Learning Distance Metrics for Interactive Search-Assisted Diagnosis of Mammograms

Liu Yang[a], Rong Jin[a], Rahul Sukthankar[b,c], Bin Zheng[d,e], Lily Mummert[b], M. Satyanarayanan[c], Mei Chen[b,c], and Drazen Jukic[d,e]

[a] Michigan State University, East Lansing, MI 48824
[b] Intel Research Pittsburgh, Pittsburgh, PA 15213
[c] Carnegie Mellon University, Pittsburgh, PA 15213
[d] University of Pittsburgh, Pittsburgh, PA 15213
[e] University of Pittsburgh Medical Center, Pittsburgh, PA 15213

ABSTRACT

The goal of interactive search-assisted diagnosis (ISAD) is to enable doctors to make better decisions about a given case by providing a selection of similar annotated cases. For instance, a radiologist examining a suspicious mass could study labeled mammograms with similar conditions and weigh the outcome of their biopsy results before determining whether to recommend a biopsy. The fundamental challenge in developing ISAD systems is the identification of similar cases, not simply in terms of superficial image characteristics, but in a medically-relevant sense. This task involves three aspects: extraction of a representative set of features, identifying an appropriate measure of similarity in the high-dimensional feature space, and return the most similar matches at interactive speed. The first has been an active research area for several decades. The second has largely been ignored by the medical imaging community. The third can be achieved using the Diamond framework, an open-source platform that enables efficient exploration of large distributed complex data repositories. This paper focuses on the second aspect. We show that the choice of distance metric affects the accuracy of an ISAD system and that machine learning enables the construction of effective domain-specific distance metrics. In the learned distance, data points with the same labels (e.g., malignant masses) are closer than data points with different labels (e.g., malignant vs. benign). Thus, the labels of the near neighbors of a new case are likely to be informative. We present and evaluate several novel methods for distance metric learning and evaluate them on a database involving 2522 mass regions of interest (ROI) extracted from digital mammograms, with ground truth defined by biopsy results (1800 malignant, 722 benign). Our results show that learned distance metrics improve both classification (ROC curve) and retrieval performance.

Keywords: methods: classification and classifier design; modalities: mammography; diagnostic task: diagnosis (mass classification); CAD/ISAD; machine learning: (distance metrics, boosting); Diamond.

1. INTRODUCTION

Computer-aided detection (CAD) of breast cancer is rapidly becoming a well-accepted clinical practice to assist radiologists in interpreting screening mammograms.[1,2] A number of studies have determined that radiologists' attitude toward and acceptance of CAD-cued micro-calcification clusters and masses were substantially different.[3,4] Due to the high sensitivity (i.e., $> 98\%$[5,6]), radiologists heavily rely on CAD-cued results while searching for and identifying micro-calcification clusters, which substantially improves the efficiency of radiologists in interpreting screening mammograms and also helps them detect more subtle cancers associated with micro-calcifications.[7] However, the lower CAD sensitivity for mass detection and the higher false-positive rates reduces the usefulness of CAD-cued masses. For example, two studies reported that CAD detected 77% (89 of 115) false-negative cancers[8] and 65% (80 of 123) of cancers depicted on prior images in which the masses are considered visible in retrospective reviews.[9] Although CAD schemes can detect a substantial fraction

Author contact information: (please send correspondence to Rahul Sukthankar)
Liu Yang, yangliu1@cse.msu.edu; Rong Jin, rongjin@cse.msu.edu; Rahul Sukthankar, rahuls@cs.cmu.edu; Bin Zheng, zhengb@upmc.edu; Lily Mummert, lily.b.mummert@intel.com; M. Satyanarayanan, satya@cs.cmu.edu; Mei Chen, mei.chen@intel.com; Drazen Jukic, jukicdm@upmc.edu.

of masses missed by radiologists in their initial interpretation, users in a busy clinical environment frequently discard CAD-cued detections for two reasons: (1) CAD systems are known to have a high false positive rate; (2) subtle masses are typically cued by CAD only on one of the views. In one prospective study involving 6,111 screening examinations radiologists discarded 7 of 8 CAD-cued false-negative masses (cancers).[10] In another recent prospective study, radiologists detected 43 of 48 cancers without using CAD. CAD detected 3 of the 5 missed cancers (two were micro-calcifications and one was a mass). Because the mass was detected only on one view, it was ultimately discarded by the radiologist and the two micro-calcification clusters were retained. As a result, radiologists detected 45 cancers (4.7% increase in sensitivity) with 15% increase in recall rate by using CAD.[11]

In order to improve CAD performance for mass detection and increase radiologists' confidence in CAD-cued mass regions, the development of interactive computer-aided diagnosis (ICAD) schemes has been attracting wide research interest.[12–15] The purpose of developing ICAD systems is to provide radiologists "visual aids" and increase their confidence in accepting CAD-cued subtle masses. For the development of ICAD systems, a large and diverse image reference library with verified pathology results is first assembled. Each selected region of interest (ROI) depicts a verified mass (either malignant or benign). In the application of ICAD systems, once a suspected mass region is identified or queried by the radiologist, the CAD scheme computes a set of features representing the region and its surrounding tissue. Then, the scheme searches for and identifies a set of reference regions that are considered "most similar" to the queried lesion. Several approaches have been investigated for automated similarity measurement, including the use of computer-extracted image features,[12] content-based image retrieval using a neural network,[13] multi-feature based k-nearest neighbor (KNN) algorithm,[14] and information theory (e.g., pixel value based mutual information).[15] The CAD-generated detection and/or classification scores, as well as the CAD-selected similar reference regions along with their verified outcome (malignant or benign) are displayed side by side with the queried image (or region) on an ICAD workstation.[12,14] By comparing the queried (suspected mass) region to the set of CAD-retrieved "similar" reference regions, radiologists can incorporate CAD-generated detection and classification scores into their decision making.

