Distance Guided Selection of the Best Base Classifier in an Ensemble with Application to Cervigram Image Segmentation

Wei Wang and Xiaolei Huang
Department of Computer Science and Engineering
Lehigh University, Bethlehem, PA 18015
wew305@lehigh.edu, xih206@lehigh.edu

Abstract

We empirically evaluate a distance-guided learning method embedded in a multiple classifier system (MCS) for tissue segmentation in optical images of the uterine cervix. Instead of combining multiple base classifiers as in traditional ensemble methods, we propose a Bhattacharyya distance based metric for measuring the similarity in decision boundary shapes *between a pair of statistical classifiers. By generating an ensemble of base classifiers trained independently on separate training images, we can use the distance metric to select those classifiers in the ensemble whose decision boundaries are similar to that of an unknown test image. In an extreme case, we select the base classifier with the* most similar *decision boundary to accomplish classification and segmentation on the test image. Our approach is novel in the way that the* nearest neighbor *is picked and effectively solves classification problems in which base classifiers with good overall performance are not easy to construct due to a large variation in the training examples. In our experiments, we applied our method and several popular ensemble methods to segmenting acetowhite regions in cervical images. The overall classification accuracy of the proposed method is significantly better than that of a single classifier learned using the entire training set, and is also superior to other ensemble methods including majority voting, STAPLE, Boosting and Bagging.*

1. Introduction

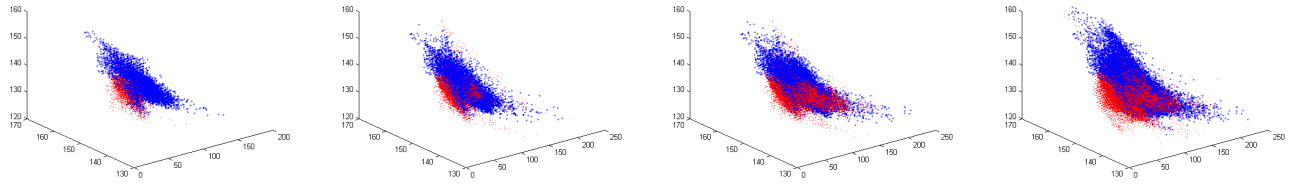
Reliable segmentation and labeling of different regions in images are important to make images searchable by content in large medical image archives. In this work, we consider the task of automatically segmenting the biomarker AcetoWhite (AW) regions in an archive of 60,000 images of the uterine cervix. These images are optical cervigram images acquired by Cervicography using specially-designed cameras for visual screening of the cervix, and they were collected from the NCI Guanacaste project [6] for the study of visual features correlated to the development of precancerous lesions. The most important observation in a cervigram image is the AW region, which is caused by whitening of potentially malignant regions of the cervix epitheliuem, following application of acetic acid to the cervix surface. Since the texture, size and location of AW regions have been shown to correlate with the pathologic grade of disease severity, accurate identification and segmentation of AW regions in cervigrams have significant implications for diagnosis and grading of cervical lesions.

Accurate tissue segmentation in cervigrams is a challenging problem due to large variations in image appearance. Illumination, specular reflection, and color changes caused by pathology all contribute to such appearance variations. As a result, the color and texture feature distributions of a tissue class in one image often overlap with those of a different tissue class in other images. Figure 1 demonstrates this problem by displaying Acetowhite (AW) and Squamous Epithelium (SE) feature sample distributions from a varied number of images.

Previous work on cervigram segmentation has reported limited success using K-means clustering [17], Gaussian Mixture Models [4], Support Vector Machine (SVM) classifiers [5]. Shape priors are also proposed [8] although such priors are applicable to cervix boundary but not to other important region boundaries such as AW since AW regions could be of arbitrary shape. Supervised learning based segmentation [14, 13] holds promise, especially with increasing number of features. However, due to the intrinsic diversity between images and the overlap between feature distributions of different classes, it is difficult to learn a single classifier that can perform tissue classification with low error for a large image set. Our empirical evaluation shows that overfitting is a serious problem when training a single (SVM) classifier using all training examples; the average sensitivity of the classifier is as low as 8%.

