Pattern Recognition 42 (2009) 1126–1132

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis ☆

Liyang Wei [a], Yongyi Yang [a,*], Robert M. Nishikawa [b]

[a] Department of Electrical and Computer Engineering, Illinois Institute of Technology, 3301 South Dearborn Street, Chicago, IL 60616, USA
[b] Department of Radiology, University of Chicago, 5841 South Maryland Avenue, Chicago, IL 60637, USA

ARTICLE INFO

Article history:
Received 23 November 2007
Received in revised form 17 August 2008
Accepted 21 August 2008

Keywords:
Microcalcification classification
Adaptive support vector machine
Image retrieval

ABSTRACT

In this paper, we propose a microcalcification classification scheme, assisted by content-based mammogram retrieval, for breast cancer diagnosis. We recently developed a machine learning approach for mammogram retrieval where the similarity measure between two lesion mammograms was modeled after expert observers. In this work, we investigate how to use retrieved similar cases as references to improve the performance of a numerical classifier. Our rationale is that by adaptively incorporating local proximity information into a classifier, it can help to improve its classification accuracy, thereby leading to an improved "second opinion" to radiologists. Our experimental results on a mammogram database demonstrate that the proposed retrieval-driven approach with an adaptive support vector machine (SVM) could improve the classification performance from 0.78 to 0.82 in terms of the area under the ROC curve.

1. Introduction

Breast cancer remains to be a leading cause of death among women in the developed countries. The American Cancer Society [1] estimates that in 2008 approximately 182,460 women in the US will be diagnosed with invasive breast cancer. About 40,480 women will die from this disease this year. Currently mammography is the dominant method for detection of breast cancer. But it is still far from being perfect. The high sensitivity of screening mammography is compromised by its low specificity to benign lesions, which often appear mammographically similar to malignant lesions. This results in approximately 70% of biopsies performed on benign lesions [2,3].

Clustered microcalcifications (MC) can be an important early sign of breast cancer. As an example, Fig. 1 shows a mammogram with a cluster of MC. They appear as bright spots of calcium deposits. Individual MCs are sometimes difficult to detect because of the surrounding breast tissue, their variation in shape (from granular to rod shapes), orientation, brightness and diameter size. Due to the subtlety in the appearance of individual MCs, there is a significant risk that a radiologist may misclassify some cases in breast cancer diagnosis [4]. It has been reported that 10–30% of lesions are misinterpreted during routine screening of mammograms [5]. In recent years, there have been significant efforts in development of computerized methods for automated classification of MCs. For example, Hamid et al. [6] investigated the performance of four different texture and shape feature extraction methods for classification of benign and malignant MCs. Massimo et al. [7] proposed a multiple-expert approach for classifying MCs wherein the final output was obtained from a combination of multiple experts. In Ref. [4] we investigated several state-of-the-art machine learning algorithms and found that a support vector machine (SVM) classifier could achieve the best performance among several well-known methods for MC classification.

Recently, we developed a content-based mammogram retrieval system as a diagnostic aid to radiologists in their interpretation of mammograms [8]. We conjecture that by presenting perceptually similar mammograms with known pathology to the one being evaluated, the radiologists could reach a better informed decision in their diagnosis. Our proposed mammogram retrieval system involves two major components: (1) retrieving similar mammogram images from a database by using learning based similarity measure and (2) classifying the query mammogram image based on retrieved results (retrieval-driven classification). This retrieval framework is illustrated with a functional diagram in Fig. 2.

In our retrieval system [8], we explored a similarity measure for mammogram retrieval based on supervised learning from expert readers. We evaluated the approach using data collected from an observer study with a set of clinical mammograms. It was demonstrated that the proposed machine learning approach can be used to model the notion of similarity as judged by expert readers in their interpretation of mammogram images and that it can outperform alternative similarity measures derived from unsupervised learning.

☆ This work was supported in part by NIH/NCI Grant CA89668.
* Corresponding author. Tel.: +1 312 567 3423; fax: +1 312 567 8976.
E-mail address: yangyo@iit.edu (Y. Yang).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.08.028

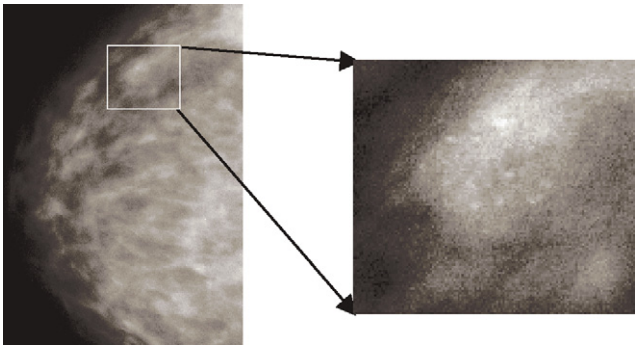

Fig. 1. Left: a mammogram in craniocaudal view. Right: expanded view showing clustered microcalcifications (MCs).