Previous studies have shown that CAD schemes with good performance can enhance radiologists' abilities whereas CAD schemes with poor performance can actually detract from them.[16,17] Therefore, a key step in the development of ICAD schemes for mammography is to improve their performance in classifying between malignant and benign mass regions. We show that the choice of distance metric affects the accuracy of an ISAD system and that machine learning enables the construction of effective domain-specific distance metrics. In the learned distance, data points with the same labels (e.g., malignant masses) are closer than data points with different labels (e.g., malignant vs. benign). Thus, the labels of the near neighbors of a new case are likely to be informative. We present and evaluate several novel methods for distance metric learning and evaluate them on a large and diverse image database using ROC and precision rank retrieval analysis.

The paper is organized as follows. Section 2 introduces the idea of *interactive-search based diagnosis* (ISAD) and outlines the challenge of learning similarity from data. Section 3 details the idea of supervised distance metric learning and presents three algorithms. Section 4 describes our implementation, including the UPMC dataset, choice of features and the search system. Section 5 presents experimental evaluations of distance metric learning against traditional Euclidean distance metrics. Section 6 concludes the paper.

2. INTERACTIVE SEARCH-ASSISTED DIAGNOSIS

Interactive search-assisted diagnosis (ISAD) is a form of interactive computer-aided diagnosis (ICAD) that focuses on retrieving medically-relevant annotated images from a large reference repository. Unlike traditional ICAD systems that primarily cue the radiologist with suspicious masses, ISAD aims to improve medical diagnosis by providing the user with additional metadata, such as biopsy results and outcome information, from similar historical cases. ISAD is conceptually similar to content-based image retrieval (CBIR),[18] where the goal is to retrieve images that match a particular semantic concept. However, in contrast to CBIR, where the query is typically text and the result a set of images, an ISAD query consists of an image and the result a set of textual annotations along with the corresponding similar images from the repository.

ISAD poses three research challenges. First, how should one characterize the image content in the ROI? Second, what criterion should be used to define similarity between two ROIs? Third, how can we efficiently perform near-neighbor searches over large repositories for novel queries? The first has been an active research problem in medical imaging for several years. This paper makes no contributions in that area; for our experiments, we employ the feature set proposed by Zheng *et al.*[19] for CAD.

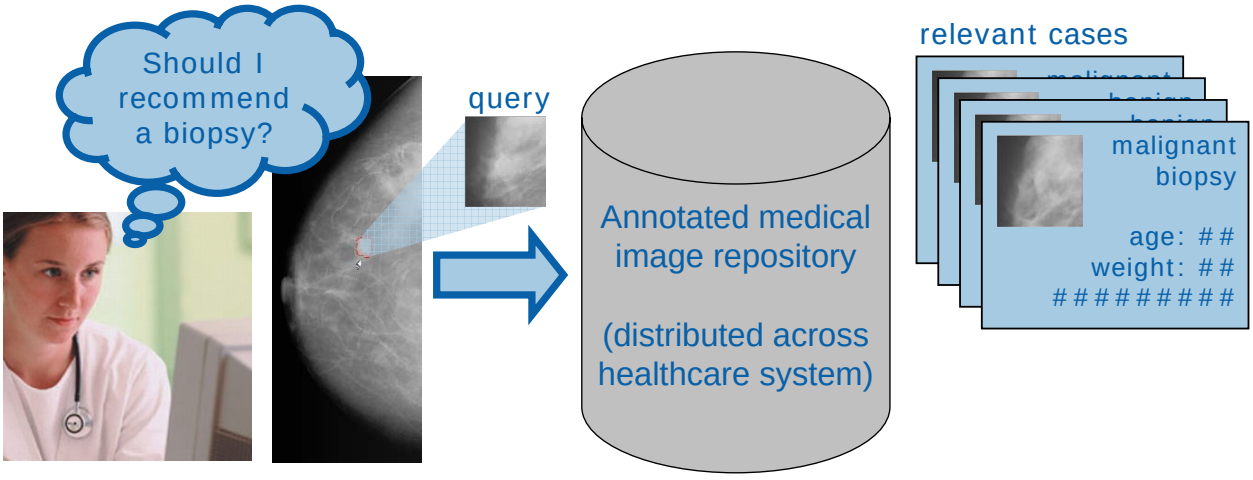


Figure 1: The goal of Interactive Search Assisted Diagnosis (ISAD) is to enable radiologists to make better decisions about a given case by presenting relevant annotated cases from large medical repositories.

The second challenge is the primary focus of this paper. Traditionally, one maps an ROI described using a set of m features to a point in m-dimensional space, $\Re^m$. In such a representation, the ROIs in the reference library generate a high-dimensional cloud of points. Given a novel query ROI, the ISAD system can perform feature extraction and map the query into a new point in the feature space. Similar ROIs in the reference library should correspond to near-neighbors of the query in feature space. A natural choice of distance metric in this space is the Euclidean distance. However, recent research in machine learning has shown that one can develop specialized distance metrics that improve classification and retrieval accuracy by exploiting any available side information. We explore several methods for learning suitable distances for ISAD.

The third challenge is to make near-neighbor search in large repositories efficient enough for interactive queries. The standard approach to efficient search (e.g., in web search) has been to employ indexing. Unfortunately, standard indexing techniques such as KD-trees,[20] fail in high-dimensional feature spaces due to the *curse of dimensionality*.[21] Consequently, practical systems have eschewed the use of near-neighbor searches in large, high-dimensional repositories. Fortunately, near-neighbor searches are amenable to parallel execution. Our ISAD application, described in Section 4, is implemented using the Diamond[22] distributed search framework. By partitioning the repository of ROI images over a set of servers, Diamond can parallelize the search and provide timely results for each ISAD query.

3. DISTANCE METRIC LEARNING

Supervised distance metric learning has recently become an active area of research in machine learning.[23–26] In this framework, the feature data is supplemented by side information in the form of pairwise "similarity" and "dissimilarity" relationships between objects. For instance, two ROI images that were visually similar to a radiologist could be tagged as "similar". In the absence of human labeling, we can tag reference cases with the same biopsy label (i.e., "benign" or "malignant") as being similar, and those with different biopsy labels as dissimilar. The goal is then to learn a distance function that best satisfies these constraints. In other words, a good distance function should ensure that the local neighborhood of a given object contains similar rather than dissimilar objects. Specifically, we expect that the local neighborhood of a suspicious mass will be dominated by malignant reference ROIs, if the query mass is malignant.

