BOOSTED CLASSIFICATION OF BREAST CANCER BY RETRIEVAL OF CASES HAVING SIMILAR DISEASE LIKELIHOOD

Juan Wang and *Yongyi Yang*

Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616

ABSTRACT

In diagnostic imaging, recent studies have shown that retrieval of cases that are similar to the case being evaluated can boost its classification performance. In this work we investigate how to improve the utility of the retrieved cases by considering the similarity both in the image features and in the pathology when comparing the cases. To demonstrate the benefit of this retrieval strategy, we propose a boosted Adaboost classifier which can be adapted to the retrieved cases at a low computational cost. The proposed approach was tested on a set of 981 mammogram cases (449 malignant, 532 benign). The results show that the retrieval-boosted Adaboost classifier can significantly outperform its baseline counterpart, and that inclusion of pathology information (measured by the likelihood of malignancy) in the retrieval can further improve the classification accuracy.

Index Terms— Image retrieval, computer-aided diagnosis (CAD), classification, Adaboost

1. INTRODUCTION

One important early sign of breast cancer in women is the appearance of clustered microcalcifications (MCs) in mammograms [1]. MCs are tiny calcium deposits that exhibit as small bright spots in a mammogram image (e.g. Fig. 1). They can be found in both malignant and benign cases in screening mammography [2]. However, due to their subtlety in appearance in mammogram images, accurate diagnosis of MC lesions as benign or malignant is a very challenging clinical task for radiologists. In the literature, there have been significant efforts in development of computer-aided diagnosis (CADx) methods for differentiating between malignant and benign MC lesions [3-4].

In recent years, content-based image retrieval (CBIR) is increasingly explored as a CADx tool in diagnostic imaging [e.g. 5-7]. The goal of a CBIR system is to provide radiologists with examples of lesions (retrieved from a reference library of known cases) that are similar to the one being evaluated. Such an approach has been studied for different lesion types and imaging modalities [5-7].

Motived by the concept of CBIR, we have been developing a case-adaptive approach to CADx, in which the classification of a query case is boosted by a set of similar cases [8-9]. This is different from the traditional approach in CADx, where a pattern classifier is first trained on a set of existing cases (called training samples), and subsequently applied for classification of future (unknown) cases. In our adaptive approach, for a case to be classified (i.e. query), we will first retrieve a set of known cases that are similar to the query (from an existing library), and then use these retrieved cases to adapt the classifier based on its classification accuracy on these similar cases. In essence, the retrieved cases are used to refine the decision function of the classifier in the local neighborhood around the query.

Built on the previous studies [8-9], in this work we investigate how to improve the retrieval of similar cases such that their benefit on boosting the classification of a query is maximized. We conjecture that truly similar cases should be similar not only in their image features but also in pathology. Based on this conjecture, we devise a new retrieval strategy for similar cases, which will take into account both the image features and the pathology information of the cases.

Furthermore, a potential drawback of using retrieval-boosted classification is that the classifier needs to be retrained for each query based on the retrieved cases, which will inevitably increase the computational burden. This issue becomes particularly severe for a non-linear classifier such as SVM [8]. To address this issue, in this study we will develop our retrieval-boosted approach based on the widely used Adaboost, for which the added computational cost for updating the adaptive classifier can be rather low when decision stumps are used.

2. METHODS

Our retrieval assisted approach for boosting diagnosis can be stated as follows: for a given query case **x** under consideration, we first obtain a set of similar cases with known pathology from a reference library; then we make use of these retrieved cases to improve the classification accuracy of an existing baseline classifier on **x**.

2.1 Similarity in pathology for retrieval

In boosting the performance of a classifier, the retrieved cases play the role of refining the classifier function for a given query. Thus, how to retrieve these similar cases is essential to the classification on the query case. In our

This work was supported by NIH/NIBIB grant R01EB009905.

previous study [9], the Euclidean distance was used for retrieving cases with image features similar to the query, the purpose of which was to improve the accuracy of the classifier in the local neighborhood around the query.

