Consensus of Ambiguity: Theory and Application of Active Learning for Biomedical Image Analysis

Scott Doyle and Anant Madabhushi[*]

Department of Biomedical Engineering,
Rutgers University, USA
scottdo@eden.rutgers.edu, anantm@rci.rutgers.edu

Abstract. Supervised classifiers require manually labeled training samples to classify unlabeled objects. Active Learning (AL) can be used to selectively label only "ambiguous" samples, ensuring that each labeled sample is maximally informative. This is invaluable in applications where manual labeling is expensive, as in medical images where annotation of specific pathologies or anatomical structures is usually only possible by an expert physician. Existing AL methods use a single definition of ambiguity, but there can be significant variation among individual methods. In this paper we present a consensus of ambiguity (CoA) approach to AL, where only samples which are consistently labeled as ambiguous across multiple AL schemes are selected for annotation. CoA-based AL uses fewer samples than Random Learning (RL) while exploiting the variance between individual AL schemes to efficiently label training sets for classifier training. We use a consensus ratio to determine the variance between AL methods, and the CoA approach is used to train classifiers for three different medical image datasets: 100 prostate histopathology images, 18 prostate DCE-MRI patient studies, and 9,000 breast histopathology regions of interest from 2 patients. We use a Probabilistic Boosting Tree (PBT) to classify each dataset as either cancer or non-cancer (prostate), or high or low grade cancer (breast). Trained is done using CoA-based AL, and is evaluated in terms of accuracy and area under the receiver operating characteristic curve (AUC). CoA training yielded between 0.01-0.05% greater performance than RL for the same training set size; approximately 5-10 more samples were required for RL to match the performance of CoA, suggesting that CoA is a more efficient training strategy.

[*] Funding provided by the Wallace H. Coulter Foundation, New Jersey Commission on Cancer Research, National Cancer Institute (R01CA136535-01, R21CA127186-01, R03CA128081-01), Cancer Institute of New Jersey, US Dept. of Defense (W81XWH-08-1-0145), Bioimagene, Inc.

T.M.H. Dijkstra et al. (Eds.): PRIB 2010, LNBI 6282, pp. 313–324, 2010.

1 Introduction

1.1 Using Consensus Methods for Certainty and Ambiguity

Ensemble classification algorithms such as bagging, boosting [1], and random forests [2] rely on some concept of consensus among several "weak" classifiers to generate a single "strong" result. Consensus, in the context of ensemble learning, describes agreement among several classification algorithms. For example, given a data object $\mathbf{x} \in \mathbb{R}^N$ belonging to one of c classes, $\omega_1, \cdots, \omega_c$, we can construct L classifiers $\mathcal{C}_l(\mathbf{x})$, for $l \in \{1, 2, \cdots, L\}$. The probability that $\mathbf{x}$ belongs to class ω_j, for $j \in \{1, 2, \cdots, c\}$, according to classifier l is denoted $p_l(\omega_j|\mathbf{x})$. While several classifier ensemble strategies seek to combine the weak learners using different rules, the underlying spirit of these methods is to assign the sample to the class ω_j for which $\arg\max_j \left[\frac{1}{L}\sum_{l=1}^{L} p_l(\omega_j|\mathbf{x})\right]$; that is, the class predicted by the majority of the classifiers. We refer to this as a consensus of certainty, and is a way of exploiting the uncorrelated variance in each of the individual classifiers.

However, in some cases it is desirable to know when there is no consensus, or more specifically when the ensemble cannot return a confident classification. Here we are not interested in knowing whether weak learners agree or disagree about the class of $\mathbf{x}$, but rather about the degree of confidence the weak learners have in assigning $\mathbf{x}$ to one of ω_j, $j \in \{1, \cdots, c\}$. The problem may be restated to ask whether $\mathbf{x}$ should belong to an "ambiguous" class or not, where ambiguousness refers to the difficulty (or lack of confidence) in classifying a sample.