A variety of algorithms have been proposed for supervised distance metric learning. In general, these techniques formulate distance metric learning as an optimization problem (e.g., minimize error on the training set subject to the pairwise constraints). We briefly review some of these algorithms and present our novel approach, *boosted distance metric* learning (BDM) below.

3.1. Global Distance Metric Learning

The intuition behind global distance metric learning[26] (GDM) is straightforward: keep all of the "similar" data pairs close while separating pairs that are "dissimilar". This can be formulated as a optimization problem where the objective is to minimize the distance between "similar" pairs subject to the constraint that "dissimilar" pairs are well separated.

Thus, we formulate distance metric learning as follows. Let $\mathcal{C} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ be a collection of data points, where n is the number of samples and each $\mathbf{x}_i \in \mathbb{R}^m$ is a vector of m features. Let the set of similarity constraints and the set of dissimilarity constraints denoted by

$$\begin{aligned} \mathcal{S} &= \{(\mathbf{x}_i, \mathbf{x}_j) |\ \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same class — i.e., both malignant or both benign}\}, \\ \mathcal{D} &= \{(\mathbf{x}_i, \mathbf{x}_j) |\ \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to different classes}\}, \end{aligned}$$

respectively. Let the distance metric be denoted by matrix $\mathbf{A} \in \mathbf{R}^{m \times m}$, and the distance between two points $\mathbf{x}$ and $\mathbf{y}$ be expressed by

$$d^2_{\mathbf{A}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2_{\mathbf{A}} = (\mathbf{x} - \mathbf{y})^\top \mathbf{A} (\mathbf{x} - \mathbf{y}).$$

Then, our goal is to solve the optimization problem:

$$\begin{aligned} \min_{\mathbf{A} \in \mathbf{R}^{m \times m}} \quad & \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}} \|\mathbf{x}_i - \mathbf{x}_j\|^2_{\mathbf{A}} \\ \text{s. t.} \quad & \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}} \geq 1, \\ & \mathbf{A} \succeq 0. \end{aligned} \tag{1}$$

An attractive property of GDM is that Eqn. 1 is a convex problem, which means that it has a single global optimum and can be solved efficiently using standard techniques.

Unfortunately, multimodal data distributions in the feature space can lead to problems for GDM. In such cases, when data points from one class are interleaved with those from another, it may be impossible to simultaneously satisfy the goals of separating dissimilar examples and contracting similar ones. In such cases, the global distance metric may simply collapse the data into a lower-dimensional space — resulting in a detrimental impact on classification accuracy. This observation has stimulated interest in algorithms for local distance metric learning.

3.2. Local Distance Metric Learning

As discussed above, for many realistic data distributions, simultaneously satisfying all of the given similarity/dissimilarity constraints may be impossible. Local distance metric learning (LDM)[24] is conceptual modification of GDM where greater importance is given to satisfying *local constraints* (i.e., those constraints between nearby data points). Thus, by weighting constraints based on the distances between pairs of data points, the algorithm attempts to ensure that the local neighborhood of each data point will contain similar points. Clearly, employing the notion of "locality" in distance metric learning leads to a circular definition: the weights on each constraint depend upon the learned metric, but learning the metric requires that these weights be specified.

Consider a data point $\mathbf{x}$ that is involved in one of the constraints in the set $\mathcal{S}$ and the set $\mathcal{D}$. Let $\Phi_S(\mathbf{x}) = \{\mathbf{x}_i | (\mathbf{x}, \mathbf{x}_i) \in \mathcal{S}\}$ include all of the data points that pair with $\mathbf{x}$ in the similarity constraints. Similarly, let $\Phi_D(\mathbf{x}) = \{\mathbf{x}_i | (\mathbf{x}, \mathbf{x}_i) \in \mathcal{D}\}$ include all of the data points that pair with $\mathbf{x}$ in the dissimilarity constraints. Now, according to the kernel-based KNN, the probability of making the right prediction for $\mathbf{x}$, denoted by $\Pr(+|\mathbf{x})$, can be written as

$$\Pr(+|\mathbf{x}) = \frac{\sum\limits_{\mathbf{x}_i \in \Phi_S(\mathbf{x})} f(\mathbf{x}, \mathbf{x}_i)}{\sum\limits_{\mathbf{x}_i \in \Phi_S(\mathbf{x})} f(\mathbf{x}, \mathbf{x}_i) + \sum\limits_{\mathbf{x}_j \in \Phi_D(\mathbf{x})} f(\mathbf{x}, \mathbf{x}_j)} \tag{2}$$

where the kernel function $f(\mathbf{x}, \mathbf{x}')$ is defined as

$$f(\mathbf{x}, \mathbf{x}') = \exp\left(-\|\mathbf{x} - \mathbf{x}'\|^2_{\mathbf{A}}\right). \tag{3}$$

Using leave-one-out estimation and Eqn. 2, we can write the log likelihood for both $\mathcal{S}$ and $\mathcal{D}$ as

$$\begin{aligned}\mathcal{L}_l(\mathbf{A}) &= \sum_{\mathbf{x}\in\mathcal{T}} \log \Pr(+|\mathbf{x}) \\ &= \sum_{x\in\mathcal{T}} \log \left(\frac{\sum\limits_{\mathbf{x}_i\in\Phi_S(\mathbf{x})} f(\mathbf{x},\mathbf{x}_i)}{\sum\limits_{\mathbf{x}_i\in\Phi_S(\mathbf{x})} f(\mathbf{x},\mathbf{x}_i) + \sum\limits_{\mathbf{x}_j\in\Phi_D(\mathbf{x})} f(\mathbf{x},\mathbf{x}_j)} \right)\end{aligned} \tag{4}$$

where the set $\mathcal{T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ includes all of the data points involved in the constraints given in sets $\mathcal{S}$ and $\mathcal{D}$. Using maximum likelihood estimation, we cast local distance metric estimation into the following optimization problem

$$\begin{aligned}\min_{\mathbf{A}\in\mathbf{R}^{m\times m}} \quad & \mathcal{L}_l(\mathbf{A}) \\ \text{s. t.} \quad & \mathbf{A} \succeq 0.\end{aligned} \tag{5}$$

Remark: Note that in Eqn. 5, it is the *ratio* between the kernel function $f(\mathbf{x},\mathbf{x}_i)$ evaluated at different data points $\mathbf{x}_i$ that determines the probability $\Pr(+|\mathbf{x})$. When a data point $\mathbf{x}_i$ is relatively far from $\mathbf{x}$ compared to other data points in $\Phi(\mathbf{x})_S$ and $\Phi(\mathbf{x})_D$, its kernel value $f(\mathbf{x},\mathbf{x}_i)$ will be relatively smaller than the kernel value of other data points. Hence, local constraints (those involving data pairs that are close) will have a much greater impact on the objective function $\mathcal{L}_l(\mathbf{A})$ than constraints involving distant data points.

