Multiple Clustered Instance Learning for Histopathology Cancer Image Classification, Segmentation and Clustering

Yan Xu*[1,2], Jun-Yan Zhu*[2,3], Eric Chang[2] and Zhuowen Tu[2,4]
[1]State Key Laboratory of Software Development Environment,
Key Laboratory of Biomechanics and Mechanobiology of Ministry of Education, Beihang University
[2]Microsoft Research Asia
[3]Dept. of Computer Science and Technology, Tsinghua University
[4]Lab of Neuro Imaging and Dept. of Computer Science, UCLA
{xuyan04, junyanzhu89}@gmail.com, {echang, zhuowent}@microsoft.com

Abstract

Cancer tissues in histopathology images exhibit abnormal patterns; it is of great clinical importance to label a histopathology image as having cancerous regions or not and perform the corresponding image segmentation. However, the detailed annotation of cancer cells is often an ambiguous and challenging task. In this paper, we propose a new learning method, multiple clustered instance learning (MCIL), to classify, segment and cluster cancer cells in colon histopathology images. The proposed MCIL method simultaneously performs image-level classification (cancer vs. non-cancer image), pixel-level segmentation (cancer vs. non-cancer tissue), and patch-level clustering (cancer subclasses). We embed the clustering concept into the multiple instance learning (MIL) setting and derive a principled solution to perform the above three tasks in an integrated framework. Experimental results demonstrate the efficiency and effectiveness of MCIL in analyzing colon cancers.

1. Introduction

High resolution histopathology images provide reliable information differentiating abnormal tissues from normal ones, and thus, it is a vital technology for recognizing and analyzing cancers [21, 19, 9, 15]. Recent development in specialized digital microscope scanners makes digitization of histopathology readily accessible. Several systems for classifying and grading cancer histopathology images have been recently proposed. These methods focus on feature design of various types, such as fractal features [11], texture features [12], and object-level features [5]. Various classifiers (Bayesian , KNN and SVM) are used for prostate cancer [11] recognition.

There is also a rich body of literature on supervised learning for image segmentation and classification [22, 23]. However, supervised approaches require a large amount of accurately annotated data; usually, high-quality manual delineations are not only labor-intensive and time-consuming to obtain, but also intrinsically ambiguous. This situation is more conspicuous for cancer tissue classification/segmentation in histopathology images, where obtaining the very detailed annotation is a challenging task even for pathologists. Unsupervised learning methods [7], on the other hand, ease the burden of manual annotation, but often at the cost of inferior results.

In the middle of the spectrum is the weakly supervised learning scenario. The idea is to use coarse-grained annotations to aid automatic exploration of fine-grained information. In our case, it is relatively easy for a pathologist to label a histopathology image as having cancer or not. Our goal is to automatically learn the image models from weakly supervised histopathology images to recognize cancers. The weakly supervised learning direction is closely related to semi-supervised learning problems in machine learning [28]. One particular form of weakly supervised learning is *multiple instance learning* (MIL) [16, 2] where a training set consists of a number of bags; each bag includes many instances and the bag-level label is given but not the instance-level label; the goal of MIL is to learn to predict both bag-level and instance-level labels.

The current literature in MIL assumes single model/cluster/classifier for the target of interest [24], single cluster within each bag [3, 26], or multiple components of the same object [6]. Here, we aim to develop an integrated system to perform pixel-level segmentation (cancer vs. non-cancer) and image-level classification; moreover, it is desirable to discover/identify the subclasses

*indicates equal contributions

of various cancer tissue types as a universal protocol for cancer tissue classification[11] is not all available; this results in patch-level clustering of the cancer tissues; however, the existing MIL frameworks are not able to do these tasks altogether. In this paper, we derive a principled approach, named multiple clustered instance learning (MCIL), to simultaneously perform classification, segmentation, and clustering.

Common histopathology cases include colon, prostate, breast, and neuroblastoma cancers. Here, we focus on colon histopathology images but our method is general and it can be applied to other image types.

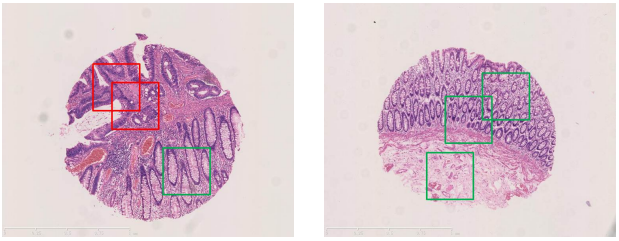
2. Related Work

Related work can be broadly divided into two categories: (1) medical image classification and segmentation in the medical imaging field, and (2) multiple instance learning in the learning and vision field.