1.2 Active Learning for Cost-Effective Training

Active Learning (AL) is a method of intelligently training a classifier, mitigating several drawbacks of the more standard Random Learning (RL), where samples are randomly selected for labeling [3]. RL assumes that large amounts of labeled data are already available, but for biomedical domains, manual labeling is costly and time-consuming. For example, digital images of pathology slides can be several gigabytes in size. To build a classifier to detect disease in these images, an expert pathologist needs to provide precise annotation of disease extent in the image. This results in a large training cost if RL is employed. In contrast, AL selects samples from an unlabeled pool for annotation based on the ambiguity of a sample: samples that are difficult to classify are not currently well-represented within the training set, so by targeting these samples, fewer training samples are needed to achieve high accuracy. Thus by finding only the most difficult to classify samples, we identify the most critical for labeling and inclusion in the training set.

1.3 Current Active Learning Approaches

There are several AL methods for selecting training samples [3,4,5], each relying upon a single measurement of ambiguity. The Query-By-Committee (QBC) method by Seung, et al. [5] trains a group of L weak learners, each of which votes on the class of sample $\mathbf{x}$. In the two-class case, if the sample receives approximately $\frac{L}{2}$ votes for both classes, then $\mathbf{x}$ is considered ambiguous (difficult to classify). Li, et al. [4] utilized a support-vector machine approach, whereby samples appearing close to a decision hyperplane in high-dimensional space are considered ambiguous. There is no guarantee that each of these methods will identify the same samples as "difficult to classify," since samples that are close to a decision hyperplane may still be unanimously identified as a single class by a QBC algorithm. Thus, the set of ambiguous samples may depend heavily on the AL method.

1.4 Novel Contributions of This Paper

In this paper, we present the concept of a consensus of ambiguity (CoA) whereby several measures of ambiguity are combined to identify the most difficult to classify samples from an unlabeled pool. This framework extends beyond the traditional AL methods by identifying ambiguousness explicitly rather than as a function of classification error. We define a consensus ratio that measures the degree of overlap between multiple algorithms for finding ambiguity, and we find that using multiple algorithms ensures that the overlap between methods decreases; the use of multiple algorithms ensures that only the most difficult to classify samples are detected by the algorithm.

We evaluate the efficacy of the algorithm by using the CoA-based AL method to train a probabilistic boosting tree (PBT) classifier on three separate medical image datasets. We use the performance of the PBT, measured in terms of accuracy and area under the receiver operating characteristic curve (AUC), to ensure that the training set created by CoA-based AL can yield higher performance compared to a randomly-selected training set of equal size. The three datasets considered in this work are: (1) Digitized prostate histopathology (100 images) are broken up into 12,000 image regions, each of which is classified as cancer / non-cancer using texture features. (2) 18 prostate dynamic contrast-enhanced MRI (DCE-MRI) images (256x256 pixels) are quantified using textural and functional intensity features to find cancer in a pixel-wise fashion. (3) 9,000 regions of interest (ROIs) are extracted from two large breast histopathology patient studies, with each ROI corresponding to either high or low Bloom-Richardson cancer grades. ROIs are quantified by graph-based nuclear architectural features. Each of these datasets represents different modalities, tissues, and features, but all are time-consuming and expensive to annotate; thus, we expect that AL training algorithms can reduce the expense required to obtain reliable training sets versus a random learning scheme.

2 Theory of CoA

2.1 Active Learning Strategy Overview

We denote by X a set of data containing samples $\mathbf{x} \in X$. Each sample is associated with a class label $y \in \{\omega_1, \omega_2, \cdots, \omega_c\}$. A supervised classifier is denoted

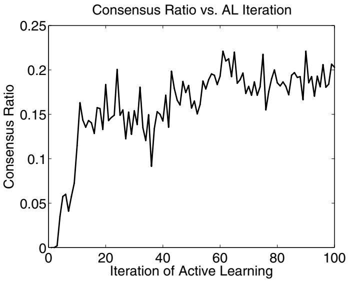


Fig. 1. Plot of the consensus ratio $\mathcal{R}$ as a function of t, for $t \in \{1, 2, \cdots, 100\}$. After $t = 50$, the consensus ratio plateaus at approximately 0.2. This indicates that there is relatively little consensus between three AL methods: Φ_1 (QBC), Φ_2 (BAY), and Φ_3 (SVD).

$\mathcal{C}(\mathbf{x}) \in \{\omega_1, \omega_2, \cdots, \omega_c\}$. The classifier returns a hypothesis for a sample and is trained on a training set S^{tr} and tested on an independent testing set. The goal of the AL algorithm is to build S^{tr} from a set of unlabeled samples in X. To do this, a training function $\Phi(\mathbf{x})$ returns a measure of ambiguity for $\mathbf{x}$.