In this study, we investigate how to improve the strategy for retrieving similar cases in order to maximize their benefit on boosting the classification on the query. While a simple metric such as the Euclidean distance is effective for retrieving cases that have similar quantitative image features, we find that cases having similar image features can often be different in pathology. This is because for a particular case not all the features can be discriminative between cancer and benign classes. This is also the reason that the problem of classification of MC lesions is fundamentally challenging.

Given that cases with different pathology can have similar image features, we conjecture that those cases that are truly similar to the query should not only be similar in image features but also in pathology, and thus can be more useful for boosting the classifier on the query. Therefore, it would be desirable to retrieve those cases that have the same pathology as the query. Of course, this is impractical because our very purpose is to determine the pathology of the query (which is unknown).

To deal with this difficulty, we will consider the following alternative approach in this study: instead of the true pathology, we first use a pre-trained classifier to estimate the likelihood of malignancy of the query case, and then use this estimated likelihood to retrieve similar cases for boosting the classifier on the query. That is, we seek to retrieve cases which are similar to the query not only in terms of their image features, but also in terms of their predicted disease likelihood.

Based on the above consideration, in our experiments, we devised the following retrieval procedure for a given query **x**: first, compute the likelihood of malignancy of **x** by a pre-trained classifier, denoted by $g(\mathbf{x})$; second, retrieve $2N_r$ cases that are closest to $g(\mathbf{x})$ according to their predicted malignancy likelihood from the same classifier; third, select from these cases the top N_r cases based on their Euclidean distance to the query **x**. The selected cases are used subsequently for boosting the classification on the query.

2.2 Retrieval-boosted classification with Adaboost

To demonstrate the use of our retrieval strategy for boosting the classification performance, in this study we consider the Adaboost classifier [10] owing to its computational advantage, as described subsequently.

Adaboost is a boosting learning algorithm to form a committee-based decision function. Mathematically, the classifier function from Adaboost can be written as:

$$f(\mathbf{x}) = \sum_{k=1}^{M} \alpha_k f_k(\mathbf{x}) \qquad (1)$$

where $f_k(\mathbf{x}), k = 1,\cdots,M$, are a sequence of weak learners determined from training (whose individual performance is only slightly better than random guessing), and M is the number of them. In (1), the coefficients $\alpha_k, k = 1,\cdots,M$, are weighting factors determined based on the accuracy of their corresponding classifiers $f_k(\mathbf{x})$ on the training samples (a higher weight is assigned for a classifier that is more accurate) [10]. Moreover, the individual classifiers $f_k(\mathbf{x})$ are trained in a sequential manner such that subsequent classifiers are trained with more emphasis on those training samples that have been misclassified by their predecessors [10].

Now consider a query MC lesion $\mathbf{x}$. Let $\left\{\left(\mathbf{x}^{(r)}, y^{(r)}\right), r = 1,2,\cdots,N_r\right\}$ be a set of N_r retrieved cases for $\mathbf{x}$. We propose to further adapt the committee classifier in (1) based on the accuracy of the individual weak classifiers on these retrieved cases. The idea is to place a higher weight on those weak classifiers that are more accurate on the retrieved cases, so as to improve the accuracy of the committee classifier in the local neighborhood of the query.

Specifically, the classifier function in (1) is adapted as follows:

$$f(\mathbf{x}) = \sum_{k=1}^{M} \gamma_k \alpha_k f_k(\mathbf{x}) \Big/ \sum_{k=1}^{M} \gamma_k \qquad (2)$$

where $\gamma_k, k = 1,\cdots,M$ are penalty factors introduced according to the accuracy of the weak classifiers $f_k(\mathbf{x})$ on the retrieved cases, which is defined below.