As mentioned before, methods developed in the medical imaging field are mostly focused on feature design in supervised settings. Fractal features are used in prostate cancer detection [11]; Kong *et al*. proposed a multi-resolution framework to classify neuroblastic grade using texture information [12]; color graphs were applied in [1] to detect and grade colon cancer in histopathology images; Boucheron *et al*. proposed a method using object-based information for histopathology cancer detection [5]; multiple features including color, texture, and morphometric cues at the global and histological object levels were adopted in prostate cancer detection [21].

Due to the intrinsic ambiguity and difficulty in obtaining human labeling, MIL approaches have its particular advantages in automatically exploiting the fine-grained information and reducing efforts in human annotations. The MIL method has also been adopted in the medical domain [10] with the focus mostly on the medical diagnosis. A multiple instance learning approach was used in [4] to detect accurate pulmonary embolism among the candidates; a computer aided diagnosis (CAD) system [14] was developed for polyp detection with the main focus on supervised learning features, which were then used for multiple instance regression; MIL [8] was adopted for cancer classification in histopathology slides. However, these existing MIL approaches are for medical image diagnosis and none of them perform segmentation, which is crucial in medical image analysis and a specific advantage of our method. Moreover, the integrated classification/segmentation/clustering tasks have not been addressed.

From another perspective, Zhang *et al*. [26] developed a multiple instance clustering (MIC) method to learn the instance clusters as hidden variables. MIC however takes no negatives and each bag contains one cluster only. In our case, multiple clusters of different cancer types might exist within one bag (histopathology image). Babenko *et al*. [3] assumed a hidden variable, pose, to each face (only one) in an image. In [6], multiple components were studied for a single object class, which also differs from our method since we have multiple instances and multiple classes within each bag. The MIL assumption was integrated into multiple-label learning for image/scene classification in [27, 25]. However, multi-class labels were given for supervision in their method while in MCIL, multiple clusters are hidden variables to be explored in an unsupervised way. In [20], the clusters and segmentations were explored for the configuration of object models, which is quite different to the problem setting here. Again, MCIL is able to perform classification, segmentation, and clustering altogether. In addition, our method can be applied in other MIL tasks other than medical imaging applications.



(a) cancer image (b) non-cancer image

Figure 1: Examples of bags and instances in our problem: (a) positive bag (cancer image); (b) negative bag (non-cancer image). Red rectangles: positive instances (cancer tissues); Green rectangles: negative instances (non-cancer tissues).

3. Methods

We follow the general definition of bags and instances in the MIL setting [24]. In this paper, we treat cancer and non-cancer images as positive and negative bags respectively; the image patches densely sampled from the images thus correspond to the instances. Figure 1 shows the definition of positive/negative bags and positive/negative instances. In this problem, a bag is labeled as positive if the bag contains at least one positive instance (cancer tissue); similarly, in histopathology cancer image analysis, if a small part of image is considered as cancer tissues, the histopathology should be diagnosed as positive by pathologists.

An advantage brought by MIL is that if an instance-level classifier is learned, automatic pixel-level segmentation could be performed; bag-level (image-level) classifier could be directly obtained under the MIL setting. The main difference between the case in medical imaging and previous weakly supervised object/face detection [24, 3] is that objects are distinct while tissues in histopathology images form segmentations with no clear boundary shapes.

In the following sections, we first overview the MIL literature, especially recent gradient decent boosting based MIL approaches. Then we integrate the clustering concept into the MIL setting and derive a new formulation, MCIL, under the boosting framework; properties of MCIL with various variations are provided. We also show how classification, segmentation and clustering could be simultaneously conducted in our MCIL algorithm, which is the key contribution of our method.

