A Multiple Instance Learning Approach toward Optimal Classification of Pathology Slides

M. Murat Dundar
IUPUI
Indianapolis, IN USA
dundar@cs.iupui.edu

Sunil Badve
Indiana University
Indianapolis, IN USA

Vikas C. Raykar
Siemens Medical Solutions
Malvern, PA USA

Rohit K. Jain
IUPUI
Indianapolis, IN USA

Olcay Sertel
The Ohio State University
Columbus, OH USA

Metin N. Gurcan
The Ohio State University
Columbus, OH USA

Abstract

Pathology slides are diagnosed based on the histological descriptors extracted from regions of interest (ROIs) identified on each slide by the pathologists. A slide usually contains multiple regions of interest and a positive (cancer) diagnosis is confirmed when ***at least one of the ROIs*** *in the slide is identified as positive. For a negative diagnosis the pathologist has to rule out cancer for* ***each and every ROI*** *available. Our research is motivated toward computer-assisted classification of digitized slides. The objective in this study is to develop a classifier to optimize classification accuracy at the slide level. Traditional supervised training techniques which are trained to optimize classifier performance at the ROI level yield suboptimal performance in this problem. We propose a multiple instance learning approach based on the implementation of the large margin principle with different loss functions defined for positive and negative samples. We consider the classification of intraductal breast lesions as a case study, and perform experimental studies comparing our approach against the state-of-the-art.*

1. Introduction

Pathology diagnoses are made according to a set of criteria defined by the World Health Organization (WHO). While these criteria are generally easy to identify for most lesions, there are borderline cases where it becomes difficult to determine with absolute certainty whether a lesion is malignant or benign. Often times this difficulty is in one or more of the criteria being somewhat ambiguous, thus leading to a diversity of opinions among pathologists. Diagnoses are typically made using formalin fixed paraffin embedded tissue specimens, which are counterstained with hematoxylin or a mixture of hematoxylin/eosin (H&E). Morphological characterization of the cells within the tissue specimen being examined, helps the pathologist decide whether the lesion is cancerous or not. This evaluation requires qualitative as well as quantitative analysis of the specimen and then combining evidence obtained from different parts of the diagnostic process into a coherent strategy used for final diagnosis; a task that we believe can be more reliably and accurately performed with assistance from a computer-assisted diagnosis (CAD) system.

In our earlier work, we have developed image analysis tools for computer-assisted classification of different gradings in neuroblastoma [5] and follicular lymphoma [7], achieving prediction accuracies of 88% and 86% respectively for these problems. One bottleneck that remains to be addressed in this domain is the optimization of the system classifier. A pathology slide is characterized in terms of the regions of interest (ROIs) identified in that slide. More specifically, if we denote a slide as a sample, and an ROI as an instance, each sample is characterized by a multitude of instances. Our objective is to develop a classifier to optimize classification accuracy at the slide level, which is different than the instance-based optimization strategy used in traditional supervised training techniques. The problem of training a binary classifier in the presence of multiple instances for each sample is known in the literature as *multiple instance learning* (MIL). In MIL, a sample is classified as positive if at least one of the instances is classified as positive and negative when all instances are classified as negative.

In [2] we have developed a multiple-instance learning approach based on the convex-hull idea and performed experimental studies on two different computer-assisted detection applications which demonstrated that our approach significantly improves detection accuracy when compared to other MIL techniques proposed in the literature. Since this approach was developed for problems where only positive samples are characterized by multiple instances, it is not directly applicable to the current problem where negative cases also contain multiple instances.

In this study, we will extend our prior work in [2] by defining a pair of asymmetric loss functions for positive and negative samples in a large-margin framework to adapt it for the current problem of classifying pathology slides. The large margin framework is selected for developing the proposed MIL approach, mainly because the difference between rendering positive and negative classification of samples can be directly incorporated into the optimization problem to be solved without sacrificing much from the algorithmic advantages the original problem offers. Our study is motivated by the classification of intraductal breast lesions. We will perform experimental studies comparing the proposed approach with the state-of-the-art on a dataset containing 40 digitized breast tissue specimens.