We solve this optimization problem efficiently using the iterative algorithm described in Yang *et al.*[24] In a manner analogous to Expectation-Maximization (EM),[27] the algorithm is initialized using a randomly-generated distance metric (corresponding to a legal positive semi-definite matrix $\mathbf{A}$). In each iteration, the algorithm alternates between: (1) computing weights on each constraint based on the current distance metric; and (2) re-estimating the parameters for a better distance metric by solving an optimization problem. As with EM, the algorithm is guaranteed to converge to a locally-optimal solution (but not to a global optimum).

3.3. Boosted Distance Metric Learning

Boosted distance metric (BDM) learning is a novel approach to supervised distance metric learning. Unlike standard algorithms that learn variants on the Euclidean distance, the BDM generates a weighted Hamming distance (i.e., a weighted sum of binary features). The approach is motivated by the recent success of boosted classifiers in machine learning. The key idea behind boosting[28] is that one can construct a very accurate classifier using an ensemble of appropriately-selected *weak* classifiers, where each weak classifier need only be slightly better than random chance. Boosting works in an iterative manner as follows. First, the ensemble is initialized with a weak classifier trained on the original dataset. Next, at the start of each iteration, the training data is re-weighted to increase the worth of any (previously-)misclassified exemplars and used to train a new weak classifier. As a result of the weighting, the new classifier has an incentive to focus on solving misclassified cases. The new classifier is added to the ensemble and the procedure repeated until the desired level of accuracy (on the training set) has been obtained. The output of the classifier ensemble is simply a weighted combination of the outputs of individual classifiers.

Whereas boosting is traditionally used to train classifiers (i.e., the input is a single data point and the output is a label), BDM employs boosting to learn a distance function (i.e., the input is a *pair* of data points and the output is a positive real number). The intuition behind BDM is that it projects the data into a space of binary features (Hamming space), where each dimension corresponds to the output of a weak classifier. Ideally, two data objects that are very similar will generate the same (binary) outputs from many of the weak classifiers and will therefore project to nearby regions in Hamming space. In other words, the binary features corresponding to semantically-similar data objects are likely to match in many bits. And while no single binary feature is particularly reliable, the output of the ensemble can become an accurate measure of the semantic distance between data objects. In each iteration, we first identify the subset of data points that should be near (based on label information) but are far apart in the current representation. We then identify a best binary projection that moves these points closer while keeping data points from different classes well separated. Each projection generates one bit in the representation, and the iterative procedure is repeated until either desired accuracy or storage constraints have been reached.

BDM is formalized as follows. As in LDM, the goal is to move the data points from the same classes close to each other while keeping data from different classes well separated. Let the set of labeled example pairs be denoted by $\mathcal{P} = \{(\mathbf{x}_i, \mathbf{x}_j, y_{i,j}) | \mathbf{x}_i \in \mathcal{D}, \mathbf{x}_j \in \mathcal{D}, y_{i,j} \in \{-1, 0, +1\}\}$ where the class label $y_{i,j}$ is defined as follows:

$$y_{i,j} = \begin{cases} +1 & \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same class;} \\ -1 & \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to different classes;} \\ 0 & \text{the relationship between the class} \\ & \text{labels of } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ is unknown.} \end{cases}$$

Since our goal is to move the data points from same classes close to each other while keeping the data points from different classes well separated, we use the following objective function for the BDM framework:

$$F(\mathcal{P}) = \sum_{i,j,k=1}^{n} \delta(y_{i,j}, -1)\delta(y_{i,k}, +1) \exp\left(d(\mathbf{x}_i, \mathbf{x}_k) - d(\mathbf{x}_i, \mathbf{x}_j)\right). \quad (6)$$

Each term in this function is evaluated based on the difference between $d(\mathbf{x}_i, \mathbf{x}_k)$ and $d(\mathbf{x}_i, \mathbf{x}_j)$. The former is the distance between two data points from different classes, and the latter is the distance between two data points from the same class. Hence, by minimizing the objective function $F(\mathcal{P})$, we can ensure that data points from the same classes will be kept closer to each other compared to data points of different classes. Note that naive approaches to this optimization problem are computationally expensive since the number of terms in the objective function is on the order of $\mathcal{O}(n^3)$. This motivates the need for efficient approaches to the problem.

To minimize the objective function in Eqn. 6, we need to define a distance function $d(\mathbf{x}_i, \mathbf{x}_j)$ that (1) is non-negative, and (2) satisfies the triangle inequality. Let $f(\mathbf{x}) : \mathbf{R}^n \rightarrow \{-1, +1\}$ denote the classification model that will be used to construct the distance function. For each example $\mathbf{x}$, the classifier $f(\cdot)$ will assign $\mathbf{x}$ to either the negative class (i.e., -1) or the positive class (i.e., $+1$). Let $f_t, t = 1, 2, \ldots, T$ denote the binary classifiers that are learned in successive iterations of the boosting algorithm. We then construct the distance function as a weighted Hamming distance:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{t=1}^{T} \alpha_t \left(f_t(\mathbf{x}_i) - f_t(\mathbf{x}_j)\right)^2, \quad (7)$$

where $\alpha_t \geq 0, t = 1, 2, \ldots, T$ are the combination weights. Clearly, $d(\mathbf{x}_a, \mathbf{x}_b)$ in Eqn. 7 is non-negative and also satisfies the triangle inequality.