A potential solution is to use a Multiple Classifier System (MCS) [12, 1], which trains a set of diverse classifiers

978-1-4244-3993-5/09/$25.00 ©2009 IEEE



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

Figure 1. AW vs. Non-AW Cervix color sample distributions in L*a*b* space. (a) samples from one image. (b) samples from two images. (c) samples from three images. (d) samples from six images. (red) AW color samples. (blue) Cervix color samples.

that disagree on their predictions and effectively combines the predictions in order to reduce classification error. In MCS, a wide variety of classifier ensemble methods, including error-correcting output coding [10], Bagging, and Boosting [15], have been proposed with demonstrated success in reducing variance and bias. These methods differ in the way an ensemble of base classifiers is formed, by modifying the data, the learning task, or by exploiting algorithm characteristics such as randomized components and tree structures. For combining predictions, average voting, weighted voting, and stacking [9] are commonly used. Previous study shows that with a proper ensemble strategy, combining base classifiers could perform comparably with or even better than selecting the best base classifier from the ensemble by cross validation [16]. A necessary condition for the above ensemble methods is that all the base classifiers should provide sufficiently good performance, usually 50% or higher sensitivity and specificity in order to support the ensemble [16]. In the cervigram segmentation problem, a natural way of constructing base classifiers is to use each training image to train a base classifier that distinguishes between different region types in the image: *AW*, *SE*, *Columnar Epithelium (CE)*, *specular (SR)*, *OS*, and *Others*. However, this creates a problem because we have observed a large variance in base classifier performance for classifying tissue in a test image: While the best base classifier from the ensemble always gives good performance, many others—even those learned using SVM, AdaBoost, and bagging—commonly have low sensitivity (30% or lower; see Tables 1, 4).

Thus, instead of using traditional ensemble methods, we propose a novel and effective way which emphasizes on selecting and combining the best base classifiers. In an extreme case, we predict the best single base classifier. The novel contribution is in our procedure to seek the *best* classifier based on class-dependent information extracted from the test image. Our assumption is that, a good base classifier that performs well on the test image should have similar decision boundary as the classifier learned from the test image itself (if the ground truth segmentation of the test image were known). We characterize the similarity in decision boundary by measuring the distance between a training and the test images' feature distributions of the same tissue class. In particular, we apply the Bhattacharyya distance [7], to measure the distance between testing and training distributions. The base classifier learned using the training example with the smallest distance from the test image is then selected as the best base classifier. More generally, we can use the distance measure to rank the base classifiers and combine the top-K (e.g.$K = 3$ or $K = 5$) in a MCS.

Our experimental results have shown very impressive performance by this new approach. Our estimated best classifier performs very close to the actual best-performing base classifier. The accuracy is far better than that of a single SVM classifier learned using all training data [5], and also better than classifier ensemble methods such as those reported in [1] and [18].

2. Methodology

The basic idea of our approach is simple: Given a test image, we seek to find the best base classifier in an ensemble. We train base classifiers by learning a multi-label, linear-kernel SVM classifier from each training image example. [1] Therefore, given N training images, we will have N base classifiers.

To perform segmentation on a test image, we first apply all N base classifiers to classifying tissue on the test image. We then seek to estimate the best base classifier, which is the most *similar in terms of classification decision boundary* to the classifier that could have been learned from the test image itself if the ground truth segmentation were known. Obviously this similarity can not be measured straightforwardly since we do not know the ground truth for the test image. We address this problem by taking advantage of the observation that, if two classifiers are similar, then after performing classification, their resulting class-dependent feature distributions will be similar.

[1] The choice of multi-label, linear-kernel for SVM was based on empirical evaluation which shows such SVMs have comparable or slightly better performance than 2-label, RBF or Polynomial kernels, on cervigram tissue classification [1].

Without loss of generality, let us consider two base classifiers, C_1 and C_2, learned from two training images T_1 and T_2, respectively. Suppose each base classifier has m classes of labels $y \in \{y_1, y_2, ...y_m\}$. Let us denote the distribution of features, x (e.g. color, texture), of the jth class in example T_1 by $p^j_{T_1}(x)$. Similarly, the feature distribution of the jth class in example T_2 is denoted by $p^j_{T_2}(x)$.

On a test image I, we apply both C_1 and C_2 to classifying each pixel to have one of the m labels based on the pixel's feature value. Therefore we have the class-dependent feature distributions in I according to C_1: $p_I^{j(C_1)}(x), j = 1, ..., m$ and those according to C_2: $p_I^{j(C_2)}(x)$, where $j(C_i)$ means the jth class according to classifier C_i. Our conjecture is that, if C_1 is a better classifier than C_2 for image I, which implies T_1 is more similar to I in terms of tissue classification decision boundary, then we have for any class j,

$$\mathcal{D}(p_I^{j(C_1)}(x), p^j_{T_1}(x)) < \mathcal{D}(p_I^{j(C_2)}(x), p^j_{T_2}(x)) \quad (1)$$

where $\mathcal{D}$ is a distance measure between two probability distributions. Therefore we can use such distances between class-dependent feature distributions of a training image and the test image (after classification using the base classifier learned from the training image) to determine the dissimilarity between the training and the test image. A ranking of the base classifiers can also be obtained according to the dissimilarity measure. Figure 2 shows the system flow of the proposed method.