Definition 1. *A sample $\mathbf{x} \in X$ is considered ambiguous if $a < \Phi(\mathbf{x}) < b$, where a, b are lower and upper thresholds for Φ, respectively.*

2.2 Consensus of Ambiguity: Definition and Properties

The CoA approach employs multiple algorithms, $\Phi_1, \Phi_2, \cdots, \Phi_M$, each of which returns a corresponding set of ambiguous samples $S_1^{\mathrm{E}}, S_2^{\mathrm{E}}, \cdots, S_M^{\mathrm{E}}$.

Definition 2. *Given nonempty sets of ambiguous samples, S_i^E, $i \in \{1, \cdots, M\}$, the consensus ratio is defined as $\mathcal{R} = \frac{U}{V}$, where $U = |\bigcap_{i=1}^{M} S_i^E|$ and $V = |\bigcup_{i=1}^{M} S_i^E|$.*

Proposition 1. *Given nonempty sets of ambiguous samples, S_i^E, where $i \in \{1, \cdots, M\}$, $\mathcal{R} = 1$ indicates perfect consensus and $\mathcal{R} = 0$ indicates no consensus across Φ_i.*

Proof. In the case of absolutely no consensus (i.e. no samples are considered ambiguous by all M algorithms), then $\bigcap_{i=1}^{M} S_i^{\mathrm{E}} = \emptyset$, so $\mathcal{R} = 0$. Conversely, when Φ_i, $i \in \{1, \cdots, M\}$ are in perfect agreement (every algorithm identifies exactly the same samples as ambiguous), then $S_1^{\mathrm{E}} = \cdots = S_M^{\mathrm{E}}$, so $\bigcap_{i=1}^{M} S_i^{\mathrm{E}} = \bigcup_{i=1}^{M} S_i^{\mathrm{E}}$ and $\mathcal{R} = 1$. □

Property 1. When $\mathcal{R} \approx 0$, there is low consensus and high variance among Φ_i, $i \in \{1, \cdots, M\}$, indicating that any agreement among the algorithms will be highly informative and suggesting a benefit to using a consensus approach. Figure 1 shows a graph of $\mathcal{R}$ as a function of t, which identifies the iterations of the AL algorithm. Beginning with $t = 0$, the AL algorithm grows a training set by selecting and labeling ambiguous samples and adding them to the training set. The process iterates for $t \in \{1, \cdots, 100\}$ times in this experiment. Three different AL algorithms were used: QBC, BAY, and SVD (Section 3.2). After 50 iterations, $\mathcal{R}$ levels off at approximately 0.2, indicating that there is little consensus among the methods. Thus, a consensus algorithm is likely to be informative.

Definition 3. *A sample* $\mathbf{x} \in X$ *will be considered strongly ambiguous if* $\mathbf{x} \in \widehat{S}^E = \bigcap_{i=1}^M S_i^E$*; that is, if the sample is designated as ambiguous by all* Φ_i *for* $i \in \{1, \cdots, M\}$*.*

Definition 3 is a version of strong ambiguity wherein all M algorithms must select the sample. It is possible that, on any particular AL iteration, no samples will satisfy this criteria. Definition 3 can easily be modified to include samples selected by a majority of algorithms, or any sample identified by more than one algorithm, and so on.

Proposition 2. *As the number of algorithms* Φ_i*,* $i \in \{1, \cdots, M\}$*, being combined increases, the consensus ratio* $\mathcal{R}$ *will monotonically decrease.*

Proof. An added algorithm, denoted Φ_{M+1}, identifies a set of samples denoted S_{M+1}^{E}. If S_{M+1}^{E} is a subset of the current set of ambiguous samples, $\bigcup_{i=1}^M S_i^{\mathrm{E}}$, then the denominator of $\mathcal{R}$ does not change since the union will not increase in size. The denominator of $\mathcal{R}$ will decrease, since any elements in $\bigcap_{i=1}^M S_i^{\mathrm{E}}$ that are not found in S_{M+1}^{E} will be removed in the new intersection, $\bigcap_{i=1}^{M+1} S_i^{\mathrm{E}}$. Thus $\mathcal{R}$ will decrease in value.