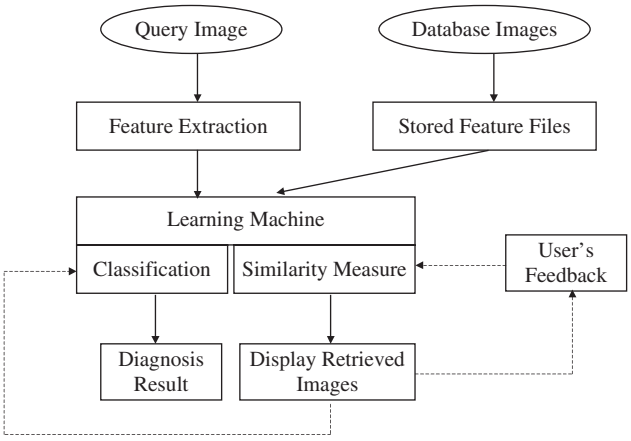

Fig. 2. The proposed content-based mammogram retrieval and classification framework.

In this work, we focus on the second component of our proposed mammogram retrieval and classification system: MC classification assisted by retrieval. The traditional approach in this field is to present the human observer with examples in the database that are similar to the one being examined. In this study, we consider how to use the retrieved similar cases as references to improve a numerical classifier's performance. We conjecture that by adaptively incorporating proximity information to the cost function of a classifier, it can help to improve its classification accuracy, thereby leading to an improved "second opinion" to radiologists. Toward this goal, we propose a retrieval-driven approach with an adaptive SVM (Ada-SVM) for improving the classification performance. We choose SVM since it has been demonstrated to outperform several competing methods in MC classification [4].

2. Methods

2.1. Adaptive SVM

SVM is a constructive learning procedure rooted in statistical learning theory [9]. It is based on the principle of structural risk minimization, which aims at minimizing the bound on the generalization error. The SVM decision function is pre-determined through a training process using a set of examples before it can be applied to data outside the training set. As a result, an SVM tends to perform well when applied to data outside the training set. Indeed, in recent years SVM learning has found a wide range of real-world applications, including handwritten digit recognition [10], object recognition [11], speaker identification [12], face detection in images [13], text categorization [14], etc. In our recent work [15], we developed an SVM based approach for detection of clustered MC in mammograms, and demonstrated using clinical mammogram data that such an approach could outperform several well-established methods in the literature. In Ref. [4], we applied SVM for classification of benign vs. malignant clustered MC.

Despite its success, the performance of an SVM classifier can be hampered by several factors in practice. First, the nature of the problem is often complicated and not well understood, as is the case of breast cancer diagnosis [4], and it is not even clear that the classification task could be well described by a single decision function. Secondly, even when such a decision function indeed exists, the "true" decision boundary is rarely obtainable because of the limited number of available training samples. In such a case, a challenging problem is how to strike a balance between over-fitting and under-fitting in the classifier model. This is especially the case when it is too expensive or simply impossible to obtain enough training samples in many practical problems. Consequently, it becomes impossible to determine the "optimal" classifier function. For example, in Ref. [4] the best classification performance achieved by the SVM was still far from being perfect, which we believe is largely due to the presence of many difficult-to-classify cases in the database we used [4].

In this work, we propose a locally Ada-SVM classification scheme. In the proposed scheme, we attempt to adapt the decision function of the SVM classifier according to how it performs on samples that are close to the one being examined (called query). Specifically, before the SVM decision function is applied to the query, it is first tested and adapted based on the knowledge of the samples that are in its neighborhood. Our motivation is as follows: if the SVM function is found to perform poorly on known samples close to the query, it indicates that the decision function has not been well trained for samples in the neighborhood of the query, and thus, it will also likely not perform well on the query. In such a case, we will adjust the SVM classifier using these similar samples accordingly, which in turn can lead to improved classification accuracy on the query.

In our retrieval-driven classification scheme the SVM classifier is adaptive in that its decision boundary is adjusted according to the "local" information of the case to be classified (i.e., retrieved similar cases). To demonstrate the concept, we show a classification example in Fig. 3. In this binary classification problem, there are 300 samples in each class. Fig. 3(a) shows the decision boundary between the two classes of an SVM (Gaussian RBF kernel, $\sigma = 2.5$, $C = 100$) trained from a subset of 100 training samples; Fig. 3(b) shows the decision boundary obtained using the proposed Ada-SVM classifier (with the same parametric setting as the SVM in (a)). In this example, the five nearest neighbors according to the Euclidean distance were used as the similar cases to the query sample. As can be seen, the proposed Ada-SVM could achieve improved classification over the SVM in this example.

We note that there exist several algorithms related to Ada-SVM classification in the literature which aim to improve a classifier's performance. For example, in Ref. [16] an adaptive-margin SVM was proposed in which the classifier margin was adjusted for each training sample. In our own previous work [17], we used the concept of Ada-SVM in a content-based image retrieval system, where the SVM regression function was adjusted according to the relevance feedback samples provided by the user. To our best knowledge, these methods are quite different from our proposed approach here.