2.1. Similarity measurement

To measure the dissimilarity between two probability density functions, we adopt an information-theoretic distance measure, the Chernoff Information. It has been shown this measure is the exponential rates of optimal classifier performance probabilities [2]. The Chernoff Information between p_1 and p_2 is defined by:

$$C(p_2||p_1) = \max_{0 \le t \le 1} -\log \mu(t) \quad (2)$$

where $\mu(t) = \int [p_1(x)]^{1-t} [p_2(x)]^t \, dx$. A special case of Chernoff distance is the Bhattacharyya distance, in which t is chosen to be $\frac{1}{2}$, i.e., the Bhattacharyya distance between p_1 and p_2 is:

$$B(p_2||p_1) = -\log \mu(\frac{1}{2}) \quad (3)$$

In order to facilitate notation, we write:

$$\rho(p_2||p_1) = \mu(\frac{1}{2}) = \int [p_1(x)]^{\frac{1}{2}} [p_2(x)]^{\frac{1}{2}} \, dx \quad (4)$$

Clearly, when the value for ρ ranges from one to zero, the value for the Bhattacharyya distance B goes from zero to infinity.

2.2. Best model estimation

In our work, we have the N base classifiers, $C_i, i = 1, ...N$, each learned from features and their corresponding class labels in a training image T_i. We apply all N classifiers to a test image I to acquire N different classification results. Denote the feature distribution of the jth class in the ith training image T_i by $p_i^j = p^j_{T_i}(x)$. And write the feature distribution of the jth class in the test image according to the classification result by C_i as $q_i^j = p_I^{j(C_i)}(x)$. We compute the following Bhattacharyya distances between training images' feature distributions and the test image's feature distributions: $\mathcal{D}(q_i^j, p_i^j) = B(q_i^j||p_i^j)$, for all $i = 1, ...N$ and all $j = 1, ..., m$.

We define the cost function of any base classifier C_i on the test image as:

$$E(C_i) = \left\| [B(q_i^1||p_i^1), B(q_i^2||p_i^2), ..., B(q_i^m||p_i^m)] \right\|_2 \quad (5)$$

where $[B(q_i^1||p_i^1), B(q_i^2||p_i^2), ..., B(q_i^m||p_i^m)]$ is a distance vector that consists of the Bhattacharyya distances between feature distributions of the test image I and those of the training image T_i for all m classes.

Therefore we seek the best base classifier (or model) for the test image:

$$\tilde{C} = C_{\tilde{i}}, \quad (6)$$

where

$$\tilde{i} = \arg\min_i \left\| [B(q_i^1||p_i^1), B(q_i^2||p_i^2), ..., B(q_i^M||p_i^M)] \right\|_2 . \quad (7)$$

And the classification result by $\tilde{C}$ is chosen as the final classification for test image I.

We can also rank all the base classifiers based on their cost function values (Eq. 5), and select the top-K models to be used in a multiple classifier system to derive the final result. For instance, a user interface can be developed to display segmentation results on the test image based on classifications from the top-K models. Either the user can manually pick one as the final segmentation, or the top few segmentations can be combined using an ensemble method such as majority voting or STAPLE [18].

3. Experimental Results

We implemented our algorithm in Matlab 2007b on a computer with Intel Core2 E6850 CPU. 939 cervigram images from the NCI/NLM archive with multiple-expert boundary markings are available for training and validation purposes. We used 100 images of diverse appearance for training and testing. 50 randomly selected images are used for training and the remaining 50 are used for testing and validation. One multi-label, linear kernel SVM base classifier is learned based on color feature samples (in L*a*b* color space) and their tissue class labels in each training

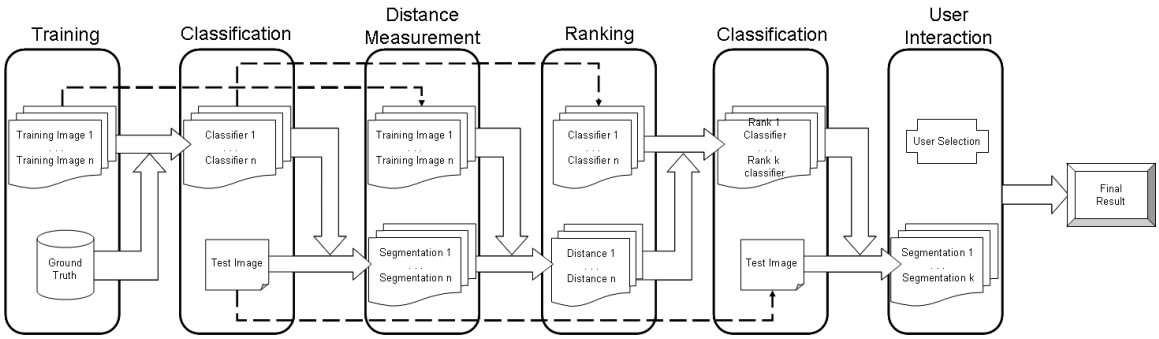


Figure 2. Overview diagram of the proposed cervigram segmentation system.

image; the multi-label SVM classifier can simultaneously segment several important tissue regions in cervigrams, including the AW, Squamous Epithelium (SE) and Columnar Epithelium (CE).