Given the distance function in Eqn. 7, our goal is to learn appropriate classifiers $f_t(\mathbf{x}), t = 1, 2, \ldots, T$ and combination weights $\alpha_t, t = 1, 2, \ldots, T$. In order to efficiently learn the parameters and functions, we follow the idea of boosting and take the greedy approach for optimization. More specifically, we start with a constant function for distance, i.e., $d_0(\mathbf{x}_i, \mathbf{x}_j) = 0$, and learn a distance function $d_1(\mathbf{x}_i, \mathbf{x}_j) = d_0(\mathbf{x}_i, \mathbf{x}_j) + \alpha_1 \left(f_1(\mathbf{x}_i) - f_1(\mathbf{x}_j)\right)^2$. Using this distance function, the objective function in Eqn. 6 becomes a function of α_1 and $f_1(\mathbf{x})$, and can be optimized efficiently using bound optimization. Given distance function $d_1(\mathbf{x}_i, \mathbf{x}_j)$, we then proceed to learn α_2 and $f_2(\mathbf{x})$ by considering $d_2(\mathbf{x}_i, \mathbf{x}_j)$ that is computed as

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = 0 + \alpha_1 \left(f_1(\mathbf{x}_i) - f_1(\mathbf{x}_j)\right)^2 + \alpha_2 \left(f_1(\mathbf{x}_i) - f_2(\mathbf{x}_j)\right)^2 d_1(\mathbf{x}_i, \mathbf{x}_j) + \alpha_2 \left(f_1(\mathbf{x}_i) - f_2(\mathbf{x}_j)\right)^2.$$

In general, given a distance function $d_{t-1}(\mathbf{x}_i, \mathbf{x}_j)$ that is learned in iteration $t-1$, we will learn α_t and $f_t(\mathbf{x})$ by using the following distance function:

$$d_t(\mathbf{x}_i, \mathbf{x}_j) = d_{t-1}(\mathbf{x}_i, \mathbf{x}_j) + \alpha_t \left(f_t(\mathbf{x}_i) - f_t(\mathbf{x}_j)\right)^2.$$

Using the above expression for distance function, the objective function in Eqn. 6 becomes a function of α_t and $f_t(\mathbf{x})$, i.e.,

$$F(\mathcal{P}) = \sum_{i,j,k=1}^{n} \left\{\delta(y_{i,j}, -1)\, \delta(y_{i,k}, 1) \exp(d_{i,k} - d_{i,j} + \alpha(f(\mathbf{x}_i) - f(\mathbf{x}_k))^2 - \alpha(f(\mathbf{x}_i) - f(\mathbf{x}_j))^2)\right\} \quad (8)$$

where $d_{i,j}$, α and $f(.)$ denote $d_{t-1}(\mathbf{x}_i, \mathbf{x}_j)$, α_t and $f_t(.)$, respectively. Hence, the key question is how to find the classifier $f(\mathbf{x})$ and weight α. Appendix A details our efficient optimization algorithm for solving the problem.

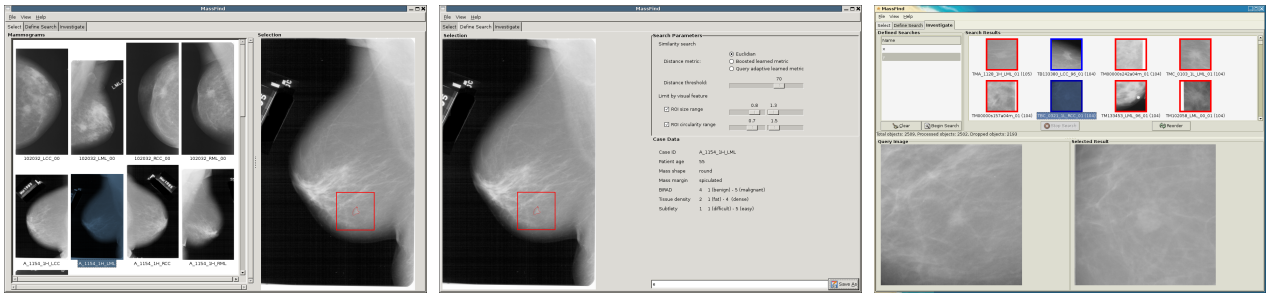


Figure 2: MassFind is an application that enables interactive search-assisted diagnosis on digitized mammograms. Once a query ROI has been selected, MassFind performs an efficient near-neighbor search over a large repository of annotated reference images, distributed over several Diamond[22] nodes. The retrieved images and their associated metadata provide radiologists with additional information about the current case. Our experiments show that learned distance metrics can significantly improve retrieval accuracy.

4. IMPLEMENTATION

4.1. Image Database and Features

The UPMC dataset consists of 2522 regions of interest (ROIs) depicting verified masses from a reference library established at the Radiographic Imaging Research Center, University of Pittsburgh. Among these, 1800 regions are associated with pathology-proven malignant masses and the remaining 722 regions are associated with benign masses. Each ROI is a 512×512 pixel region extracted from a digitized mammogram (with each pixel mapping to $100\mu m \times 100\mu m$). The mass boundary contour was first automatically detected using an adaptive topographic region growth algorithm.[19] Based on local contrast estimation, this region growth algorithm grows three topographic layers to define the final boundary contour of the mass region. The growth region (segmentation result) was then visually examined and manually corrected (if needed) by experienced observers. The computer scheme computed a set of 36 morphological and intensity (pixel value) distribution based features to represent each selected mass region. Among these 36 features, 8 were computed from the whole breast area depicted on one image ("global" features) and the remaining 24 were computed from the segmented mass region and its surrounding tissue background ("local" features). More information on the features is available in Zheng *et al.*[14,19]