2.2. Algorithm

Consider a general two-class classification problem of assigning a class label $y \in \{-1, +1\}$ to an input feature vector $\mathbf{x} \in R^N$. We are given input–output training data pairs $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.

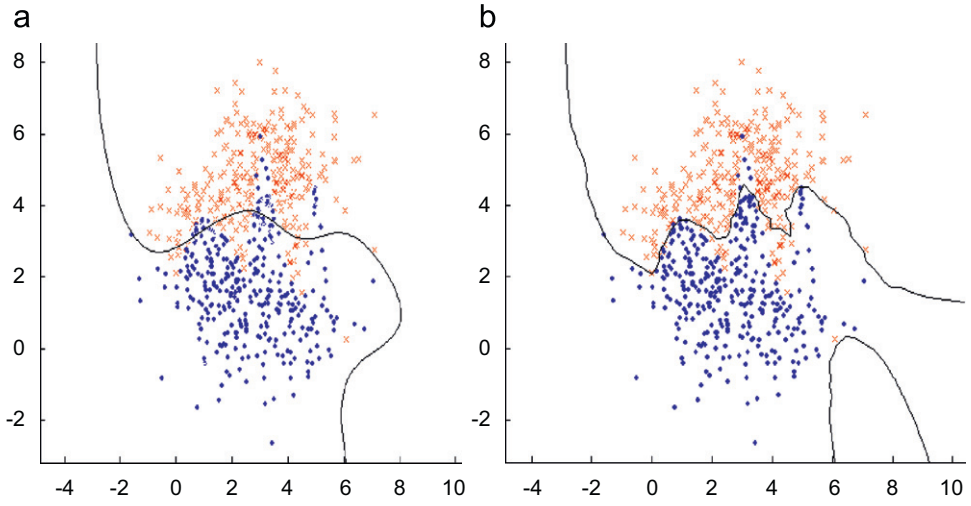

Fig. 3. (a) The classification boundary learned by SVM; (b) the classification boundary learned by Ada-SVM.

In an SVM classifier, the classification function can be written in the following form [9]:

$$f_{SVM}(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}) + b \tag{1}$$

where the parameters $\mathbf{w}$, b are determined from the training data samples. This is accomplished through minimization of the following so-called *structural risk* function:

$$J(\mathbf{w},\xi) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t.}\quad y_i f_{SVM}(\mathbf{x}_i) \geqslant 1-\xi_i$$
$$\xi_i \geqslant 0; \quad i = 1,2,\ldots,N \tag{2}$$

where C is a user-specified, positive parameter, ξ_i are slack variables. The cost function in Eq. (2) constitutes a balance between the empirical risk (i.e., the training errors reflected by the second term) and model complexity (the first term). A larger C corresponds to assigning a higher penalty to the training errors.

In the proposed Ada-SVM, we modify the SVM cost function as follows:

$$\tilde{J}(\mathbf{w},\xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C^{(s)}\sum_{\mathbf{x}_i\in N(\mathbf{x})}\xi_i + C\sum_{\mathbf{x}_i\notin N(\mathbf{x})}\xi_i$$

$$\text{s.t.}\quad y_i f_{SVM}(\mathbf{x}_i) \geqslant 1-\xi_i$$
$$\xi_i \geqslant 0; \quad i = 1,2,\ldots,N \tag{3}$$

where $N(\mathbf{x})$ denotes the set of training samples that are in a defined neighborhood of a query sample $\mathbf{x}$, and $C^{(s)}$ is a penalty parameter introduced for the training samples in $N(\mathbf{x})$.

In the modified cost function $\tilde{J}(\mathbf{w},\xi)$ above, the training samples in $N(\mathbf{x})$ are closer (hence more similar) to the query $\mathbf{x}$ than the others. We write $C^{(s)} = tC$, where $1 < t < \infty$ is a *penalty factor*. This will have the effect of imposing a greater emphasis ($C^{(s)}$) on those samples similar to the query $\mathbf{x}_i$ over other samples. The rationale is that those similar samples should have a greater impact on the classification of the query. Thus, a larger penalty is assessed in the cost function $\tilde{J}(\mathbf{w},\xi)$ when a similar sample is misclassified. Indeed, when $t \to 1$, the Ada-SVM simply becomes a regular SVM where the same factor C is used for all training samples; on the other hand, when $t \to \infty$, the cost function $\tilde{J}(\mathbf{w},\xi)$ will be dominated by the samples similar to the query. In this latter case, the Ada-SVM decision function will depend on only the similar samples. Interestingly, this would be similar in spirit to the well-known k-nearest neighbor (KNN) classifier [18], which makes use of only local neighborhood information in the decision function. In this sense, the Ada-SVM functions can be viewed as a hybrid of a global SVM classifier and a local classifier.

In the SVM cost function, the purpose of using model complexity to regularize the optimization of empirical risk is to avoid overfitting, a situation in which the decision boundary too precisely corresponds to the training data, and thereby may fail to perform well on data outside the training set. As in the choice of the parameter C in regular SVM, the newly introduced penalty factor t in the Ada-SVM will have to be determined during the training phase. This can be determined by using a cross-validation procedure on the set of samples similar to the query, as we discuss later in Section 2.4.