3.1. Multiple Instance Learning

Here we briefly discuss the MIL problem formulation and study boosting based [17] MIL approaches[24, 3], which serve as the basis for MCIL. In MIL, training data is represented by a set of m vectors[1], often called a bag, $X_i = \{x_{i1}, \ldots, x_{im}\}$ while each bag is associated with a label $y_i \in \mathcal{Y} = \{-1, 1\}$. Each instance $x_{ij} \in \mathcal{X}$ in a bag $X_i \in \mathcal{X}^m$ has a true label $y_{ij} \in \mathcal{Y}$ as hidden variable, which remains unknown during training. In the binary case, a bag is labeled positive if and only if at least one instance in the bag is positive, which could be formulated as:

$$y_i = \max_j (y_{ij}) \tag{1}$$

where $\max$ is essentially equivalent to an OR operator since for $y_{ij} \in \mathcal{Y}$, $\max_j (y_{ij}) = 1 \iff \exists j, \text{ s.t. } y_{ij} = 1$.

The goal of MIL is to learn an instance-level classifier $h(x_{ij}) : \mathcal{X} \to \mathcal{Y}$. A bag-level classifier $H(X_i) : \mathcal{X}^m \to \mathcal{Y}$ could be built with the instance-level classifier:

$$H(X_i) = \max_j h(x_{ij}) \tag{2}$$

Viola *et al*. [24] first introduced MIL-Boost by combining MIL cost functions and AnyBoost framework [17]. Babenko *et al*. [3] re-derived and generalized it later. Here we adopt the loss function defined in the AnyBoost:

$$\mathcal{L}(h) = -\sum_{i=1}^{n} w_i (\mathbf{1}(y_i = 1) \log p_i + \mathbf{1}(y_i = -1) \log (1 - p_i)) \tag{3}$$

$\mathbf{1}(\cdot)$ is an indicator function. The loss function is the standard negative log likelihood. $p_i \equiv p(y_i = 1|X_i)$ and w_i is the prior weight of the i^{th} training data. It is often useful to train with an initial distribution over the data, *e.g*. if more positive than negative training data we are available.

A *softmax* function, a differentiable approximation of $\max$, is then introduced. We summarize four models used in MIL-Boost and MCIL in Table 1: noisy-or (NOR), generalized mean (GM), log-sum-exponential (LSE), and integrated segmentation and recognition (ISR). Parameter r controls sharpness and accuracy in LSE and GM model:

[1] Although each bag may have different number of instances, for clarity of notation, we use m for all the bags here.

	$g_l(v_l)$	$\partial_{g_l}(v_l)/\partial v_i$	domain
NOR	$1 - \prod_l (1 - v_l)$	$\frac{1-g_l(v_l)}{1-v_i}$	$[0, 1]$
GM	$(\frac{1}{m} \sum_l v_l^r)^{\frac{1}{r}}$	$g_l(v_l) \frac{v_i^{r-1}}{\sum_l v_l^r}$	$[0, \infty]$
LSE	$\frac{1}{r} \ln \frac{1}{m} \sum_l \exp(r v_l)$	$\frac{\exp(r v_i)}{\sum_l \exp(r v_l)}$	$[-\infty, \infty]$
ISR	$\frac{\sum_l v'_l}{1+\sum_l v'_l}, v'_l = \frac{v_l}{1-v_l}$	$(\frac{1-g_l(v_l)}{1-v_i})^2$	$[0, 1]$

Table 1: Four softmax approximations $g_l(v_l) \approx \max_l(v_l)$

$g_l(v_l) \to v^*$ as $r \to \infty$. For m variables $\boldsymbol{v} = \{v_1, \ldots, v_m\}$, *softmax* function $g_l(v_l)$ is defined as follows:

$$g_l(v_l) \approx \max_l(v_l) = v^*, \quad \frac{\partial_{g_l}(v_l)}{\partial v_i} \approx \frac{\mathbf{1}(v_i = v^*)}{\sum_l \mathbf{1}(v_l = v^*)} \tag{4}$$

$m = |\boldsymbol{v}|$. Note that for the rest of the paper $g_l(v_l)$ indicates a function g which takes all v_l indexed by l; $g_l(v_l)$ is not a function merely on v_l.

The probability p_i of bag X_i is computed as the *softmax* of probability $p_{ij} \equiv p(y_{ij} = 1|x_{ij})$ of all the instances x_{ij}: $p_i = g_j(p_{ij}) = g_j(\sigma(2h_{ij}))$ where $h_{ij} = h(x_{ij})$ and $\sigma(v) = \frac{1}{1+\exp(-v)}$ is the sigmoid. The weights w_{ij} and the derivatives $\frac{\partial \mathcal{L}}{\partial h_{ij}}$ could be written as:

$$w_{ij} = -\frac{\partial \mathcal{L}}{\partial h_{ij}} = -\frac{\partial \mathcal{L}}{\partial p_i} \frac{\partial p_i}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial h_{ij}} \tag{5}$$

3.2. Multiple Cluster Assumption

Histopathology cancer images include multiple types, which are not addressed by the single model/cluster/classifier in the previous MIL algorithms.