To quantify the accuracy of $f_k(\mathbf{x})$ on the set of retrieved cases, we use a weighted error which is defined as

$$e_k = \sum_{r=1}^{N_r} \beta^{(r)} I\left(y^{(r)} \neq f_k\left(\mathbf{x}^{(r)}\right)\right) \Big/ \sum_{r=1}^{N_r} \beta^{(r)} \qquad (3)$$

where $I(\cdot)$ is the indicator function, and the factors $\beta^{(r)}$ are used to place more emphasis on cases that are more similar to the query $\mathbf{x}$.

Afterward, the penalty factors in (2) are calculated as

$$\gamma_k = \exp\left(-e_k \big/ \lambda\right) \qquad (4)$$

where λ is a parameter controlling the sensitivity to the error term. In our experiment, λ was set at 0.5. At this setting, the value of γ_k varies from 1 (when $e_k = 0$; no penalty) to 0.1353 (when $e_k = 1$; worst case).

For the similarity factors $\beta^{(r)}$ in (3), in our experiments we used the Gaussian kernel function as in [9] to measure the similarity between a retrieved case and the query:

$$\beta^{(r)} = \exp\left(-\frac{\| \mathbf{x}^{(r)} - \mathbf{x} \|^2}{2\sigma^2}\right) \qquad (5)$$

where σ is a scaling factor (which was set as the 10^{th} percentile of the pair-wise distances of all the cases in the training set [9]).

In this study, we use the decision stumps for the weak classifiers $f_k(\mathbf{x})$, which are commonly used for Adaboost

due to its simplicity [11]. Specifically, a decision stump has the following form:

$$f_k(\mathbf{x}) = \begin{cases} 1 & x_{(k)} \geq T_{(k)} \\ -1 & x_{(k)} < T_{(k)} \end{cases} \quad (6)$$

where $x_{(k)}$ is the associated decision feature and $T_{(k)}$ is the decision threshold, both of which are determined during training.

It can be readily seen that the computational complexity of the boosted classifier in (2) is mostly associated with the calculation of the error terms $e_k, k = 1,2,\cdots,M$. With decision stumps as the weak classifiers, this takes only $2MN_r$ comparison operations, $M(N_r+1)$ multiplication operations, and $M(N_r-1)$ summation operations. Thus, the retrieval-boosted Adaboost can be updated with very low complexity.

3. EXPERIMENTS AND RESULTS

3.1 Mammogram dataset

In this study, we made use of a large set of mammogram cases collected from two different datasets. The first dataset was collected by the Department of Radiology at the University of Chicago, which includes 333 cases (161 malignant, 172 benign). The second was from the DDSM dataset maintained at the University of South Florida, which includes 648 cases (288 malignant, 360 benign). Altogether, there are a total of 981 cases (449 malignant, 532 benign) in the collection, all containing MC lesions. All of the mammogram images were digitized with a spatial resolution of 0.1 mm/pixel. The use of a large number of cases is to ensure that there are enough cases available for retrieval.

3.2 Experiment implementation

In our implementation, the dataset was randomly divided into three subsets, denoted by A, B and C, respectively. Subsets A and B have 200 cases each (100 malignant, 100 benign), and subset C has the remaining 581 cases (249 malignant, 332 benign). Subset A was used for training and optimizing the baseline classifier, and subset B was used exclusively for testing the classification performance. Subset C was used together with A as the reference library of known cases.

To quantify the MC lesions in the dataset, the individual MCs in each marked lesion was first detected by an SVM detector [12] with bi-thresholding scheme [13]; afterward, a set of descriptive features (consisting of 11 cluster features and 40 MC features [3, 9, 14-16]) was extracted to characterize the clustered MCs within each lesion. To determine the most salient features for discriminating between cancer and benign lesions, we applied a sequential forward selection procedure with logistic regression using the cases in subset A. This resulted in the following nine features: (1) scatterness of cluster [14], (2) eccentricity of cluster [14] (3) Fourier descriptor II of cluster, (4) moment feature $F_3' - F_1'$ of cluster [15], (5) rotation invariant moment I_1 of cluster [16], (6) the mean area of MCs in the cluster, (7) mean of the effective thickness of MCs in the cluster [3], (8) mean of the scatterness of MCs in the cluster, and (9) the standard deviation of the scatterness of MCs in the cluster.