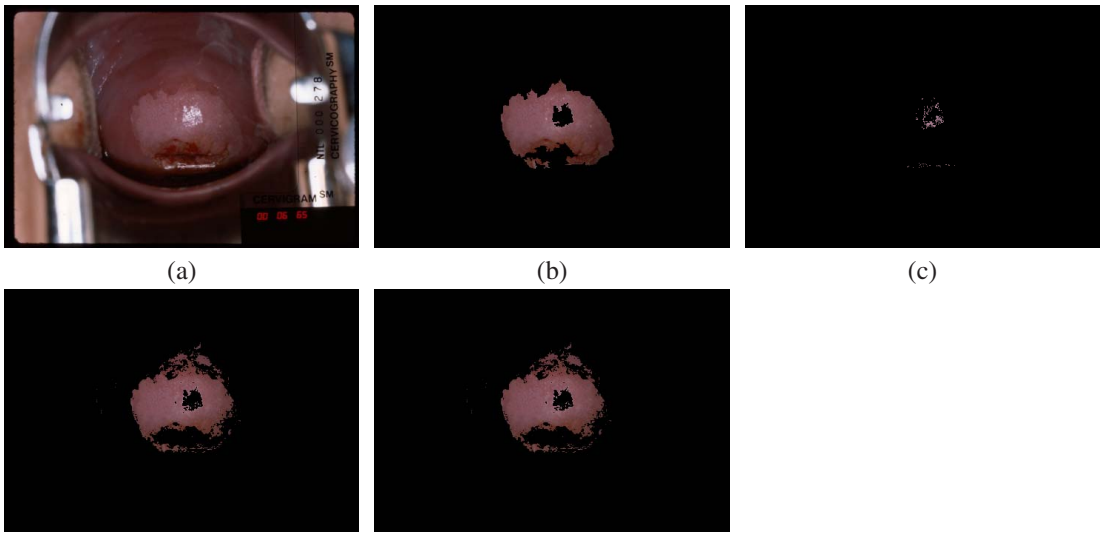
Given a test image, we apply all the base classifiers to label pixels in the image. Utilizing the classification results, we compute the cost function of each classifier based on measuring the class-dependent Bhattacharyya distances between the test image's feature distributions and the classifier's corresponding training image's feature distributions (Eq. 5). We consider two tissue classes in computing the distances ($m = 2$) in Eq. 5: AW and SE. The best classifier $\tilde{C}$ is selected as the one with minimum cost (Eq. 6). We also obtain a ranking of the base classifiers in ascending order of the cost function value so that segmentation results from the top-K models can be presented to the user.

We evaluate our novel class-dependent *nearest neighbor* approach by comparing its performance with those of: (1) training a single SVM classifier using all training data, (2) another *nearest neighbor* approach based on direct class-independent dissimilarity between test image and training image, (3) other ensemble methods including majority voting [9], STAPLE [18, 1], Boosting [3, 14], and Bagging [15].

In the first experiment, we trained an overall multi-label, linear SVM classifier using feature samples from all 50 training images. We then compared results of this overall single SVM classifier with the results by our distance-guided classification method which selects the best base classifier based on one *nearest neighbor* training image. The performance measure comparison on the validation image set of 50 images is shown in Table 1. The mean and standard deviation of p and q (sensitivity and specificity) and dice similarity coefficient (DSC) [11] are computed for each test image. Figure 3 also visually demonstrates such comparison on a test image. One can see that the proposed method provides much better segmentation results than the overall classifier. The overall classifier has unacceptably low sensitivity and DSC, which could be because of overfitting. Our distance-guided approach estimates the best base classifier based on a training image that has the *closest classification decision boundary* as the test image, and it achieves a performance level significantly better than the overall classifier, and similar to that of the actual best base classifier (Table 1).