2.3. Insight on the Ada-SVM

Below, we examine how the modified SVM cost function can impact on the SVM decision function. Recall the SVM formulation in Eq. (2), a training sample $(\mathbf{x}_i, y_i)$ is called a support vector when $y_i f_{SVM}(\mathbf{x}_i) \leqslant 1$. More specifically, $(\mathbf{x}_i, y_i)$ is called a margin support vector for $y_i f_{SVM}(\mathbf{x}_i) = 1$, and an error support vector for $y_i f_{SVM}(\mathbf{x}_i) < 1$. In the latter case, the data point $\mathbf{x}_i$ is inside the decision margin, though it may still be correctly classified.

Introducing a so-called kernel function $K(\cdot,\cdot)$, we can rewrite the SVM function $f_{SVM}(\mathbf{x})$ in Eq. (1) as follows:

$$f_{SVM}(\mathbf{x}) = \sum_{i=1}^{N_s}\alpha_i K(\mathbf{x},\mathbf{s}_i) + b \tag{4}$$

where $\mathbf{s}_i, i = 1,2,\ldots,N_s$, denote the *support vectors*. In general, support vectors constitute only a small fraction of the training samples $\mathbf{x}_i, i = 1,2,\ldots,N$. From Eq. (4) it is readily seen that the SVM decision function is formed by only the support vectors. Then, we have the following result:

Proposition 1. *Let* $\mathbf{x}$ *denote a query sample to be classified, and* $N(\mathbf{x})$ *denote a set of training samples in its neighborhood. If none of the samples in* $N(\mathbf{x})$ *are error support vectors in the regular SVM cost function* $J(\mathbf{w},\xi)$*, then the Ada-SVM decision function resulting from the modified cost function* $\tilde{J}(\mathbf{w},\xi)$ *will coincide with the regular SVM.*

To demonstrate the above proposition, consider a training sample $\mathbf{x}_i$ in $N(\mathbf{x})$. Suppose that $\mathbf{x}_i$ is not an error support vector of the regular SVM with $J(\mathbf{w}, \xi)$. By definition, we have $y_i f_{SVM}(\mathbf{x}_i) \geqslant 1$, which implies that the corresponding slack variable $\xi_i = 0$ in the optimal solution of Eq. (2). Thus, the sample $\mathbf{x}_i$ has no contribution to the empirical risk term in $J(\mathbf{w}, \xi)$. Therefore, none of the samples in $N(\mathbf{x})$ will have any contribution to $J(\mathbf{w}, \xi)$ when none of them are error support vectors. Now consider the modified cost function $\tilde{J}(\mathbf{w}, \xi)$ in Eq. (3). It can be seen that in such a case the optimal solution of $J(\mathbf{w}, \xi)$ will also be the optimal solution of the modified $\tilde{J}(\mathbf{w}, \xi)$. Consequently, the resulting Ada-SVM decision function for the query sample $\mathbf{x}$ will be identical to the regular SVM when the same cost factor C is used.

The above proposition offers a rather intuitive insight on the nature of the Ada-SVM. Recall that for a sample $\mathbf{x}_i$ that is not an error support vector, we have $y_i f_{SVM}(\mathbf{x}_i) \geqslant 1$, that is, $\mathbf{x}_i$ is correctly classified by the decision margin of the classifier. Thus, the Ada-SVM will simply reduce to the regular SVM when the latter can correctly classify by the decision margin those training samples in the neighborhood of the query $\mathbf{x}$; otherwise, the SVM function will be adjusted with an increased emphasis toward those misclassified samples.

Furthermore, recall that the support vectors are associated with the decision boundary of the decision function $f_{SVM}(\mathbf{x})$. When the training samples in the neighborhood of the query $\mathbf{x}$ are away from the decision boundary, so is $\mathbf{x}$. Thus, the result in Proposition I implies the following: the Ada-SVM makes the same decision as the regular SVM classifier $f_{SVM}(\mathbf{x})$ for a query sample $\mathbf{x}$ that it is located far away from its decision boundary; on the other hand, when $\mathbf{x}$ is near the decision boundary of $f_{SVM}(\mathbf{x})$ and that the samples in its neighborhood are not well-classified (signaled by the presence of error support vectors), the SVM will be retrained with an emphasis on the samples in the local neighborhood of $\mathbf{x}$. As we explain next, this observation can lead to a computationally efficient implementation of the Ada-SVM classifier.

2.4. Implementation issues

Compared to the regular SVM, it may seem that the Ada-SVM will be much more demanding computationally, because the modified cost function $\tilde{J}(\mathbf{w}, \xi)$ in Eq. (3) would vary with the query sample $\mathbf{x}$, which would need to be re-optimized for every $\mathbf{x}$. However, based on the result in Proposition I, this is not the case. Instead, we can greatly reduce the extra computation burden by employing a regular SVM. Specifically, we adopt the following procedure for training the Ada-SVM: for each query, we first apply a regular SVM classifier on its similar cases. If it can correctly classify all of them, then we apply this SVM classifier to the query as well; otherwise, we invoke the Ada-SVM procedure. The rationale behind this is that if the SVM classifier performs well on the similar cases, it will likely perform well on the query as well. Our experiments show that this can result in marked saving in computation time.