However, if S_{M+1}^{E} contains unique samples not in the current ambiguous sample set, the union will increase in size; that is, $|\bigcup_{i=1}^M S_i^{\mathrm{E}}| < |\bigcup_{i=1}^{M+1} S_i^{\mathrm{E}}|$. Thus the denominator of $\mathcal{R}$ will increase. The numerator of $\mathcal{R}$ will not change, since any samples in S_{M+1}^{E} that are not in $\bigcap_{i=1}^M S_i^{\mathrm{E}}$ will be removed in the new intersection, $\bigcap_{i=1}^{M+1} S_i^{\mathrm{E}}$. In this case, $\mathcal{R}$ will decrease. □

Property 2. Adding additional algorithms to the ensemble, will decrease or maintain $\mathcal{R}$. By Property 1, ensembles with a low consensus ratio $\mathcal{R}$ ensure that only samples with a very high degree of ambiguity will be identified. Thus increasing M will ensure that only extremely ambiguous samples are included in $\widehat{S}^{\mathrm{E}}$. However, if $S_{M+1}^{\mathrm{E}} \cap \widehat{S}^{\mathrm{E}} = \emptyset$, then no samples will be considered strongly ambiguous.

3 Experimental Setup

3.1 Overview of Datasets

Experiment 1 - Prostate cancer on digitized histopathology: Over a million annual prostate biopsies are performed in the US, each of which must

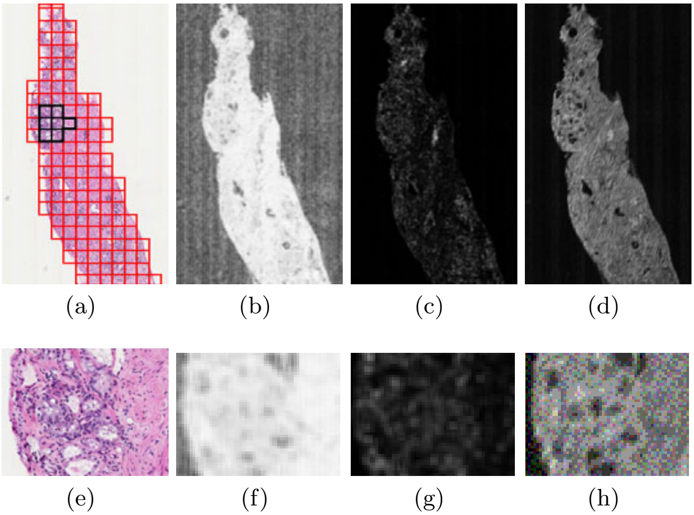


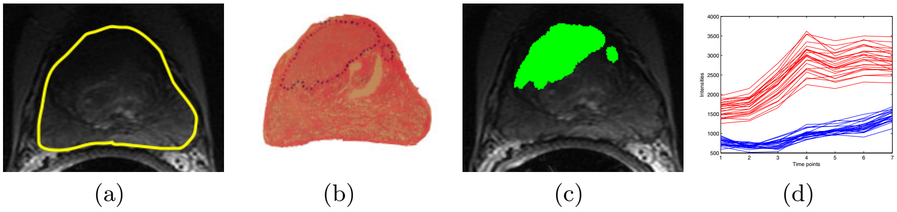
Fig. 2. Image data from Experiment 1. The original image (a) has a red 30-pixel square grid superimposed, with cancer labeled in black. Texture images are extracted corresponding to first-order greylevel statistics (b), second-order Haralick co-occurrence features (c), and Gabor steerable filter features (d). Shown in the second row (e)-(h) are magnified regions of the cancer region in each image.

be analyzed manually under a microscope [7]. A quantitative system capable of automatically detecting disease can greatly increase the speed and accuracy with which patients are diagnosed for cancer. Digitized glass slides can be over 2 GB in size (several million pixels), with benign and cancer regions appearing close to one another, and so annotation of these samples is difficult. The objective of this experiment is to apply CoA-based AL to build a classifier able to distinguish between cancerous and non-cancerous patches of biopsy tissue.