The rest of the paper is organized as follows. In Section 2 we will review the large margin principle. In Section 3 we will discuss the proposed MIL approach. Experimental results will be presented in Section 4. We will conclude with the discussion of the results and future research directions in Section 5.

2. Large Margin Principle

In traditional supervised training we are given a training dataset $D = \{x_i, y_i\}_{i=1}^{N}$ containing N samples where $x_i \in \mathcal{X} = \Re^d$ is an instance (d-dimensional feature vector) characterizing sample i and $y_i \in \mathcal{Y} = \{\pm 1\}$ is the corresponding known label. The task is to optimize a classification function $f : \mathcal{X} \rightarrow \mathcal{Y}$. Large margin principle optimizes the classifier, i.e. f, as a linear hyperplane separating two classes in the feature space so as to maximize the margin between the classes while minimizing the number of samples on the wrong side of the margin. The hyperplane is defined by $f(x) = w \cdot x + w_0 = 0$ and positive and negative margins are set at $f(x) = w \cdot x + w_0 = 1$ and $f(x) = w \cdot x + w_0 = -1$ respectively, where $\cdot$ denotes a dot product, and the margin between positive and negative classes is expressed in terms of w as $\frac{1}{\Phi(w)}$, where $\Phi(\cdot)$ is some norm function. Therefore, the formulation for a classifier based on the large margin principle can be obtained by solving the following optimization problem:

$$\mathcal{J}(w, w_0) = \Phi(w) + C \sum_{i=1}^{N} (1 - y_i(w \cdot x_i + w_0))_+ \tag{1}$$

where $(\cdot)_+ = max(0, \cdot)$ represents the hinge loss, and C is the cost preassigned to the misclassification associated with x_i.

For a convex function $\Phi(w)$ (1) is also convex. For $\Phi(w) = \|w\|_2^2$, where $\|.\|_2$ is the 2-norm, (1) results in the conventional Quadratic-Programming Support Vector Machines (SVM) [9], and for $\Phi(w) = |w|$, where $|.|$ is the 1-norm it yields the sparse Linear-Programming-SVM.

3. Classifier Training with Multiple Instances

Optimizing the classifier at the slide level will require defining different loss functions for positive and negative samples such that a positive sample is penalized only when all of its instances, i.e., all ROIs in that slide, are classified negative whereas a negative sample is penalized when at least one of its instances is classified positive. In equation (1) the same loss function is used for positive and negative samples and the objective function is penalized for every misclassified ROI. Since this loss function is defined with individual ROIs as opposed to all ROIs in a slide, the naive large margin formulation presented in (1) does not incorporate the difference between rendering positive and negative classification into training. To address this issue, we define two new loss functions for positive and negative samples. We first update our notation for the training dataset, D, to account for the multiple instances in each sample by $D = \left\{ \left\{ x_i^j, y_i \right\}_{j=1}^{M_i} \right\}_{i=1}^{N}$, where x_i^j is the feature vector characterizing the j^{th} ROI in slide i and M_i is the number of ROIs in slide i.

Loss function for negative samples: For negative cases loss is induced when at least one of the ROIs in a slide is classified as positive. This requirement can be incorporated into the training of the classifier by replacing the hinge loss function used in (1)with its multivariable counterpart. The new loss function for negative cases is defined as $\max\left(0, e_i^1, \ldots, e_i^{M_i}\right)$, where $e_i^j = 1 + \left(w \cdot x_i^j + w_o\right)$. This ensures that the loss induced by a negative slide i is zero only if $\forall j : \left(w \cdot x_i^j + w_o\right) \leq -1$, i.e., all ROIs in the slide are correctly classified as negative.