Except for annotation difficulty, unclear definition of cancer tissue type in medical research also motivates us to propose MCIL. There are many individual classification, segmentation and clustering approaches in the computer vision and medical imaging community; however, most of the existing algorithms are designed for one particular purpose and therefore do not fit our task. Here, we simultaneously perform three tasks in an integrated learning framework under the weakly supervised scenario.

We are still given a training dataset containing bags $X_i = \{x_{i1}, \ldots, x_{im}\}$ and bag labels $y_i \in \mathcal{Y} = \{-1, 1\}$; here, we integrate the clustering concept into the MIL setting by assuming the existence of hidden variable $y_{ij}^k \in \mathcal{Y}$ which denotes whether the instance x_{ij} belongs to the k^{th} cluster. Similar to MIL constraints, if one instance belongs to one of K clusters, this instance could be considered as a positive instance; and only if at least one instance in a bag is labeled as positive, the bag is considered as positive. This forms the MCIL assumption, which could be formulated as follows:

$$y_i = \max_j \max_k (y_{ij}^k) \tag{6}$$

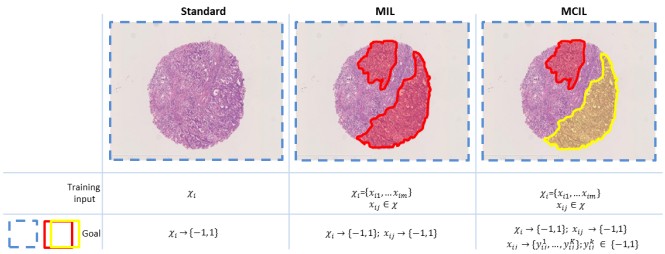


Figure 2: Distinct learning goals of supervised learning, MIL and MCIL. MCIL could perform image-level classification ($x_i \rightarrow \{-1, 1\}$), pixel-level segmentation ($x_{ij} \rightarrow \{-1, 1\}$) and patch-level clustering($x_{ij} \rightarrow \{y_{ij}^1, \ldots, y_{ij}^K\}, y_{ij}^k \in \{-1, 1\}$).

Again the max is equivalent to an OR operator where $\max_k (y_{ij}^k) = 1 \iff \exists k$, s.t. $y_{ij}^k = 1$.

The goal of MCIL is to learn K instance-level classifiers $h^k(x_{ij}) : \mathcal{X} \rightarrow \mathcal{Y}$ for K clusters. Corresponding bag-level classifier for the k^{th} cluster could be built as $H^k(X_i) : \mathcal{X}^m \rightarrow \mathcal{Y}$. The overall cancer classifier could be constructed as $H(X_i) : \mathcal{X}^m \rightarrow \mathcal{Y}$:

$$H(X_i) = \max_k H^k(X_i) = \max_k \max_j h^k(x_{ij}) \tag{7}$$

Figure 2 illustrates the distinction between standard supervised learning, MIL and MCIL.

3.3. Multiple Clustered Instance Learning

In this section, based on the previous derivations, we give the full formulation of our MCIL method. The probability p_i is computed as the *softmax* of $p_{ij} \equiv p(y_{ij} = 1|x_{ij})$ of all the instances and the instance probability p_{ij} could be obtained by *softmax* of $p_{ij}^k = p(y_{ij}^k = 1|x_{ij})$ which measures:

$$p_i = g_j(p_{ij}) = g_j(g_k(p_{ij}^k)) \tag{8}$$

where the p_{ij}^k means the probability of the instance x_{ij} belonging to the k^{th} cluster. We use *softmax* to rewrite the MCIL assumption (eqn. (6)) and give the Remark 1:

$$g_j(g_k(p_{ij}^k)) = g_{jk}(p_{ij}^k) = g_k(g_j(p_{ij}^k)) \tag{9}$$

Again, functions of $g_k(p_{ij}^k)$ can be seen in Table 1; it indicates a function g which takes all p_{ij}^k indexed by k; similarly, functions of $g_{jk}(p_{ij}^k)$ could be understood as a function g including all p_{ij}^k indexed by k and j. Remark 1 can be checked with care and we put the verification into the appendix.