Biopsy samples are stained with Hematoxylin and Eosin (H & E) to visualize cell cytoplasm and nuclei and digitized using a whole-slide digital scanner. For each image, a 30x30 pixel grid is superimposed on the tissue, generating regions of interest (ROIs) of prostate tissue. In previous work [8], we have identified 14 texture features that can easily distinguish between cancer and non-cancer regions of tissue on a pixel-wise basis. These features include: (1) First-order gray-level statistics quantify simple statistics calculated from pixel values in the images [8]. (2) Second-order Haralick features [9] are based on the co-occurrence of pixel values, and are calculated over each ROI. (3) Gabor filter features, also known as steerable filters, operate at a specific orientation and spatial frequency to yield a filter response from the image. Each of the 14 discriminating features is extracted from the image, and the modal value for each 30-by-30 ROI is used as its feature value. 100 images are used to generate 12,000 ROIs which are classified as cancer or non-cancer tissue.

Experiment 2 - Prostate cancer on DCE-MRI: In addition to biopsy, *in vivo* imaging, particularly magnetic resonance imaging (MRI), can be mined for quantitative diagnostic information [10,11]. Dynamic Contrast Enhanced (DCE) MRI is a technique whereby a contrast agent is injected into a patient with MR images taken at specific time points. The contrast agent is taken up and removed from different tissues at different rates, indicating the presence of disease at a pixel-wise level. A classification system for this modality could be used for automated *in vivo* screening for cancer and treatment, but labeled samples are difficult to obtain since cancer cannot be annotated directly on the MRI. Histopathology is used to find cancer ground truth, which is mapped onto the MR images.

We apply CoA-based AL to a dataset of 6 patients with confirmed prostate cancer on needle biopsies. Prior to radical prostatectomy, MR imaging was performed using an endorectal coil in the axial plane and included T2-w and DCE protocols. Prostatectomy specimens were later sectioned and stained with H & E. An expert pathologist annotated the spatial extent of prostate cancer on the whole-mount prostatectomy sections, and identified 18 corresponding histopathology and MRI sections. A multimodal registration scheme, COLLection of Image-derived Non-linear Attributes for Registration Using Splines (COLLINARUS) [12], was used to register histology sections onto the corresponding MRI data, thus mapping the cancer ground truth onto the MR images. Structural information from T2-w MRI and functional intensity information from DCE MRI are combined to distinguish between cancer and non-cancer pixels.



(a) (b) (c) (d)

Fig. 3. Examples of data from Experiment 2. Shown are (a) T2-w MRI image with the prostate boundary in yellow, (b) the corresponding histopathology slice with cancer mapped in blue, and (c) the cancer extent mapped onto the T2-w MRI after registration via COLLINARUS [12]. Also shown are (d) intensity vs. time curves for dynamic contrast; blue curves represent pixel locations in benign tissues, while red curves are inside cancer ground truth ((c)).

Experiment 3 - Breast cancer on digitized histopathology: Breast cancer is the second-leading cause of cancer death in women in the United States [7]. Mammogram screening followed by a biopsy is the current standard for definitive diagnosis. Similar to the motivation in Experiment 1, an automated image analysis system can assist pathologists in detecting and diagnosing breast cancer.

Images of H & E stained breast biopsies are classified between low and high Bloom-Richardson grades of breast cancer. Two patient studies were used to

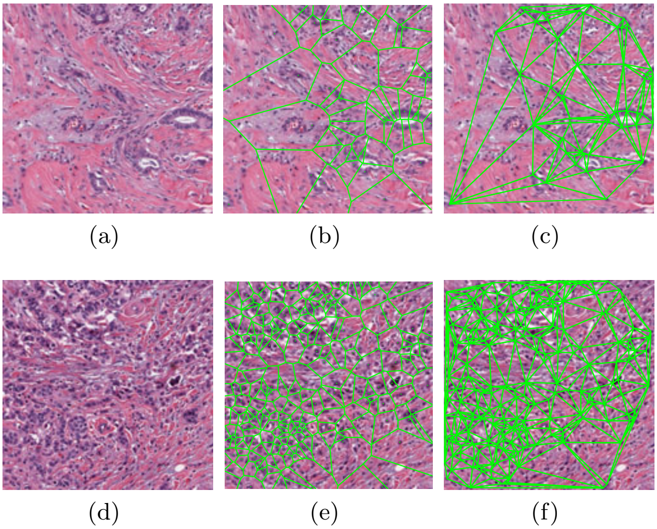


Fig. 4. Examples of image data from Experiment 3, where we distinguish low-grade breast cancer tissue ((a)-(c)) from high-grade tissue ((d)-(f)). Nuclei are detected from breast biopsy tissue (a), (d) and used to generate graphs such as the Voronoi tesselation (b), (e) and Delaunay triangulation (c), (f). Features from these graphs are used to quantify each image patch.