In the second experiment, we compared our class-dependent distance measure and cost strategy with a direct class-independent dissimilarity measure. For the class-independent measure, we simply compute the Bhattacharyya distance between the distribution of each training image's features (regardless of feature labels) and that of the test image's features. The *nearest neighbor* model image then is selected as the training image that has the smallest distance from the test image and this model image is used to train a multi-label, linear SVM classifier which is applied to classifying the test image. This class-independent distance is a direct measure of overall training-testing image appearance dissimilarity. In contrast, our class-dependent distance measures dissimilarity in tissue classification decision boundary. The experimental comparison between the two are shown in Table 2 and Figure 4. From Table 2, one can see that our proposed class-dependent measure gives significantly better segmentation results ($\bar{p} = 0.7361, \bar{q} = 0.8304, \overline{DSC} = 0.5822$) than the direct appearance dissimilarity measure ($\bar{p} = 0.3443, \bar{q} = 0.8233, \overline{DSC} = 0.2366$). The comparison can also be seen in Figure 4. Figure 4(a) displays the difference between mean DSC (on the 50 validation test images) of the actual best-performing base classifier and (1) mean DSC of the estimated best classifier using our proposed approach, (2) average of mean DSCs of top-ranked classifiers by our approach. Figure 4(b) displays similar kinds of DSC differences, but the best and top-ranked base classifiers are learned using training images selected according to the direct class-independent image dissimilarity measure.

To further improve performance, instead of selecting the estimated best model, we can present the top-K (e.g. $K = 3$ or $K = 5$) models' results to the user and allow the user to interactively pick the final result. Table 3 shows the performance comparison with and without user



(a) (b) (c)

(d) (e)

Figure 3. Comparison of classification results for AW segmentation using different kinds of classifiers: (a) Original Image, (b) Expert-marked ground truth for AW, (c) Result by the single overall classifier trained on all training images, (d) Result by the estimated best base classifier trained on a single image (proposed method), (e) Result by the actual best classifier trained on a single image.

Methods	$\overline{p}$	σ of p	$\overline{q}$	σ of q	$\overline{DSC}$	σ of DSC
Over-all classifier	0.0820	0.0956	0.9802	0.0237	0.1183	0.1196
Actual best base classifier	0.6950	0.1613	0.8838	0954	0.6250	0.1814
Estimated best base classifier (proposed)	0.7361	0.1551	0.8304	0.1225	0.5822	0.2049
Average of all base classifiers	0.3573	0.0921	0.8248	0.0849	0.2491	0.0968

Table 1. Performance level comparison among: Over-all SVM classifier trained on all training images, Actual best base classifier trained on a single image (i.e. the classifier that actually gives the best performance when compared to ground truth), Estimated best base classifier trained on a single image (using our proposed method), Average of all base classifiers' performance measures.

interaction. In Figure 4(a), we also plot the difference in DSC between the actual best classifier and the estimated best model (when $K = 1$) or the user-selected model (when $K = 2, ...50$). We observe that: (1) The automatically estimated top 1 (best) classifier always performs similar to the actual best classifier. The difference in DSC is less than 0.03 on average, with a very small standard deviation. (2) Based on the average performance curve of top-K, the finding is that the more classifiers included, the worse the average performance gets. In fact, the nicely descending curve for average performance demonstrates the exceptional ability of our proposed distance measure in ranking the better classifiers higher. There is a clear contrast between our approach (Fig. 4(a)) and the direct *image dissimilarity* measure (Fig. 4(b)) whose average top-K performance shows no trend. (3) With user interaction (when $K > 1$), the performance difference between the user-selected best model and the actual best model was further decreased; in our experiments, this DSC difference is 0.02 when $K = 3$, 0.01 when $K = 5$ and almost 0 when $K = 10$.

We also compared the proposed method using the estimated best base classifier with other classifier ensemble methods including majority voting [9] and STAPLE [1, 18]. The performance level comparisons are shown in Table 4 and Figure 6. The same base classifiers are used for our approach, majority voting and STAPLE; our method applies the distance-guided measure to estimate the best model while the other two combine the predictions of all base classifiers by either majority voting or Expectation Maximization (EM). The proposed method achieved significantly better performance measures (p, q, DSC) than both STAPLE ensemble and majority voting.

Finally, we compared our method with other ensemble methods including Boosting [3, 14] and Bagging [15]. Table 4 and Figure 6 show the comparison results. The base (weak) classifiers in Boosting and Bagging are learned using Naive Bayesian classifiers. From the results, it is clear that the proposed approach produces much better accuracy than other state-of-the-art ensemble learning algorithms on the problem of cervigram tissue segmentation. It could be that the large variations in tissue appearance in the cervigram image database make it difficult to form a proper en-