Based on the above equation, we could rewrite eqn. (8) as follows:

$$p_i = g_j(g_k(p_{ij}^k)) = g_{jk}(p_{ij}^k) = g_{jk}(\sigma(2h_{ij}^k)), h_{ij}^k = h^k(x_{ij}) \tag{10}$$

w_{ij}^k / w_i	$y_i = 1$	$y_i = -1$
NOR	$-2p_{ij}^k$	$\frac{-2p_{ij}^k(1-p_i)}{p_i}$
GM	$-\frac{2p_i}{1-p_i}\frac{(p_{ij}^k)^r-(p_{ij}^k)^{r+1}}{\sum_{j,k}(p_{ij}^k)^r}$	$2\frac{(p_{ij}^k)^r-(p_{ij}^k)^{r+1}}{\sum_{j,k}(p_{ij}^k)^r}$
LSE	$-\frac{2p_{ij}^k(1-p_{ij}^k)}{1-p_i}\frac{\exp(rp_{ij}^k)}{\sum_{j,k}\exp(rp_{ij}^k)}$	$\frac{2p_{ij}^k(1-p_{ij}^k)}{p_i}\frac{\exp(rp_{ij}^k)}{\sum_{j,k}\exp(rp_{ij}^k)}$
ISR	$-\frac{2\mathcal{X}_{ij}^k p_i}{\sum_{j,k}\mathcal{X}_{ij}^k}, \mathcal{X}_{ij}^k = \frac{p_{ij}^k}{1-p_{ij}^k}$	$\frac{2\mathcal{X}_{ij}^k(1-p_i)}{\sum_{j,k}\mathcal{X}_{ij}^k}, \mathcal{X}_{ij}^k = \frac{p_{ij}^k}{1-p_{ij}^k}$

Table 2: MCIL w_{ij}^k / w_i with different *softmax* functions

σ is the sigmoid function mentioned before. Therefore, we give the weights w_{ij}^k and derivatives $-\frac{\partial \mathcal{L}}{\partial h_{ij}^k}$ could be given as:

$$w_{ij}^k = -\frac{\partial \mathcal{L}}{\partial h_{ij}^k} = -\frac{\partial \mathcal{L}}{\partial p_i}\frac{\partial p_i}{\partial p_{ij}^k}\frac{\partial p_{ij}^k}{\partial h_{ij}^k} \tag{11}$$

$$\frac{\partial \mathcal{L}}{\partial p_i} = \begin{cases} -\frac{w_i}{p_i} & \text{if } y = 1 \\ \frac{w_i}{1-p_i} & \text{if } y = -1 \end{cases} \tag{12}$$

$$\frac{\partial p_i}{\partial p_{ij}^k} = \begin{cases} \frac{1-p_i}{1-p_{ij}^k} & \text{NOR,} \quad p_i\frac{(p_{ij}^k)^{r-1}}{\sum_{j,k}(p_{ij}^k)^r} \quad \text{GM} \\ \frac{\exp(rp_{ij}^k)}{\sum_{j,k}\exp(rp_{ij}^k)} & \text{LSE,} \quad (\frac{1-p_i}{1-p_{ij}^k})^2 \quad \text{ISR} \end{cases} \tag{13}$$

$$\frac{\partial p_{ij}^k}{\partial h_{ij}^k} = 2p_{ij}^k(1-p_{ij}^k) \tag{14}$$

Thus, we summarize the weight w_{ij}^k / w_i in Table 2. Recall that w_i is the given prior weight for the i^{th} bag. Details of MCIL are demonstrated in Algorithm 1. Notice that the outer loop is for each weak classifier while the inner loop is for the k^{th} strong classifier.

We introduce the latent variables y_{ij}^k, which denote the instance x_{ij} belonging to the k^{th} cluster, and we encode the concept of clustering by re-weighting the instance-level weight w_{ij}^k. If the k^{th} cluster can explain some instances well, the weight of instances and bags for other clusters decrease in re-weighting. Thus, it forms a competition among clusters.

4. Experiments

In the experiments, we apply our method on several cancer image datasets. The advantage of our integrated MCIL framework is evident in image-level classification compared with Multiple Kernel Learning (MKL) [23], MILBoost[24], standard Boosting[17], mi-SVM[2], and MISVM[2], in pixel-level segmentation compared with MILBoost and standard Boosting, and in patch-level clustering compared with Boosting + K-means[7] and MIL + K-means.