For retraining the SVM, the optimization problem for the modified cost function in Eq. (3) can be solved in a similar fashion as in the case of regular SVM in Eq. (2). Indeed, by using the method of Lagrange multipliers and applying the Kuhn–Tucker conditions, the dual problem of Eq. (3) is obtained as maximization of:

$$
\begin{aligned}
J(\vec{\alpha}) &= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\
\text{s.t.} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\
& 0 \leqslant \alpha_i \leqslant C^{(s)}, \quad \mathbf{x}_i \in N(\mathbf{x}) \\
& 0 \leqslant \alpha_i \leqslant C, \quad \mathbf{x}_i \notin N(\mathbf{x})
\end{aligned}
\tag{5}
$$

The maximization is accomplished by quadratic programming. To speed up the numerical algorithm, the so-called incremental learning technique can be applied [17], in which the cost function in Eq. (3) is treated as a perturbation from the regular SVM in Eq. (2), and thus, the regular SVM solution can be used as a starting point for the solution of Eq. (5).

As in regular SVM, the parameters of the Ada-SVM will have to be determined during the training phase. In particular, the newly introduced penalty factor t can be determined using a leave-one-out cross-validation procedure on the set of similar samples to the query. Too large a value for t can lead to over-emphasis on the training samples near the query, which may cause over-fitting. On the other hand, too small a value for t may not have enough impact on the cost function. For each query, we pick the t value that corresponds to the lowest error rate resulting from this procedure, i.e.,

$$
t = \underset{t}{\operatorname{argmin}} \sum_{\mathbf{x}_i \in N(\mathbf{x})} [y_i - \operatorname{sign}(\tilde{f}_{SVM}(\mathbf{x}_i))] \tag{6}
$$

where $\tilde{f}_{SVM}(\mathbf{x})$ is the Ada-SVM classifier. This yields a customized penalty factor for each test sample. In our experiments the penalty factor was found to be typically in the range of $2 \leqslant t \leqslant 10$.

Another issue for the Ada-SVM is how to determine the similar samples to use for the query. In this work, we use the mammogram retrieval framework reported previously in Ref. [8]. For each query mammogram image, we invoke the retrieval system to obtain a set of similar mammograms from the database, which is then used for the Ada-SVM. For comparison purposes, we also experimented with use of other distance based similarity measures, including KNN based on the Euclidean distance [18], and discriminant adaptive nearest neighbors (DANN) [19]. DANN [19] is an improved version of the KNN measure based on the Euclidean distance for computing the similarity between two images. The KNN is based on the assumption that locally the class posterior probabilities are approximately constant. This is not true in many cases especially in a high dimensional space. The DANN modifies the KNN metric, so that the resulting local neighborhood stretches out in the direction for which the class probability does not vary much. In our experiments, a neighborhood was formed at each query point and the class distribution among these neighborhood points was used to decide how to adapt the metric. The adapted metric was then used in a nearest-neighbor rule at the query point. In essence, the modified neighborhood extends in parallel to the local decision boundaries and shrinks in directions orthogonal to the decision boundaries.

2.5. Cascade SVM

To further reduce the computational complexity of the proposed Ada-SVM classifier, we also introduced a pre-classifier stage in our experiments. This pre-classifier stage functions in the following fashion: for each query, we first examine its retrieved similar cases. If all these retrieved cases have the same class label, then we simply assign the same label to the query; otherwise, we will invoke the Ada-SVM classifier. The motivation for this pre-classifier stage is that if all the similar cases are found to be from the same class, it is a good indication that the neighborhood around the query is away from the decision boundary between the two classes. Because the decision function of SVM depends on only the samples located in the proximity of the boundary (i.e., support vectors), we simply invoke the KNN rule to avoid the more expensive Ada-SVM classifier. The decision steps of the resulting cascade SVM (Cas-SVM) classification system are described as follows: First, similar cases to a query sample are retrieved according to a similarity measure. Next, if all these similar cases have the same class label, simply assign the same label to the query; otherwise, we invoke the Ada-SVM classifier as follows: if the SVM can correctly classify all the retrieved similar

cases, we apply it to the query; if not, we apply the Ada-SVM to adapt the SVM classifier.

3. Evaluation study

3.1. Data set

In our study, we used a database of mammogram images collected by the Department of Radiology at the University of Chicago. The database consists of a total of 200 different mammogram images from 104 cases (46 malignant, 58 benign), of which all had lesions containing clustered MC which were histologically proven. These images were digitized with a spatial resolution of 0.1 mm/pixel and 10-bit grayscale. All these images contain clustered MC, as in the example shown earlier in Fig. 1. The MCs in each image have been identified by a group of expert radiologists. Many of them are extremely difficult to classify [20]; in an observer study the average classification performance by a group of five attending radiologists on these cases yielded a value of only 0.62 in the area under the receiver-operating characteristic (ROC) curve [20].

