.1016/j.acra.2011.08.012

In a multifeature KNN classifier–based CBIR CAD scheme, the need for the reduction of the number of features for CAD schemes is well recognized (12). In mammography CAD systems, genetic algorithms (GA) and stepwise linear discriminant analysis (SLDA) are two commonly feature selection methods (13–16).

For complex classification problems, the use of multiple classifier systems usually leads to improvement of classification performance (17). A powerful multiple classifier technique is the use of classifier ensembles, alternately known as ensembles of classifiers, committees of classifiers, or multiple classifier systems (18), which are a learning paradigm in which several base classifiers are jointly used to classify an unknown data point independently, and the output results of these classifiers are combined to create an ensemble final output. Methods for creating base classifiers are usually introduced through different classification methods, different training data sets, or different feature sets (19). Three feature spaces (ie, the original features, principal component analysis [PCA], and domain feature space) are used in a random subspace method (RSM) for training the ensemble of different classifiers (17). Lee et al applied the RSM on the features selected by GA to create a set of diverse classifiers (18). Yu et al proposed ensemble based on feature selection, which selects those GA-based classifiers with the best validation performance (19). The combination method by using majority voting was applied in the previous three literatures.

Generally, according to the real-time querying requirement of CBIR CAD, the number of base classifiers to create an ensemble classifier should not be too large and should have diversity. This study is focused on improving diversity and accuracy of base classifiers in an ensemble classifier by using a new dual-stage feature selection method called dual-stages GA (DGA). We found that it will be more precise and reasonable for representing each class of mass by specific feature subset and weights. Motivated by this, we constructed five subdatabases by using domain knowledge. Then the DGA method was applied to each subset of data, and several corresponding feature subsets and weights could be obtained. While using these results, five base classifiers were created on the entire database. At last, these classifiers were combined into an ensemble classifier by an ensemble learning technique. We tested the proposed method on a publicly available database, and experimental results demonstrated the effectiveness of the method.

MATERIALS AND METHODS

Overview

We present an ensemble classifier based on domain knowledge and a dual-stage feature selection method. The proposed CBIR CAD scheme followed a general CBIR framework, which has been preliminarily researched in our previous study (20). In this scheme, there are several critical steps: mass region segmentation, feature extraction, feature selection, and classification with ensemble classifiers. The performance of the proposed method is tested and evaluated using a leave-one-out sampling scheme (21).

In experiments, to a given reference image database of m (eg, 2114) ROIs, each of m ROIs in our reference database was separately used once as a queried ROI while the other remaining m-1 ROIs were used to establish a new reference database. Then the CBIR searched for the K ROIs that were considered the most similar to the queried ROI in the new database, and used a decision algorithm that effectively combines the similarity indices and known truth of the retrieved images to calculate a decision index (ie, the likelihood score of the suspicious queried image being a positive region). In the ensemble classifier approach, five different base classifiers are used separately to obtain five decision indexes and then weighted to a detection score. On the basis of the detection scores for both true-positive and false-positive ROIs, we applied the receiver operator characteristic (ROC) data-fitting and analysis program ROCKIT (Charles E. Metz, Department of Radiology, The University of Chicago Medical Center, Chicago, IL. Available at: http://metz-roc.uchicago.edu/.) to compute the areas under the curves (Az) and 95% confidence intervals to assess the performance of the CBIR CAD scheme (22).

Reference-image Database

The mammograms used in this study were selected from a publicly available database, the Digital Database for Screening Mammography (DDSM) (23). In this study, a reference-image database includes 2114 ROIs extracted from DDSM using following three steps.

Step 1: Uncompressed the original mammograms by an uncompress software (24).

Step 2: Subsampled the images by a factor of eight, which means a pixel size is increased from 50 × 50 to 400 × 400 μm, and the range of image pixel gray level was compressed from 12 to 8 bits (from 4096 to 256 gray levels).

Step 3: Selected ROIs with a fixed size of 125 × 125 pixels. In this step, all ROIs in reference-image database were selected. Our reference-image database consisted of two classes of ROIs: one depicts biopsy-proven mass regions and the other does not. In DDSM, each image depicting abnormal mass has a corresponding overlay file that records detailed information of each mass region, including pathologic classification (malignant or benign) and mass boundaries. To regularly extract an ROI for a mass region, we first manually marked a point (pixel) near the subjectively estimated geometric center of the mass. Then, using this pixel as the center of an ROI and then extracting the ROI included 125 × 125 pixels (ie, 50 × 50 mm). Because the original outlines of lesions provided by the DDSM are not good enough for detection (25) and the features computation is based on mass contours, we applied an automatic segmentation algorithm (26) from our previous study to extract the masses contour. To acquire a set of precise segmentation results and to compute the features more