Algorithm 1 MCIL-Boost

Input: Bags $\{X_1, \ldots, X_n\}, \{y_1, \ldots, y_n\}, K, T$
Output: $h^1, \ldots, h^K$
for $t = 1 \rightarrow T$ **do**
  for $k = 1 \rightarrow K$ **do**
    Compute weights $w_{ij}^k = -\frac{\partial \mathcal{L}}{\partial h_{ij}^k}$
    Train weak classifiers h_t^k using weights $|w_{ij}^k|$
    $h_t^k = \arg\min_h \sum_{ij} \mathbf{1}(h(x_{ij}^k) \neq y_i)|w_{ij}^k|$
    Find α_t via line search to minimize $\mathcal{L}(., h^k, .)$
    $\alpha_t^k = \arg\min_\alpha \mathcal{L}(., \mathbf{h}^k + \alpha h_t^k, .)$
    Update strong classifiers $\mathbf{h}^k \leftarrow \mathbf{h}^k + \alpha_t^k h_t^k$
  end for
end for

	NC	MTA	LTA	MA	SRC
Binary	30	30	0	0	0
Multi1	30	15	9	0	6
Multi2	30	13	9	8	0

Table 3: Number of images in the datasets. The "Binary" dataset contains only one class of cancer images (MTA).

Datasets: We study three colon cancer image datasets: $binary$, $multi1$, and $multi2$. Table 3 shows the constituents of datasets. In $binary$, we demonstrate the advantage of the MIL formulations against the state-of-the-art supervised image categorization approaches. In $mulit1$ and $multi2$, we further show the advantage of MCIL in an integrated framework.
Cancer Types: Five types of colon cancer images are used: Non-cancer (NC), Middle tubular adenocarcinoma (MTA), Low tubular adenocarcinoma (LTA), Mucinous adenocarcinoma (MA), and Signet-ring carcinoma (SRC). We use the same abbreviations for each type in the following sections.
Annotations: All the histopathology images are labeled as cancer or non-cancer images by two pathologists independently. If there exists a disagreement for a certain image between two pathologists, two pathologists together with the third senior pathologist discuss the result until final agreement is reached. We also ask them to label the instance-level segmentation (cancer tissues) and the patch-level clustering (different type) for test data. Instance-level and patch-level annotations also follow the above process to ensure the quality of the ground truth.
Settings: After downsampling the histopathology images by $5\times$, we densely extract 64×64 patches from images. The overlap step size is 32 pixels for training course and 4 pixels for testing. The *softmax* function we use here is GM model and the weak classifier we use is Gaussian function. All the results are reported in a 5-fold cross validation. For parameters, we set $r = 20$, $K = 4$ and $T = 200$. With respect to patch representation, generic features for object classification rather than ones specifically designed for medical imaging are used including $L^*a^*b^*$ Color Histogram, Local Binary Pattern [18], and SIFT [13]. It is worth noting that we focus on our integrated learning formulation rather than the feature design in this work. We use the same setting for MCIL, MIL-Boost[24], standard Boosting[17], mi-SVM[2] and MI-SVM[2] in the following three experiments.

4.1. Image-level Classification

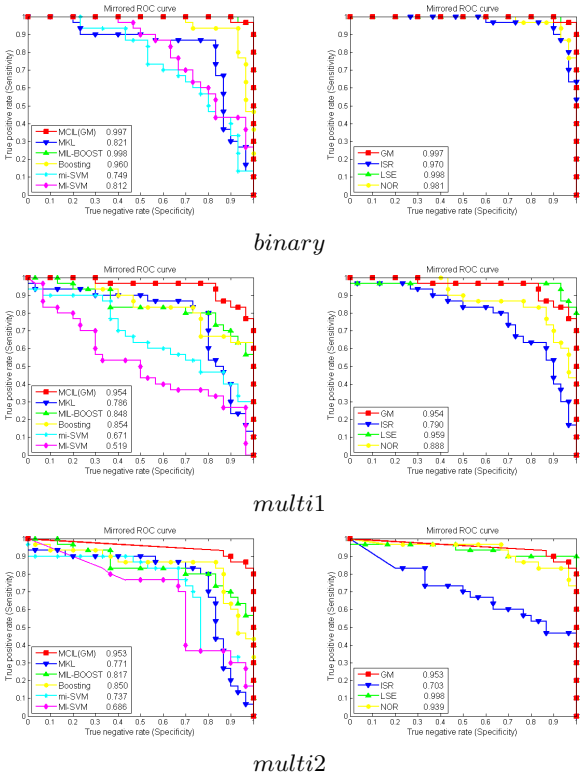
We first measure the bag-level classification (cancer vs. non-cancer). The standard learning baseline is MKL [23] which obtains very competitive results and wins the PASCAL Visual Object Classification Challenge 2009. We use their implementation and follow the same features and parameters reported in their paper. We use all the instances x_{ij} to train a standard Boosting [17] by considering instance-level labels derived from bag-level labels ($y_{ij} = y_i$, $i = 1, \ldots, n$, $j = 1, \ldots, m$). For the MIL baselines, we use MI-SVM [2], mi-SVM [2], and MIL-Boost [24].