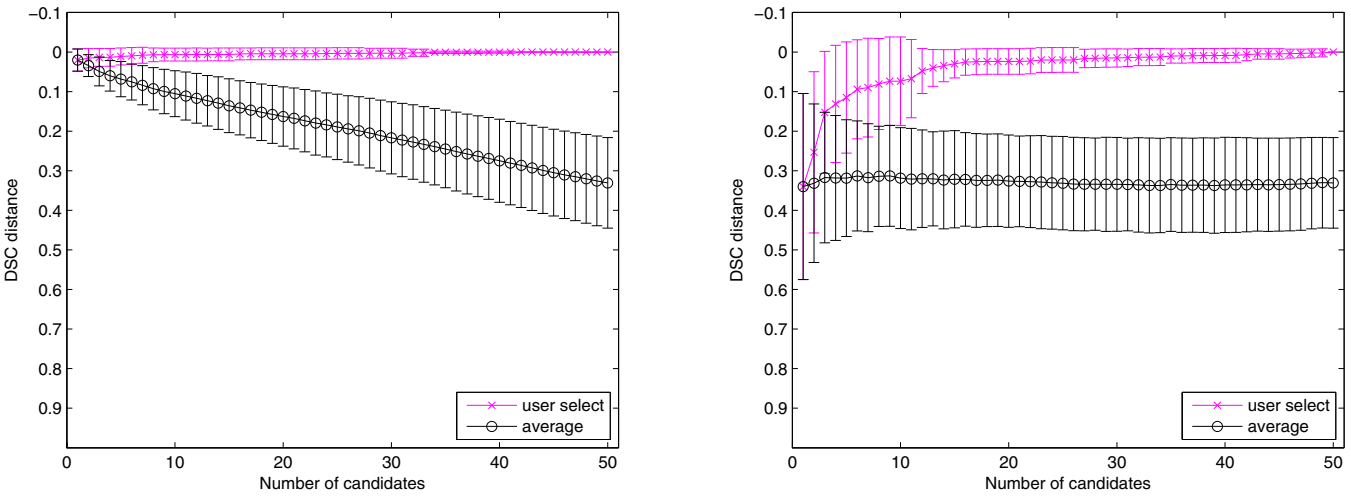


Figure 4. Comparing our distance measure with a direct class-independent *image appearance dissimilarity* distance measure. (a) DSC difference between our estimated best classifier (using proposed cost function in Eq. 5) and the actual best classifier, (b) DSC difference between the estimated best classifier using direct image dissimilarity and the actual best classifier.

Methods	$\overline{p}$	σ of p	$\overline{q}$	σ of q	$\overline{DSC}$	σ of DSC
Actual best base classifier	0.6950	0.1613	0.8838	0954	0.6250	0.1814
Estimated best base classifier (proposed)	0.7361	0.1551	0.8304	0.1225	0.5822	0.2049
Best classifier estimated based on direct image dissimilarity	0.3443	0.3317	0.8233	0.2074	0.2366	0.1945

Table 2. Performance comparison between best classifiers picked by class-dependent distance measure (proposed method) and best classifiers picked by a brute force class-independent image dissimilarity distance measure.

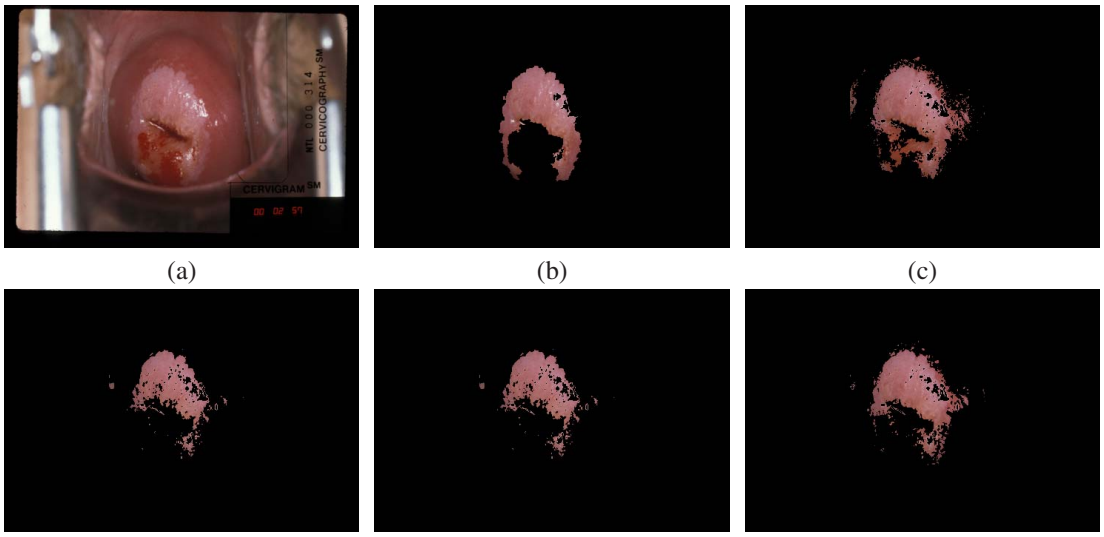
semble of classifiers, each of which provides reasonable performance to support the ensemble.