3.2. Experiment setup

For the retrieval system, we used the learning based similarity measure reported in Ref. [8], where 600 image pairs had been scored in a human observer study for training the similarity function. To test the proposed retrieval-driven Ada-SVM classifier, the 200 mammogram images were used in a leave-one-out procedure. During each round, one mammogram image was used as the test sample (i.e., query); similar mammogram images were then retrieved for this query from the database based on the learned similarity function; subsequently, the test mammogram was classified by the Ada-SVM.

It is important to note that, to avoid any potential bias, during the leave-one-out procedure the held-out image for testing was also removed from training the retrieval stage. This achieved complete isolation of the test sample from any of the training sets.

3.3. Feature selection

In our previous work [17], a set of 10 features was used based on the geometric distribution of the MCs in a cluster. However, image features of individual MCs are very important in diagnosis of clustered MCs. To better characterize the similarity data by the experts, in this work we introduced eight additional features which were demonstrated to have high discriminating power for cancer diagnosis [4]. These eight features were selected to have intuitive meanings that correlate qualitatively to features used by radiologists [21]. Consequently, there were a total of 18 features used for describing the MCs. For the purpose of selecting the most relevant features for similarity learning and classification, we applied a feature selection procedure, called sequential backward selection [22]. It is a suboptimal searching procedure, but is simple and easy to implement. The following set of 12 features was finally selected for characterizing a MC cluster:

- Compactness of the cluster: a measure of roundness of the region occupied by the cluster.
- Eccentricity of the cluster: the eccentricity of the smallest ellipse of the region (ratio of the distance between the foci and the major axis).
- The number of MCs per unit area.
- The average of the inter-distance between neighboring MCs.
- The standard deviation of the inter-distance between neighboring MCs.
- Solidity of the cluster region: the ratio between cross-sectional area and the area of the convex hull formed by the MCs.
- The moment signature of the cluster region: computed based on the distance deviation of the boundary point from the center of the region.
- The number of MCs in the cluster.
- The mean effective volume (area times effective thickness) of individual MCs.
- The relative standard deviation of the effective thickness.
- The relative standard deviation of the effective volume.
- The second highest MC-shape-irregularity measure.

In our experiment, all the feature components were normalized to have the same dynamic range (0,1).

3.4. Performance evaluation for classification

To evaluate the performance of a classifier, we used the so-called ROC analysis [23], which is now used routinely for many classification tasks. An ROC curve is a plot of the classification sensitivity (i.e., true positive fraction, TPF) as the ordinate versus the specificity (i.e., false positive fraction, FPF) as the abscissa; for a given classifier, it is obtained by continuously varying the threshold associated with the decision function. Thus, at any given FPF, an ROC curve with a higher TPF corresponds to better classification performance. As a summary measure of overall diagnostic performance, the area under an ROC curve (denoted by A_z) is used. A larger A_z means better classification.

4. Experimental results and discussions

Fig. 4 summarizes the classification results achieved by the proposed retrieval-driven approach (Ada-SVM and Cas-SVM), where the obtained A_z value is plotted against N, the number of most similar cases used for the Ada-SVM classifier. For comparison, the best A_z value obtained by a regular SVM is also shown in Fig. 4.

As can be seen from Fig. 4, the classification result (A_z) could be improved from 0.7752 (SVM) to 0.8139 (Ada-SVM with $N=3$), and 0.8223 (Cas-SVM with $N=5$). All these classifiers, i.e., SVM, Ada-SVM and Cas-SVM, were under the same parameter setting (Gaussian RBF kernel, $\sigma=2.5$, $C=100$) which was determined with a leave-one-out procedure. A statistical comparison between SVM and Ada-SVM using the ROCKIT program [23] yielded a two-tailed p-value 0.0285 (one-tailed p-value 0.0142) for rejecting the null hypothesis that their corresponding ROC curves have the same area under them; moreover, approximate 95% confidence interval for the difference

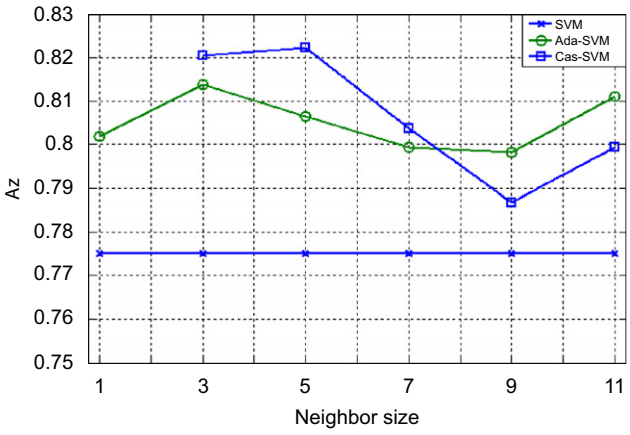

Fig. 4. Classification results by Ada-SVM, Cas-SVM and SVM.