Afterward, the baseline Adaboost classifier was trained with the cases in subset A. To determine the number of weak classifiers M, we applied a four-fold cross-validation procedure, based on which the optimal $M = 36$ was chosen. This trained Adaboost classifier was used for subsequent boosting with our proposed retrieval strategy. Moreover, it was also used for predicting the likelihood of malignancy $g(\mathbf{x})$ needed in the retrieval procedure.

Considering that the boosted classifier made use of additional cases (retrieved from C) for training, for comparison purpose, we also tested the performance of the Adaboost classifier by training with all the cases in both subsets A and C. This represents the optimal performance that could be achieved by the baseline classifier when all the available cases (except the test cases) were used for training.

To evaluate the performance of the classifiers, we conducted a receiver operating characteristic (ROC) analysis, which is now routinely used for performance evaluation in a classification task. The area under the ROC curve, denoted by AUC, is used to summarize the diagnostic performance. A larger AUC means better performance by a classifier.

To suppress the effect of case distribution in the test set, we evaluated the classification performance for 20 different random splits of subsets B and C, and the average AUC values are obtained from these 20 splits.

3.3 Results

In Fig. 2 we show the classification performance results (measured by AUC) achieved by the Adaboost classifier when boosted with the proposed strategy ("Boosted–pathology") when the number of retrieved cases N_r is varied from 6 to 500. For comparison, we also show in Fig. 2 the performance results achieved by the boosted classifier when the Euclidean distance was used for retrieval ("Boosted–Euclidean"). In addition, the perforamnce results are also shown for the baseline classifier, i.e., without boosting ("Baseline"), and the Adaboost classifier trained with all the avaiable cases in subsets A and C ("Adaboost-all"). As can be seen, the boosted classifiers (Boosted–pathology and Boosted–Euclidean) achieved noticeably higher AUC values than their baseline counterpart. Furthermore, the use of likelihood malignancy in the retrieval (Boosted–pathology) achieved further improvement in classification performance when compared with the use of the Euclidean distance.

From Fig. 2, the best performance is achieved by Boosted–pathology with $Nr = 20$, for which AUC=0.7490. However, as the number of retrieved cases is further increased, the performance of the boosted classifiers

(Boosted–pathology and Boosted–Euclidean) starts to degrade. This is due to the fact that the number of similar cases for a given query is limited in the dataset and further retrieval of additional cases becomes no longer beneficial.

Finally, it is interesting to note in Fig. 2 that the boosted classifiers even achieved improvement over the Adaboost classifier trained all the cases in both subsets A and C. This indicates the improvement by the boosted classifiers was attributed to the boosting strategy in refining the baseline classifier.

4. CONCLUSION

In this study we investigated how to use CBIR to boost the classification performance of a classifier by considering the similarity both in the image features and in the pathology for retrieving cases. We demonstrated the benefit of the proposed retrieval strategy with a boosted Adaboost classifier owing to its low computational cost. The evaluation results show that the retrieval-boosted approach can significantly outperform its baseline classifier and that inclusion of pathology information in the retrieval can further improve the classification accuracy.

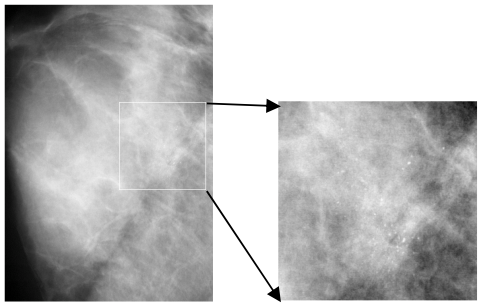

Fig. 1 A mammogram (left) and a magnified view of a lesions with clustered MCs (right).