4. Conclusion and Discussion

We introduced an approach for selecting the best base classifier in an ensemble based on measuring class-dependent distance between feature distributions of the test and training images. We compared this approach with a traditional *image dissimilarity* based distance measure and found that our novel distance measure has exceptional ability in consistently ranking better classifiers higher. We applied the method to segmenting tissue regions, especially the biomarker Acetowhite regions, in digitized uterine cervix images. Experimental results show that our method achieves significantly better accuracy than (1) a single SVM classifier learned using all training images, (2) other classifier ensemble methods including majority voting and STAPLE, and (3) boosted and bagged Naive Bayes classification. In our future work, we will also evaluate the applicability of the method for general color classification in application areas beyond cervigram segmentation.

The key observations that led to our approach are two-fold: (1) Given a training image dataset with diverse appearance, adaptively selecting the best base classifier based on test image evidence is more promising than combining all classifiers. (2) Instead of relying on general *image dissimilarity*, on the problem of tissue classification it makes more sense to *seek the training example with the most similar classification decision boundary*. Our novel class-dependent distance measure is a first attempt in the direction of evaluating the distance in decision boundary between testing and training images. We are developing a theoretical proof of this concept.

However, by requiring training a base classifier using each training image and measuring the cost function of each classifier, our method achieves better accuracy but sacrifices efficiency. Segmentation can still be done within a reasonable amount of time ($\sim$ a few minutes in Matlab for one test image). And given the significant gain in accuracy (See Tables 1, 2, 3, 4), we consider this extra computational cost acceptable. We are also working on hierarchical classifiers based on the new distance measure as well as exploiting multicore platforms and parallel C/C++ implementations, which should reduce the segmentation time of an image to seconds, even with more texture feature types.



(a) (b) (c)

(d) (e) (f)

Figure 5. Comparison of Top-K models' classification results for AW segmentation. (a) Original Image, (b) Expert-marked ground truth, (c) Result by the estimated top 1 (best) classifier, (d) Result by the best classifier picked among the estimated top 3, by user interaction, (e) Result by the best classifier picked among the estimated top 5, (f) Result by the actual best-performing base classifier.

Methods	$\overline{p}$	σ of p	$\overline{q}$	σ of q	$\overline{DSC}$	σ of DSC
Actual best base classifier	0.6950	0.1613	0.8838	0954	0.6250	0.1814
Estimated best base classifier	0.7361	0.1551	0.8304	0.1225	0.5822	0.2049
Average of estimated Top 3 classifiers w/o user interaction	0.7268	0.1219	0.8112	0.1219	0.5575	0.1948
Best among Top 3 classifiers w/ user interaction	0.7174	0.1573	0.8637	0.1047	0.6043	0.1908
Average of estimated Top 5 classifiers w/o user interaction	0.7126	0.1097	0.7982	0.1282	0.5375	0.1922
Best among Top 5 classifiers w/ user interaction	0.6967	0.1588	0.8770	0.1005	0.6105	0.1908

Table 3. Performance comparison of top-K results w/ user interaction (i.e. user selects the best among top-K) and w/o interaction (i.e. averaging top-K results).

Acknowledgements

The authors would like to thank the Communications Engineering Branch, National Library of Medicine—NIH, and the Hormonal and Reproductive Epidemiology Branch, National Cancer Institute—NIH, for providing the data and support of this work. The authors are thankful to Yaoyao Zhu, Daniel Lopresti, Zhiyun Xue, L. Rodney Long, Sameer Antani, and George Thoma, for stimulating discussions on the classification problem.

References

[1] Y. Artan and X. Huang, "Combining multiple 2ν-SVM classifiers for tissue segmentation," Proc. of *ISBI* 2008, pp. 488–491.

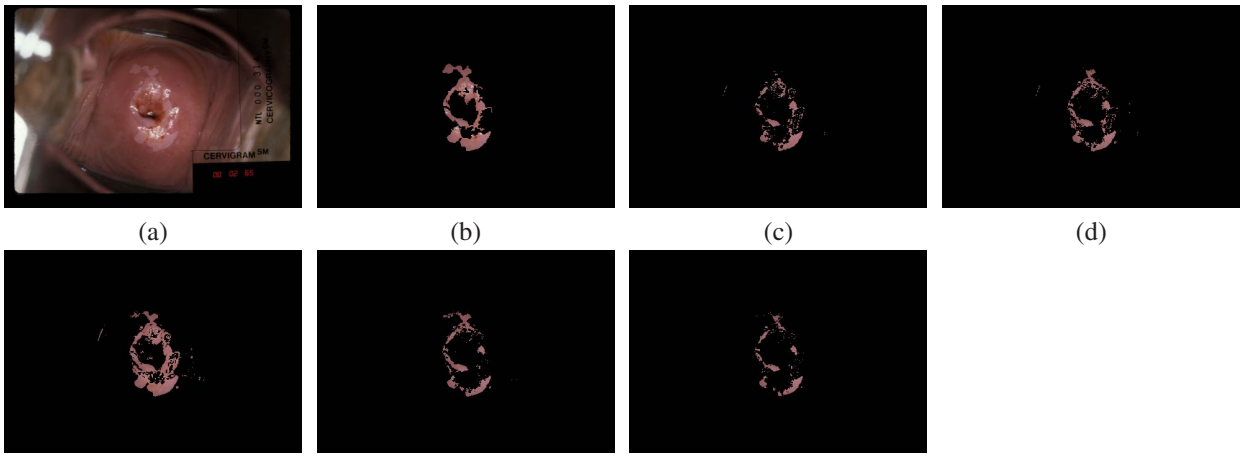
[2] Herman Chernoff, "Large-Sample Theory: Parametric Case," in *The Annals of Mathematical Statistics*, Vol. 27, pp. 1-22, Mar. 1956.