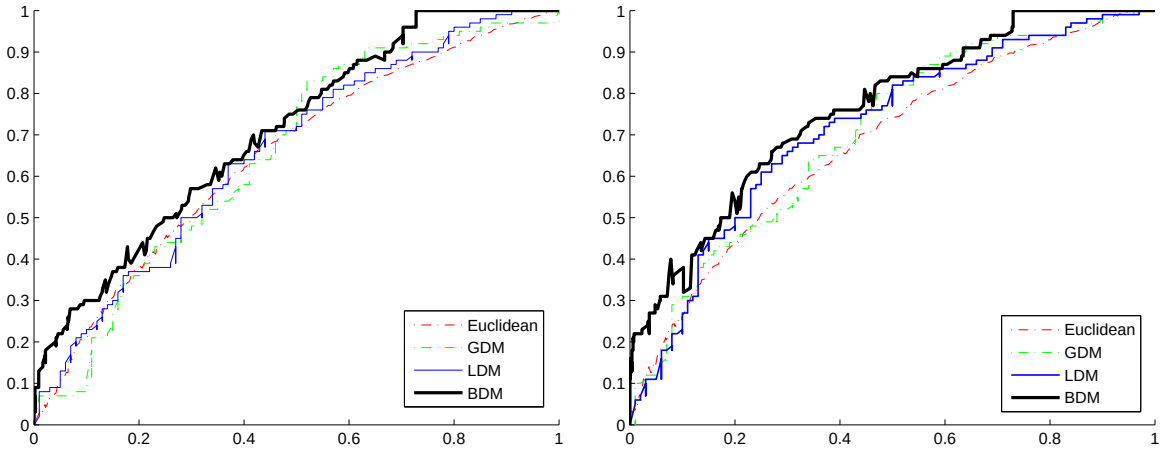
4.2. MassFind: A Prototype Application of ISAD for Breast Lesions

We have developed a prototype implementation of ISAD using the Diamond distributed search framework.[22] Figure 2 shows some screenshots of an interactive search. In Fig. 2(a), the user selects a query ROI centered on a suspicious mass from one of the mammograms in the left panel. This search is refined in Fig. 2(b) and sent to a set of Diamond servers. Each server manages a subset of the reference library and performs a near-neighbor search between the query ROI and candidate images from the library using the selected distance metric. A large fraction of the images in the reference library can be discarded since they match poorly. Even though the large dimensionality of the feature space prevents effective indexing, Diamond efficiently performs this retrieval task by distributing load over multiple compute nodes, intelligently structuring the search and exploiting cached results from similar previous queries. Since the computation is performed close to storage, Diamond can reject unlikely candidates at the source, enabling significant savings in network resources. The set of images that match better than the specified threshold are returned to the MassFind client. MassFind then aggregates the results from the Diamond nodes and sorts them in increasing order of distance for display, as shown in Fig. 2(c). The user can then compare the query ROI against retrieved ROIs for visual similarity, examine biopsy results for the reference ROIs and study the metadata associated with reference ROIs to determine the best diagnosis for the query image.

5. EXPERIMENTAL RESULTS

This section presents a comparison of learned distance metrics against the standard Euclidean metric, both in terms of classification performance and retrieval accuracy. The following methodology was employed.

- Dataset: 2522 ROIs from the UPMC dataset, described in Sec. 4.1.



(a) 200 training examples (b) 1200 training examples

Figure 3: Classification performance (ROC curves) for both small and large training set sizes for different distance metrics. Similar results are obtained for other training set sizes (not shown). Area under ROC for these experiments is shown in Table 1.

- Features: 36-dimensional feature vector, as described in Zheng *et al.*[14] Feature vectors were independently normalized in each dimension.
- Classifier: kernel-based KNN (varying the threshold on posterior probability generates ROC curves).
- Training set: varied from 200 to 1200 ROIs, in steps of 200, evenly distributed across malignant and benign cases.
- Test set: 100 randomly-selected ROIs that were not employed in training.
- All reported results are averages over 10 independent trials with different randomly-selected training and test sets.

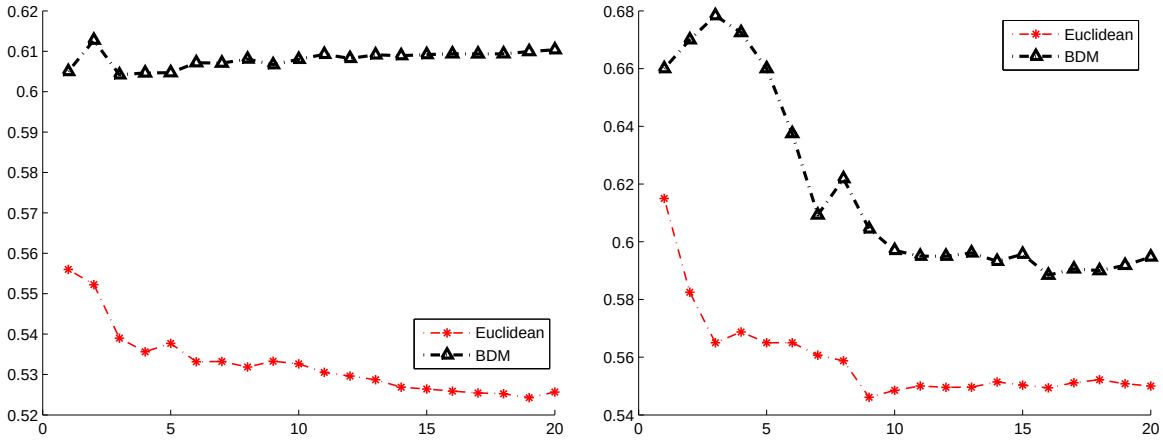
Classification and retrieval are different tasks, and are evaluated according to different established criteria. In classification, the goal is to determine whether a given ROI is malignant. Performance can be measured along two axes (1) detection rate, and (2) false-positive rate. The former is the fraction of malignant masses that were correctly classified; the latter is the fraction of benign cases that were incorrectly classified as malignant. An ideal classifier will be able to achieve a perfect detection rate with zero false positives. In reality, there is always a trade-off: for a given classifier, one can generally improve detection rate only at the expense of more false positives; conversely, reducing the false positive rate will also cause the detection rate to suffer. Varying the acceptance threshold of the classifier generates an ROC curve. Figure 3 shows ROC curves for a set of classifiers employing different distance metrics: Euclidean, GDM, LDM and BDM. We see that BDM outperforms the other distance metrics in all cases (small or large amounts of training data). The area under the ROC curve (AUR) is frequently employed as a summary of classification performance, and classifiers with a high AUR are preferred.