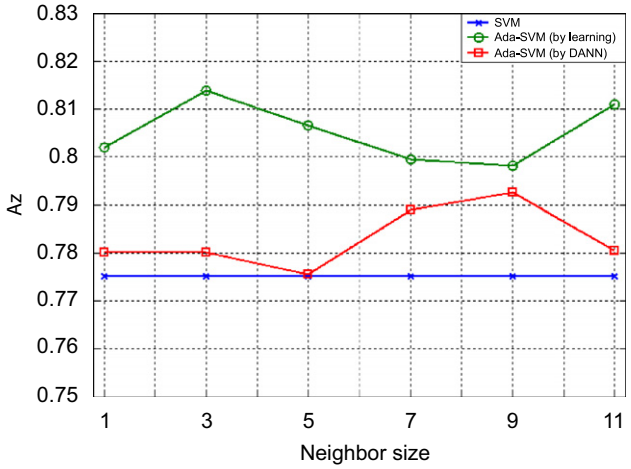

Fig. 5. Classification results by Ada-SVM using different similarity measures for retrieval.

Table 1
Retrieval-driven classification results (A_z) on mammogram database

	SVM	Ada-SVM	Cas-SVM
Learning based	0.7752	0.8139	0.8223
DANN	0.7752	0.7925	0.7883

was (0.0034, 0.0611). Also, the average error rates were 0.31 (SVM), 0.26 (Ada-SVM with $N = 3$) and 0.23 (Cas-SVM with $N = 5$), respectively. These results showed that the proposed retrieval-driven approach can lead to meaningful improvement in classification accuracy over the SVM, which was demonstrated to outperform several other state-of-the-art methods in our previous study [4].

We also note from Fig. 4 that as the size N of retrieved images is further increased the classification performance A_z value starts to decrease. We believe that this is due to the fact that the database is limited in size, which in turn limits the number of truly "similar" cases to the query. Thus, further increase of the number of retrieved images will no longer be beneficial.

For comparison, in Fig. 5 we also show the classification results obtained by the Ada-SVM using a different similarity measure, the discriminant adaptive nearest neighbors [19] (DANN), for retrieving similar images. Note that the classification result could still be improved from 0.7752 (SVM) to 0.7925 (Ada-SVM with $N = 9$). For clarity, the best classification results with different similarity measures for retrieval are shown in Table 1.

In our experiments, we also tested the proposed approach by grouping multiple views from the same cases. Specifically, the multiple views were treated as separate samples, but grouped together either for training or for testing in classification. In testing a classifier, a case is classified as malignant when any of its multiple views are classified as malignant. Under this setup, the performance was improved from 0.7691 (SVM) to 0.8050 (Ada-SVM with $N = 7$) and 0.8134 (Cas-SVM with $N = 7$). The error rates were 0.3173 (SVM), 0.2692 (Ada-SVM with $N = 7$) and 0.25 (Cas-SVM with $N = 7$), respectively. As above, SVM, Ada-SVM and Cas-SVM were using the same parameter setting (Gaussian RBF kernel, $\sigma = 2.5$, $C = 100$). Interestingly, these results are similar to those obtained above. In addition, when the distance based measure (DANN) was used in the retrieval stage, the classification result could be improved from 0.7691 (SVM) to 0.7856 (Ada-SVM with $N = 9$).

5. Conclusion

In this paper, we proposed a classification approach assisted by content-based image retrieval to improve the classification accuracy in computer aided diagnosis for breast cancer. We presented the proposed adaptive classification scheme in the context of SVM learning, which has been demonstrated to outperform several competing methods in breast cancer classification. The proposed retrieval and classification framework was developed and tested using a database of 200 mammogram images collected by the Department of Radiology at the University of Chicago. While the data set is somewhat limited in size, our results demonstrated that the proposed Ada-SVM classifier could lead to reduced generalization error. As a part of the proposed CBIR system, we demonstrated that a numerical observer's (classifier) performance could be improved by incorporating retrieved similar cases. Encouraged by this initial success, we plan to further develop and validate the proposed approach using more clinical evaluations in future studies. It is reasonable to expect that use of a significantly enlarged database will increase the number of truly "similar" cases for retrieval, which would lead to even bigger improvement in classification performance by the Ada-SVM classifier.

Acknowledgment

R.M. Nishikawa is a shareholder in Hologic, Inc. (Bedford, MA). He and the University of Chicago receive research funds and royalties from Hologic.