generate 9,000 ROIs of homogeneous tissue measuring 500x500 pixels each. We calculate features based on the architecture of the cell nuclei, in accordance with the major indicators of breast cancer grade. Color deconvolution is used to transform the RGB color space of the image into an alternate three-color space to separate out the hematoxylin, eosin, and white background of the image [13]. Using the deconvoluted image, the centroids of cell nuclei are detected, which are used to construct a series of graphs based on the Voronoi tesselation, Delaunay triangulation, and a minimum spanning tree. From each of these, a set of quantitative features is extracted to characterize the cell architecture [13]. Each ROI is classified as high or low Bloom-Richardson grades of cancer, where ground truth is determined by a pathologist.

3.2 Comparison of AL Methods

Query-By-Committee (QBC): QBC [5] involves a group of L weak classifiers that produce votes for the class of an unlabeled sample $\mathbf{x}$. Samples with approximately $\frac{L}{2}$ votes are considered difficult to classify. The output of $\Phi_1(\mathbf{x})$ is the number of votes for the target class, and a, b represent the minimum and maximum votes, respectively. A total of $L = 10$ Random Forests were generated using C4.5 decision trees [2,1] with threshold values of $a = 4$ and $b = 6$.

Bayes Likelihood (BAY): Bayes Theorem [14] models the likelihood of observing a class based on the feature values of sample $\mathbf{x}$. A probability density function is created for each of K features, where $p_k(\omega_j|\mathbf{x})$ denotes the likelihood that $\mathbf{x}$ belongs to class ω_j given feature k. Samples for which $p_k(\omega_j|\mathbf{x}) \approx 0.5$ are considered ambiguous. The output of $\Phi_2(\mathbf{x})$ is $\frac{1}{K}\sum_{k=1}^{K} p_k(\omega_1|\mathbf{x})$ where ω_1 is the target (cancer) class. Threshold parameters were set to $a = 0.4$ and $b = 0.6$.

Support Vector Distance (SVD): Support Vector Machines (SVMs) [15] create a high-dimensional projection of feature data, in which a decision hyperplane is created via training. Samples are classified by finding the position relative to the hyperplane. The output of $\Phi_3(\mathbf{x})$ is the signed distance between $\mathbf{x}$ and the hyperplane, where the sign indicates class membership. Parameters a and b define the distances within which a sample is considered ambiguous. We set a and b to ±10% of the maximum distance from the support vector.

3.3 Probabilistic Boosting Tree Classification Algorithm

CoA-based AL was used to train a probabilistic boosting tree (PBT) [16]. The PBT combines AdaBoost [17] and decision trees [1], iteratively generating a tree where each node is boosted with L weak classifiers and whose output is a likelihood for the class of sample $\mathbf{x}$. The PBT algorithm was chosen as a classifier that is different from the methods used in each of the AL algorithms described above. At each iteration of the active learning algorithm, $t \in \{1, 2, \cdots, 100\}$, ambiguous samples found by the CoA ensemble are sampled to obtain equal numbers of samples from both classes [6], which are used to train the PBT. For our experiments, each iteration added two samples (one from each class) to the growing training set. Evaluation on an independent testing set is done via area under the receiver operating characteristic curve (AUC) and accuracy.

4 Results and Discussion

Shown in Figure 5 are examples of two datasets, prostate histopathology (top row) and DCE-MRI (bottom row), used in this study. In the left column (Figures 5 (a), (d)) are the original images with the cancerous region delineated in a black contour, while the results of classification with RL training are shown in the middle column (Figures 5 (b), (e)) and training with CoA-based AL are shown in the right column (Figures 5 (c), (f)). The images were obtained when the AL algorithm had run for $t = 50$ iterations.