Table 1 summarizes the AUR for classifiers employing the set of distance metrics for a range of different training set scenarios. We make several observations. First, as expected the AUR generally increases with additional training data for all of the classifiers. Second, we note that the learned distance metrics (GDM, LDM and BDM) all outperform the standard Euclidean metric, indicating that learning distances is worthwhile from the standpoint of classification. Among the distance metrics, the boosted distance metric (BDM) is clearly and consistently the best.

However, neither ROC curves nor the AUR measure appropriately characterize the desired performance of a distance metric for the ISAD application. Recall that the goal in ISAD is not to automatically classify an ROI as either malignant or benign but rather to provide the radiologist with a small set of similar ROIs from the reference library. Unlike in classification, where the decision is made using (distance-weighted) contributions from every image in the ROI, ISAD demands that the small set of displayed images be relevant. In other words, the proportion of malignant reference images in the display set should be high if the query ROI depicts a malignant mass and low otherwise. This is captured by the precision at the desired rank (termed *precision at* n or P@n). Precision is defined as the fraction of correct objects among the retrieved objects. For instance, if the ISAD system displays the 8 most similar reference images and 6 of them are correct, then P@n would be 0.75.

Training size	200	400	600	800	1000	1200
Euclidean	0.6506	0.6606	0.6666	0.6753	0.6807	0.6823
GDM	0.6544	0.6783	0.6628	0.6963	0.7019	0.7029
LDM	0.6602	0.6704	0.6973	0.7000	0.7018	0.7103
BDM	0.6819	0.7074	0.7187	0.7334	0.7375	0.7381

Table 1: Summary of classification performance (area under ROC curve) for different distance metrics with a variety of training set sizes. Learned distance metrics show better classification performance under all conditions, and BDM clearly outperforms the other learned distances.



(a) 200 training examples (b) 1200 training examples

Figure 4: Precision against retrieval rank for small and large training set sizes. Learned distance metrics show a slight but consistent improvement over the Euclidean distance. Since BDM dominates the other learned distance metrics, curves for LDM and GDM are not shown here.

Figure 4 shows the P@n for ranks 1 to 20 for both small (200) and large (1200) numbers of training examples. Employing learned distance metrics for ISAD leads to small but consistent improvements in the precision of retrieved results.

6. CONCLUSION

This paper proposes the idea of interactive search-assisted diagnosis (ISAD) of masses in mammograms. ISAD is form of interactive computer-aided diagnosis where the system retrieves relevant data from a large reference collection of ROI images to enable radiologists to make better decisions about the given case. Developing an ISAD system entails three challenges: (1) the choice of visual features; (2) the criterion for defining similarity between ROI images; (3) efficient near-neighbor search in high dimensions on large data collections. We focus on the second challenge and investigate a variety of novel techniques for improving the quality of similarity search for ISAD. Experimental results on a large database of ROI images from UPMC indicate that the boosted distance metric (BDM) algorithm outperforms other learned distances and the standard Euclidean distance. We present a prototype system, MassFind, that enables ISAD on large real-world datasets.

APPENDIX A. OPTIMIZATION ALGORITHM FOR BOOSTED DISTANCE METRIC LEARNING

The first step toward efficient optimization is to decouple the interaction between the classification function $f(\mathbf{x})$ and the combination weight α. This can be achieved using Jensen's inequality and the convexity of exponential functions. The

resulting upper bound for the objective function $F(\mathcal{P})$ can be expressed as follows:

$$\begin{aligned} F(\mathcal{P}) - \tilde{F}(\mathcal{P}) \quad \leq \quad & \frac{\exp(-8\alpha)-1}{8} \sum_{i,j=1}^{n} \{\delta(y_{i,j},-1)\mu_i^+ \exp(-d_{i,j})(f(\mathbf{x}_i)-f(\mathbf{x}_j))^2\} + \\ & \frac{\exp(8\alpha)-1}{8} \sum_{i,j=1}^{n} \{\delta(y_{i,j},1)\mu_i^- \exp(d_{i,j})(f(\mathbf{x}_i)-f(\mathbf{x}_j))^2\}, \end{aligned} \tag{9}$$

where

$$\begin{aligned} \tilde{F}(\mathcal{P}) &= \sum_{i,j,k=1}^{n} \delta(y_{i,j},-1)\delta(y_{i,k},1)\exp\left(-d_{i,j}+d_{i,k}\right) \\ \mu_i^+ &= \sum_{j=1}^{n} \delta(y_{i,j},1)\exp(d_{i,j}) \\ \mu_i^- &= \sum_{j=1}^{n} \delta(y_{i,j},-1)\exp(-d_{i,j}). \end{aligned}$$

In the equations above, the quantity μ_i^+ indicates how far the data point $\mathbf{x}_i$ is kept from those data points that share its class label. Similarly, the quantity μ_j^- indicates how close the data point $\mathbf{x}_i$ is kept from those data points with different class labels. We can further rewrite the upper bound in Eqn. 9 in matrix form:

$$F(\mathcal{P}) - \tilde{F}(\mathcal{P}) \leq \quad \frac{\exp(-8\alpha)-1}{8}\mathbf{f}^\top L^+ \mathbf{f} + \frac{\exp(8\alpha)-1}{8}\mathbf{f}^\top L^- \mathbf{f}, \tag{10}$$

where $\mathbf{f} = (f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n))$ denote the class labels for all the examples. L^- and L^+ are the combinatorial Laplacian that are built based on the similarity matrices S^- and S^+, which can be computed as follows

$$S_{i,j}^- = \frac{1}{2}\delta(y_{i,j},1)\exp(d_{i,j})\left(\mu_i^- + \mu_j^-\right) \tag{11}$$

$$S_{i,j}^+ = \frac{1}{2}\delta(y_{i,j},-1)\exp(-d_{i,j})\left(\mu_i^+ + \mu_j^+\right). \tag{12}$$

Since the similarity S^- only depends on the pairs of data points that share the same class labels, the quantity $\mathbf{f}^\top L^- \mathbf{f}$ in Eqn. 10 measures the consistency between the binary feature $\mathbf{f}$ and the correlation between the data points of the same classes. Similarly, the quantity $\mathbf{f}^\top L^+ \mathbf{f}$ in Eqn. 10 measures the consistency between the binary feature $\mathbf{f}$ and the correlation between the data points of different classes. Given the expression in Eqn. 10, we can efficiently compute the solution $\mathbf{f}$ and α using standard optimization methods.