Loss function for positive cases: For positive cases loss is induced when all of the ROIs in a slide are classified as negative. In other words for correct diagnosis, it is sufficient to identify at least one ROI on a slide as positive. In [2], we proposed an approach based on the representation of a sample by a feature vector within the convex hull of all its instances. Let λ_i s.t. $0 \leq \lambda_i^j, e \cdot \lambda_i = 1$, be the vector containing the coefficients of the convex combination of all ROIs in slide i, and e be a vector of ones. Then the feature vector characterizing slide i is defined by $\bar{x}_i = x_i^1 \lambda_i^1 + \ldots + x_i^{M_i} \lambda_i^{M_i} = X^i \cdot \lambda_i$, where $X^i = [x_i^1 \ldots x_i^{M_i}]$ is the data matrix containing feature vectors of all ROIs within slide i. The loss function for a positive case in this new framework can be defined as $\left(1 - \left(w \cdot \left(X^i \cdot \lambda_i\right) + w_o\right)\right)_+$, which is a function of both convex-hull coefficients λ_i and classifier coefficients w.

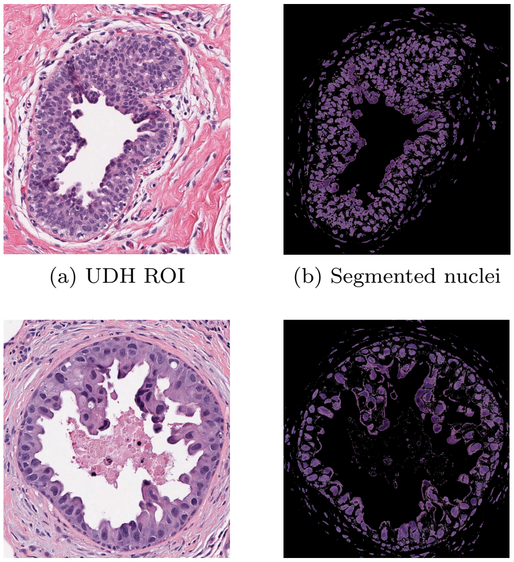
With the new loss functions for positive and negative samples added to the problem, the large-margin formulation becomes:

$$\begin{array}{rl} \min\limits_{(w, w_0, \lambda^i)} & \Phi(w) + C_- \sum_{i \in \Omega^-} \max\left(0, e_i^1, \ldots, e_i^{M_i}\right) \\ + & C_+ \sum_{i \in \Omega^+} \left(1 - \left(w \cdot \left(X^i \cdot \lambda_i\right) + w_o\right)\right)_+ \\ \text{s.t.} & 0 \leq \lambda_i \\ & e \cdot \lambda_i = 1 \end{array} \tag{2}$$

where Ω^+ and Ω^- are the corresponding sets of indices for the positive and negative samples respectively, and the two constraints are imposed to ensure that the feature vector characterizing a positive sample is always within the convex-hull of its instances. Since this problem is no longer convex, a direct solution via convex programming is not possible; however, a solution can still be obtained via biconvex programming. This involves solving two convex subproblems in an iterative way. Convergence properties of a similar problem were discussed in our earlier work [1] where we proved that the algorithm is guaranteed to converge in a suboptimal manner.

4. Experimental Results

The continuum of intraductal breast proliferations encompasses benign lesions of usual (UDH) and atypical (ADH) ductal hyperplasia and low (LG-DCIS) and high-grade (DCIS) ductal carcinoma in situ. While the histological criteria are generally easy to identify for lesions at the two ends of the spectrum, i.e. UDH and DCIS, for borderline cases, i.e. ADH and LG-DCIS, it becomes difficult to determine with absolute certainty whether a lesion is one or the other. Although the CAD system will eventually be implemented to address classification of all intraductal breast lesions, in this feasibility study, we focused our efforts on the classification of UDH and DCIS only. We have collected 20 cases of DCIS and UDH (40 total) from the Indiana University Medical Center according to the approved Institutional Review Board (IRB) protocol for this study. The digitized slides, i.e., scans, are evaluated by a research associate and several regions of interest are identified on each slide.