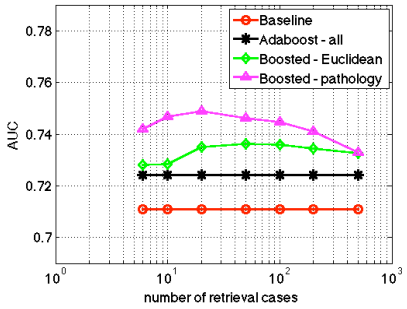

Fig. 2 Classification performance (AUC) achieved by the Adaboost classifier boosted with the proposed retrieval strategy ("Boosted–pathology"), and with retrieval using the Euclidean distanace ("Boosted–Euclidean"). For comparison, the results are also shown for the basedline Adaboost classifier ("Baseline"), and the Adabssot classifier trained with all the avaialble cases ("Adaboost-all").

5. REFERENCES

[1] American Cancer Society, "Cancer facts and figures", Atlanta, GA, 2012.

[2] W. J. H. Veldkamp and N. Karssemeijer, "Automated classification of clustered microcalcifications into malignant and benign types," *Medical Physics*, vol. 27, no. 11, 2000.

[3] Y. Jiang, R. M. Nishikawa, *et al*, "Malignant and benign clustered microcalcifications: automated feature analysis and classification," *Radiology*, vol. 198, no. 3, 1996.

[4] L. Wei, Y. Yang, and R. M. Nishikawa, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE Transaction on Medical Imaging*, vol. 24, no. 3, 2005.

[5] R. Nakayama, H. Abe, J. Shiraishi, and K. Doi, "Potential usefulness of similar images in the differential diagnosis of clustered microcalcifications on mammograms," *Radiology*, vol. 253, no. 3, pp. 625-631, 2009.

[6] G. D. Tourassi, B. Harrawood, S. Singh, L. Y. Lo, and C. E. Floyd, "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms," *Medical Physics*, vol. 34, no. 1, pp. 140-150, 2007.

[7] A. M. Aisen, L. S. Broderick, H. Winer-Muram, C. E. Brodley, A. C. Kak, C. Pavlopoulou, J. Dy, C. Shyu, and A. Marchiori, "Automated storage and retrieval of thin-section CT images to assist diagnosis: system description and preliminary assessment," *Radiology*, vol. 228, no. 1, pp. 265-270, 2003.

[8] L. Wei, Y. Yang, and R. M. Nishikawa, "Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis," *Pattern Recognition*, vol. 42, no. 6, pp. 1126-1132, 2009.

[9] J. Hao and Y. Yang, "Retrieval boosted computer-aided diagnosis of clustered microcalcifications for breast cancer," *Medical Physics*, vol. 39, no. 2, 2012.

[10] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Computational Learning Theory*, vol. 55, no. 1, pp. 119-139, 2001.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE Computer Soceicty Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 1-511, 2001.

[12] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galasanos, and R. M. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Transaction on Medical Imaging*, vol. 21, no. 12, pp. 1552-1563, 2002.

[13] J. Wang, Y. Yang, and R. M. Nishikawa, "Reduction of false positive detection in clustered microcalcifications," *IEEE International Conference on Image Processing (ICIP)*, pp. 1433-1437, 2013.

[14] J. Wang and Y. Yang, "Spatial density modeling for discriminating between benign and malignant microcalcification lesions," *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 133-136, 2013.

[15] L. Shen, R. M. Rangayyan, and J. E. L. Desautels, "Application of shape analysis to mammographic calcifications," *IEEE Transaction on Medical Imaging*, vol. 13, no. 2, pp. 263-274, 1994.

[16] M. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory*, vol. 8, no. 2, pp. 178-187, 1962.