For histopathology, brighter regions indicate higher likelihood of cancer. The RL-trained classifier identifies the majority of patches as cancer yielding a high false-positive count, while the CoA-trained classifier is able to discriminate between obviously benign regions and cancerous areas. Note that we are not commenting here on the accuracy of the final classifier, but on the performance of

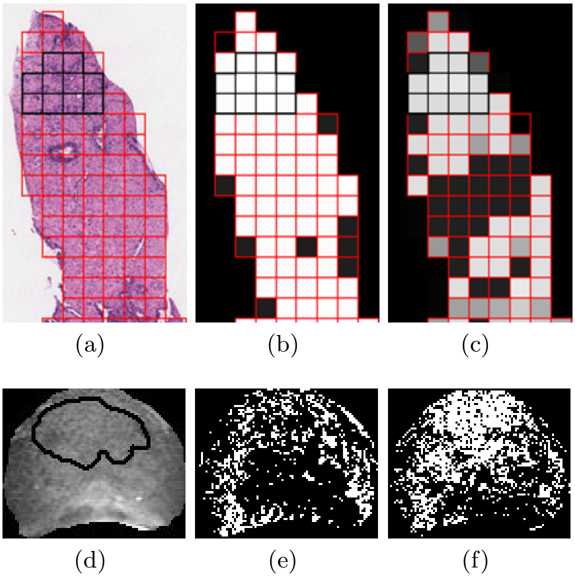


Fig. 5. Examples of images taken from the prostate histopathology (a) and DCE-MRI (d) datasets, with cancer regions indicated by black contours. Also shown are the corresponding classification results of the PBT, when using training sets built via RL ((b), (e)) and CoA-based AL ((c), (f)). Images were obtained at AL iteration $t = 50$.

one training method with respect to another. For the DCE images, images were thresholded at a likelihood of 75%. Here, the RL-trained classifier yields false-negatives with a small set of pixels classified as cancer, while the CoA-trained classifier correctly classifies many pixels near the ground truth. Again, this indicates that – given the limitations on labeling biomedical images – CoA yields better results than random training on a limited number of training samples.

The accuracy and AUC of the PBT are plotted against the AL iteration $t \in \{1, \cdots, 100\}$ in Figure 6. Shown are the results for the classifier trained using the CoA algorithm (red solid) as well as random learning (blue dotted) and each of the three AL strategies: QBC (green dot), BAY (cyan solid), and SVM (magenta dash). Each location on the independent axis indicates a training set size (increasing from left to right); we can see that for the majority of training set sizes, all of the AL-trained classifiers yield better accuracy and AUC than random learning. Additionally, AL requires fewer samples to reach that desired performance compared with RL. We note that the individual AL algorithms do not necessarily perform better than the CoA approach in terms of classifier performance, but this is not an unexpected result. The goal of using the CoA algorithm is to prune down the number of samples deemed "eligible" at each stage; we see that by constraining our search in this way, we have a smaller pool from which to choose labeled samples, while keeping performance the same as an individual algorithm (which has a much wider set of "eligible" samples).

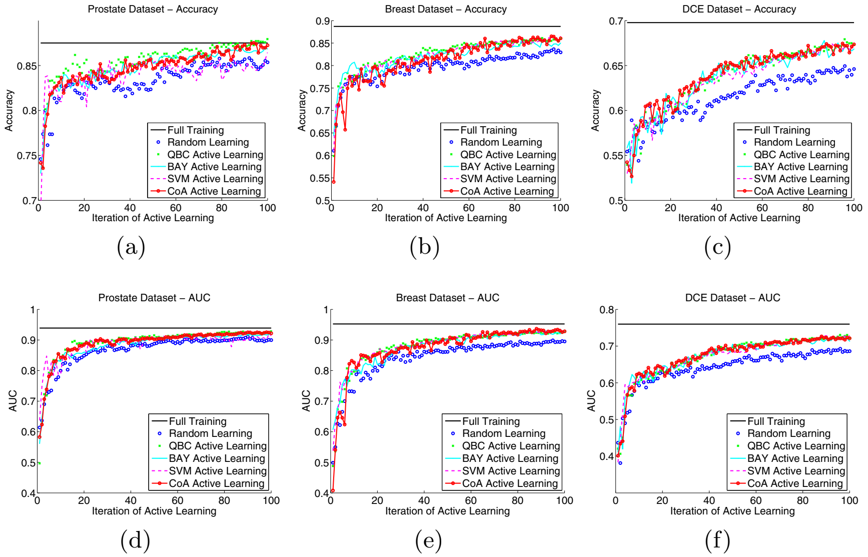


Fig. 6. Plots of the accuracy and AUC obtained by the PBT using the training derived from CoA Active Learning method (red solid line), which combines three AL schemes (QBC, BAY, and SVD), and Random Learning (blue dotted line). Shown are results for the dataset of 12,000 prostate histopathology ROIs ((a), (d)), 28,000 prostate DCE-MRI pixel samples ((c), (f)), and 9,000 breast histopathology ROIs ((b), (e)).