(a) UDH ROI

(b) Segmented nuclei

(c) DCIS ROI

(d) Segmented nuclei

Figure 1. Segmented nuclei for sample UDH and DCIS ROIs

The scans are first preprocessed to convert the RGB color space into the La*b* color space. In the La^*b^* color space each pixel is characterized with a numeric vector containing intensity and color in-

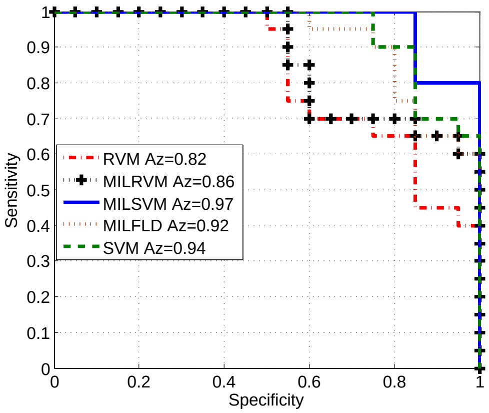


Figure 2. ROC curves obtained for classifiers using leave-one-patient-out cross-validation

formation separately. Next, using the values for L, a^* and b^* channels for each pixel in the image, a clustering algorithm is implemented to segment different cytological components of the breast tissue. The segmentation of the nuclei for sample DCIS and UDH ROIs are shown in Figure 1. The feature extraction module implements Haralick texture descriptors [4] for each channel of the La^*b^* color space to characterize the textural properties of different cytological components. Each region of interest is characterized by a feature vector of size nine. Feature vectors extracted from a total of 195 ROIs across 40 cases are pooled into a training dataset along with their labels at the slide level. This dataset is used to train a classifier using the MIL approach proposed in this study with $\Phi(\alpha) = |\alpha|$ (MILSVM). We compared the performance of this approach with two other MIL techniques from the literature, namely the MIL versions of the relevance vector machine (MILRVM) [6] and Fisher's linear discriminant (MILFLD) [2]. We choose these two techniques for their superior performance in computer-assisted detection applications studied in our earlier work over other MIL techniques considered from the literature. Two traditional supervised training techniques, namely RVM [8] and linear SVM [9] are also included in this comparative analysis. All classifiers are tuned and validated by the leave-one-patient-out cross validation approach. The receiver operating characteristics (ROC) curves obtained for the five classifiers are shown in Figure 2.

5. Conclusion

When we compare the area under the ROC curves (Az values) obtained for each classifier, the proposed approach achieves an Az value of 0.97. The second largest $Az = 0.94$ is achieved by SVM. Significance analysis between the two ROC curves using the approach in [3] yields a p-value of 0.16, which indicates the improvement is most likely not random. We believe that these preliminary results achieved by the proposed MIL algorithm are promising and shows the potential of a CAD system in classifying UDH and DCIS lesions. Future research efforts will focus on increasing the size of the dataset, mapping computed features to histological descriptors and improving the proposed approach to classify borderline lesions in addition to UDH and DCIS.

References

[1] M. M. Dundar and J. Bi. Joint optimization of cascaded classifiers for computer aided detection. In *Proceedings of CVPR 2007, Minneapolis, USA.*

[2] G. Fung, M. Dundar, B. Krishnapuram, and R. B. Rao. Multiple instance learning for computer aided diagnosis. In *NIPS*, pages 425–432, 2006.

[3] J. A. Hanley and B. J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843, 1983.

[4] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, 3(6):610–621, 1973.

[5] J. Kong, O. Sertel, H. Shimada, K. L. Boyer, J. H. Saltz, and M. N. Gurcan. Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation. *Pattern Recogn.*, 42(6):1080–1092, 2009.

[6] V. C. Raykar, B. Krishnapuram, J. Bi, M. Dundar, and R. B. Rao. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *Proceedings of ICML 2008.*

[7] O. Sertel, J. Kong, U. V. Catalyurek, G. Lozanski, J. H. Saltz, and M. N. Gurcan. Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading. *Journal of Signal Processing Systems*, 55(1-3):169–183, 2009.

[8] M. E. Tipping. The relevance vector machine. In *Proceedings of NIPS 2000.*

[9] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.