References

[1] American Cancer Society, Cancer facts and figures ⟨http://www.cancer.org⟩, 2008.
[2] A.M. Knutzen, J.J. Gisvold, Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions, Mayo Clin. Proc. 68 (1993) 454–460.
[3] D.B. Kopans, The positive predictive value of mammography, AJR 158 (1992) 521–526.
[4] L. Wei, Y. Yang, R.M. Nishikawa, A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications, IEEE Trans. Med. Imaging 24 (3) (2005) 371–380.
[5] R.N. Strickland, H.L. Hahn, Wavelet transforms for detecting micro-calcifications in mammograms, IEEE Trans. Med. Imaging 15 (1996) 218–229.
[6] H. Soltanian-Zadeh, F. Rafiee-Rad, S.P.-N. D, Comparison of multiwavelet, wavelet, haralick, and shape features for microcalcification classification in mammograms, Pattern Recognition 37 (2004) 1973–1986.
[7] M.D. Santo, M. Molinara, F. Tortorella, M. Vento, Automatic classification of clustered microcalcifications by a multiple expert system, Pattern Recognition 36 (2003) 1467–1477.
[8] L. Wei, Y. Yang, R.M. Nishikawa, M.N. Wernick, Learning of perceptual similarity from expert readers for mammogram retrieval, in: IEEE International Symposium on Biomedical Imaging, 2006, pp. 1356-1359.
[9] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[10] M. Pontil, A. Verri, Support vector machines for 3-d object recognition, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 637–646.
[11] V. Wan, W.M. Campbell, Support vector machines for speaker verification and identification, in: Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, 2000, pp. 775–784.
[12] E. Osuna, R. Freund, F. Girosi, Training support vector machines: application to face detection, in: Proceedings of the Computer Vision and Pattern Recognition, 1997, pp. 130–136.
[13] T. Joachims, Transductive inference for text classification using support vector machines, in: Proceedings of the International Conference on Machine Learning.
[14] C.J. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discovery 2 (1998) 121–167.
[15] I. El-Naqa, Y. Yang, M.N. Wernick, N.P. Galatsanos, R.M. Nishikawa, A support vector machine approach for detection of microcalcifications, IEEE Trans. Med. Imaging 21 (12) (2002) 1552–1563.
[16] R. Herbrich, J. Weston, Adaptive margin support vector machines for classification, in: Ninth International Conference on Artificial Neural Networks, vol. 2, 1999, pp. 880–885.
[17] I. El-Naqa, Y. Yang, N.P. Galatsanos, M.N. Wernick, Relevance feedback based on incremental learning for mammogram retrieval, in: International Conference on Image Processing, vol. 1, 2003, pp. 729–732.
[18] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.
[19] T. Hastie, R. Tibshirani, Discriminant adaptive nearest neighbor classification, IEEE Trans. Pattern Anal. Mach. Intell. 18 (6) (1996) 607–616.

[20] Y. Jiang, R.M. Nishikawa, R.A. Schmidt, C.E. Metz, M.L. Giger, K. Doi, Improving breast cancer diagnosis with computer-aided diagnosis, Acad. Radiol. 6 (1999) 22–33.
[21] Y. Jiang, R.M. Nishikawa, E.E. Wolverton, C.E. Metz, M.L. Giger, R.A. Schmidt, C.J. Vyborny, Malignant and benign clustered microcalcifications: automated feature analysis and classification, Radiology 198 (1996) 671–678.
[22] S. Theodoridis, K. Koutroumbas, Pattern Recognition, Academic Press, New York, 2003.
[23] C.E. Metz, B.A. Herman, J. Shen, Maximum-likelihood estimation of receiver operating (ROC) curves from continuously-distributed data, Stat. Med. 17 (1998) 1033–1053.

About the Author—LIYANG WEI received the B.S. and M.S. degrees in Biomedical Engineering from Xi'an Jiaotong University, Xi'an, China, in 1998 and 2001, respectively. She received her Ph.D. degree in Biomedical Engineering from Illinois Institute of Technology, Chicago, in 2006. Her research interests include medical image analysis, machine learning, pattern recognition and computer-aided diagnosis.

About the Author—YONGYI YANG received the B.S.E.E. and M.S.E.E. degrees from Northern Jiaotong University, Beijing, China, in 1985 and 1988, respectively. He received the M.S. degree in Applied Mathematics and the Ph.D. degree in Electrical Engineering from Illinois Institute of Technology (IIT), Chicago, in 1992 and 1994, respectively. Dr. Yang is currently on the Faculty of the Department of Electrical and Computer Engineering at IIT, where he is a Professor. Prior to this position, he was a Faculty member with the Institute of Information Science, Northern Jiaotong University. His research interests are in signal and image processing, medical imaging, machine learning, pattern recognition and biomedical applications. He is a co-author of Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics, John Wiley & Sons, Inc., 1998. He is an Associate Editor for the IEEE Transactions on Image Processing.

About the Author—ROBERT M. NISHIKAWA received his B.Sc. in physics in 1981 and his M.Sc. and Ph.D. in Medical Biophysics in 1984 and 1990, respectively, all from the University of Toronto. He is currently an Associate Professor in the Department of Radiology and the Committee on Medical Physics at the University of Chicago. He is director of the Carl J. Vyborny Translational Laboratory for Breast Imaging Research. He is also a fellow of the American Association of Physicists in Medicine (AAPM). His research interests are in computer-aided diagnosis, breast imaging, and evaluation of medical technologies.