5 Concluding Remarks

In this paper, we presented a CoA framework for identifying ambiguousness in an unlabeled pool of data. The CoA approach exploits variance between different ambiguity measurements. A consensus ratio determines the amount of variance between multiple ambiguity methods, and by combining these algorithms, this ratio decreases. This ensures that only the most ambiguous samples are selected from the unlabeled data. Finally, we applied CoA to the problem of Active Learning (AL), where ambiguous samples are selected for training a classifier. For medical image datasets (which are time-consuming and expensive to annotate), the CoA-trained classifier yields higher accuracy and AUC than RL for similar training set sizes.

We observe similar classification performance using CoA versus individual AL training schemes. However, the low consensus ratio indicates that each training algorithm is selecting mostly unique samples. Since our goal is to improve training efficiency, we wish to explore evaluation measures besides classifier performance. For example, it is possible that samples selected by one AL scheme are more difficult to annotate than those selected by another, or have significantly different feature distributions. If so, we may be able to derive an evaluation metric that is divorced from classifier performance that is able to identify the most efficient training algorithm.

References

1. Quinlan, J.R.: Decision trees and decision-making. IEEE Trans. Syst. Man Cybern. 20(2), 339–346 (1990)
2. Brieman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
3. Cohn, D., Ghahramani, Z., Jordan, M.I.: Active Learning with Statistical Models. J. of Art. Intel. Res. (4), 129–145 (1996)
4. Li, M., Sethi, I.K.: Confidence-based active learning. IEEE Trans. Patt. Anal. Mach. Intel. 28(8), 1251–1261 (2006)
5. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: 5th Annual ACM Workshop on Computational Learning Theory, pp. 287–294. ACM, New York (1992)
6. Doyle, S., Madabhushi, A., Feldman, M., Tomaszewski, J., Monaco, J.: A Class Balanced Active Learning Scheme that Accounts for Minority Class Problems: Applications to Histopathology. In: OPTIMHisE Workshiop (in conjunction with MICCAI), pp. 19–30 (2009)
7. American Cancer Society. Cancer Facts & Figures 2010. American Cancer Society, Atlanta (2010)
8. Doyle, S., Feldman, M., Tomaszewski, J., Madabhushi, A.: A Boosted Bayesian Multi-Resolution Classifier for Prostate Cancer Detection from Digitized Needle Biopsies. IEEE Transactions on Biomedical Engineering (accepted)
9. Haralick, R.M., Shanmugan, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics SMC 3, 610–621 (1973)
10. Madabhushi, A.: Digital Pathology Image Analysis: Opportunities and Challenges. Imaging in Medicine 1(1), 7–10 (2009)
11. Viswanath, S., Bloch, B.N., Rosen, M., Chappelow, J., Rofsky, N., Lenkinski, R., Genega, E., Kalyanpur, A., Madabhushi, A.: Integrating Structural and Functional Imaging for Computer Assisted Detection of Prostate Cancer on Multi-Protocol in vivo 3 Tesla MRI. In: SPIE Medical Imaging, vol. 7260 (2009)
12. Chappelow, J., Madabhushi, A., Bloch, B.: COLLINARUS: Collection of image-derived non-linear attributes for registration using splines. In: Proc. SPIE: Image Processing, vol. 7259, San Diego, CA, USA (2009)
13. Basavanhally, A.N., Ganesan, S., Agner, S., Monaco, J., Feldman, M., Tomaszewski, J., Bhanot, G., Madabhushi, A.: Computerized Image-Based Detection and Grading of Lymphocytic Infiltration in HER2+ Breast Cancer Histopathology. IEEE Transactions on Biomedical Engineering 57(3), 642–653 (2010)
14. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley Interscience, New York (2001)
15. Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning 20, 273–297 (1995)
16. Tu, Z.: Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. In: 10th IEEE International Conference on Computer Vision, pp. 1589–1596 (2005)
17. Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm. In: 13th International Conference on Machine Learning, pp. 148–156 (1996)