[3] Y. Freund, R. E. Schapire, "A Decision-theoretic Generalization of On-line Learning and an Application to Boosting," in *Journal of Computer and System Sciences*, 55(1):119-139, 1997.

[4] S. Gordon, G. Zimmerman, R. Long, S. Antani, J. Jeronimo, and H. Greenspan, "Content analysis of uterine cervix images: Initial steps towards content based indexing and retrieval of cervigrams.," in *SPIE, Medical Imaging: Image Processing*, Vol. 6144, pp. 2037-2045, 2006.

[5] X. Huang, W. Wang, Z. Xue, S. Antani, L. R. Long, and J. Jeronimo, "Tissue classification using cluster features for lesion detection in digital cervigrams," in *SPIE, Medical Imaging: Image Processing, 2008.*

[6] J. Jeronimo, L. Long, L. Neve, M. Bopf, S. Antani and M. Schiffman, "Digital tools for collecting data from cervigrams for research and training in colposcopy," in *J. of Lower Genital Tract Disease*, 10(1) (2006) 16-25.



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

(e) (f) (g)

Figure 6. Classification results for AW segmentation by different ensemble methods. (a) Original Image, (b) Expert-marked ground truth, (c) majority voting, (d) STAPLE, (e) our proposed method, (f) AdaBoost, (g) Bagging.

Methods	$\overline{p}$	σ of p	$\overline{q}$	σ of q	$\overline{DSC}$	σ of DSC
Estimated best base classifier (proposed)	0.7361	0.1551	0.8304	0.1225	0.5822	0.2049
Majority voting	0.3746	0.2342	0.9337	0.0818	0.3543	0.1858
STAPLE	0.3669	0.2118	0.9381	0.0702	0.3902	0.1790
AdaBoost+NaiveBayes	0.2574	0.1802	0.9341	0.0854	0.2609	0.1498
Bagging+NaiveBayes	0.2732	0.1933	0.9301	0.0896	0.2814	0.1652

Table 4. Performance comparison among different ensemble methods: the proposed method, majority voting, STAPLE, AdaBoost and Bagging.

[7] D. Johnson, and S. Sinanovic, "Symmetrizing the Kullback-Leibler distance," in *Technical report, Rice University*, 2001

[8] S. Gordon, S. Lotenberg and H. Greenspan, "Shape priors for segmentation of the cervix region within uterine cervix images," in *SPIE medical imaging*, 2008.

[9] Ludmila I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 281–286, 2002.

[10] N. Yamaguchi and N. Ishii, "Combining classifiers in error correcting output coding," *Syst. Comput. Japan*, vol. 35, no. 4, pp. 9–18, 2004.

[11] A. Popovic, D. La, M. Engelhardt, K. Radermacher, "Statistical validation metric for accuracy assessment in medical image segmentation," in *International Journal of Computer Assisted Radiology and Surgery*, 2 (2007) 169181

[12] M. De Santo, G. Percannella, C. Sansone, M. Vento, "A neural multi-expert classification system for MPEG audio segmentation," in *Advances in Pattern Recognition*, pp. 50–59, 2001.

[13] F. Schroff, A. Criminisi, and A. Zisserman, "Single-histogram class models for image segmentation" in *ICVGIP*, 2006

[14] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost: Joint Appearance, Shape and context Modeling for Multi-Class Object Recognition and Segmentation," in *Proc. of ECCV*, pp. 115, 2006.

[15] Thomas G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, 2000.

[16] Thomas G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, pp. 1-15, 2000.

[17] B. Tulpule, D. Hernes, Y. Srinivasan, S. Yang, S. Mitra, Y. Sriraja, B. Nutter, B. Phillips, L.R. Long, and D. Ferris, "A probabilistic approach to segmentation and classification of neoplasia in uterine cervix images using color and geometric features," in *SPIE, Medical Imaging: Image Processing*, Vol. 5747, pp. 9951003, 2005.

[18] S.K. Warfield, K.H. Zou, and W.M. Wells, III, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.