ACKNOWLEDGMENTS

Liu Yang was supported by an Intel Research summer internship. This publication was made possible by Grant Number 1 UL1 RR024153-01 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NCRR or NIH. This work is also supported in part by Grants CA77850 and CA101733 to the University of Pittsburgh from the National Cancer Institute, National Institutes of Health. We thank Todd Mowry and Limor Fix for valuable discussions on interactive search-assisted diagnosis.

REFERENCES

1. R. L. Birdwell, P. Bandodkar, and D. M. Ikeda, "Computer-aided detection with screening mammography in a university hospital setting," *Radiology* **236**, pp. 451–457, 2005.
2. S. H. Taplin, C. M. Rutter, and C. Lehman, "Testing the effect of computer-assisted detection on interpretive performance in screening mammography," *American Journal of Roentgenology* **187**, pp. 1475–1482, 2006.

3. C. J. D'Orsi, "Computer-aided detection: there is no free lunch," *Radiology* **221**, pp. 585–586, 2001.
4. B. Zheng, D. Chough, C. Cohen, J. H. Sumkin, G. Abrams, M. A. Ganott, L. Wallace, R. Shah, and D. Gur, "Actual versus intended use of CAD systems in the clinical environment," in *Proc. SPIE 6146*, 2006.
5. A. Malich, C. Marx, M. Facius, T. Bochm, M. Fleck, and W. A. Kaiser, "Tumour detection rate of a new commercially-available computer-aided detection system," *European Radiology* **11**, pp. 2454–2459, 2001.
6. R. F. Brem, J. W. Hoffmeister, G. Zisman, M. P. DeSimio, and S. K. Rogers, "A computer-aided detection system for evaluation of breast cancer by mammographic appearance and lesion size," *American Journal of Roentgenology* **184**, pp. 893–896, 2005.
7. T. M. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center," *Radiology* **220**, pp. 781–786, 2001.
8. L. J. Warren Burhenne, S. A. Wood, C. J. D'Orsi, and D. B. Kopans, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," *Radiology* **225**, pp. 554–562, 2000.
9. R. F. Brem, J. Baum, and M. Lechner, "Improvement in sensitivity of screening mammography with computer-aided detection: a multi-institutional trial," *American Journal of Roentgenology* **181**, pp. 687–693, 2003.
10. L. A. Khoo, P. Taylor, and R. M. Given-Wilson, "Computer-aided detection in the United Kingdom National Breast Screening Programme: a prospective study," *Radiology* **237**, pp. 444–449, 2005.
11. J. M. Ko, M. J. Nocholas, J. B. Mendel, and P. J. Slanetz, "Prospective assessment of computer-aided detection in interpretation of screening mammograms," *American Journal of Roentgenology* **187**, pp. 1483–1491, 2006.
12. M. L. Giger, Z. Huo, C. J. Vyborny, L. Lan, I. Bonta, R. M. Nishikawa, and I. Rosenbourgh, "Intelligent CAD workstation for breast imaging using similarity to known lesions and multiple visual prompt aids," in *Proc. SPIE 4684*, pp. 768–773, 2002.
13. I. El-Naqa, Y. Yang, N. P. Galatsanos, R. M. Nishikawa, and M. N. Wernick, "A similarity learning approach to content-based image retrieval: application to digital mammography," *IEEE Transactions on Medical Imaging* **23**, pp. 1233–1244, 2004.
14. B. Zheng, A. Lu, J. H. Sumkin, C. M. Hakim, M. A. Ganott, and D. Gur, "A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment," *Medical Physics* **33**, pp. 111–117, 2006.
15. G. Tourassi, B. Harrawood, S. Singh, J. Y. Lo, and C. E. Floyd, "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms," *Medical Physics* **34**(1), 2007.
16. B. Zheng, R. G. Swensson, S. Golla, C. M. Hakim, R. Shah, L. Wallace, and D. Gur, "Detection and classification performance levels of mammographic masses under different computer-aided detection cueing environments," *Academic Radiology* **11**, pp. 398–406, 2004.
17. E. Alberdi, A. Povyakalo, L. Strigini, and P. Ayton, "Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography," *Academic Radiology* **11**, pp. 909–918, 2004.
18. A. Smeulders and M. Worring, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(12), 2000.
19. B. Zheng, Y. H. Chang, and D. Gur, "Computerized detection of masses in digitized mammograms using single image segmentation and a multi-layer topographic feature analysis," *Academic Radiology* **2**, pp. 959–966, 1995.
20. J. Friedman, J. Bentley, and R. Finkel, "An algorithm for finding best matches in logarithmic expected time," *ACM Transactions on Mathematical Software* **3**(3), 1977.
21. R. Duda, P. Hart, and D. Stork, *Pattern Classification*, Wiley, 2001.
22. L. Huston, R. Sukthankar, R. Wickremesinghe, M. Satyanarayanan, G. Ganger, E. Riedel, and A. Ailamaki, "Diamond: A storage architecture for early discard in interactive search," in *Proceedings of USENIX Conference on File and Storage Technologies*, 2004.
23. A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning distance functions using equivalence relations," in *Proceedings International Conference on Machine Learning*, 2003.
24. L. Yang, R. Jin, R. Sukthankar, and Y. Liu, "An efficient algorithm for local distance metric learning," in *Proceedings of AAAI*, 2006.
25. M. Schultz and T. Joachims, "Learning a distance metric from comparisons," in *Proceedings Neural Information Processing Systems*, 2004.
26. E. Xing, A. Ng, M. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," in *Proceedings Neural Information Processing Systems*, 2003.

27. A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society* **39 (Series B)**, pp. 1–38, 1977.
28. R. Schapire, "Theoretical views of boosting and applications," in *International Conference on Algorithmic Learning Theory*, 1999.