Figure 3(a) shows the receiver operating characteristic (ROC) curves for different learning methods in the three datasets. In dataset $binary$, both MCIL and MIL outperform well developed MKL algorithm [23] and standard Boosting[17], which shows the advantage of the MIL formulation to the cancer image classification task. MCIL and MIL-Boost achieve similar performance on the $binary$ dataset of one class/cluster; however, when applied to datasets $multi1$ and $multi2$, MCIL significantly outperforms MIL, MKL and Boosting, which reveals multiple clustering concept integrated in MCIL framework successfully deals with the complex situation in cancer image classification.

Notice that MKL utilizes more discriminative features than that we use in MIL and MCIL. For the computational complexity, it takes several days to use MKL [23] to train a classifier for a dataset containing 60 images while it only takes about two hours using MCIL to achieve a significantly improved result.

We also compare performance based on different *softmax* models. Figure 3(b) shows that LSE model and GM model fit the cancer image recognition task best.

Different cancer types, experiment settings, benchmarks, and evaluation methods are reported in the literature. As far as we know, the code and images used in [11, 21, 9] are not publicly accessible. Hence, it is quite difficult to make a direct comparison between different algorithms. Below we only list their results as references. In [11], 205 pathological images of prostate cancer were chosen as evaluation which included 50 of grade 1-2, 72 of grade 3, 31 of grade 4, and 52 of grade 5. The highest correct classification rates based on Bayesian, KNN and SVM classifiers achieved 94.6%,



(a) classification results (b) different softmax models

Figure 3: Comparisons of image (bag)-level classification results with state-of-the-art methods on the three datasets. (a) shows the ROC curves and our proposed method (MCIL in red) has apparent advantages. (b) demonstrates the effect of using different soft-max functions.

94.2% and 94.6% respectively. In [21], 367 prostate images (218 cancer and 149 non-cancer) were chosen to detect cancer or non-cancer. The highest accuracy was 96.7%. 268 images were chosen to classify gleason grading. The numbers of grades 2-5 are 21, 154, 86 and 7, respectively. The highest accuracy was 81%. In [9], a total of 44 non-cancer images and 58 cancer images were selected to detect cancer or non-cancer. The sensitivity of 90%-95% and the specificity of 86%-93% were achieved according to various features.

4.2. Pixel-level Segmentation

We now turn to instance-level experiment. Since it is both time-consuming and intrinsically ambiguous for pathologists to label detailed cell annotations for all the images and MCIL does not require any instance-level supervision, we report instance-level results in a subset of $multi1$. In particular, we randomly select 11 cancer images and 11 non-cancer images to construct the subset. Pathologists provide careful instance-level annotations for cancer images.

MCIL generates a probability map P_i for each bag X_i (image). We use the F-measure for segmentation measurement. Given the ground truth map G_i, Precision $= |P_i \cap G_i|/|P_i|$, Recall $= |P_i \cap G_i|/|G_i|$ and F-measure $= \frac{2\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$.

The segmentation baselines are MIL-Boost [24] and standard Boosting [17] we mentioned before. Unsupervised segmentation techniques cannot be used in comparison since they do not output labels for each segment. The F-measures of MCIL, MIL-Boost, and standard Boosting are 0.588, 0.231, and 0.297 respectively, which suggests the great advantage of MCIL against previous supervised and MIL-based segmentation approaches. Figure 4 shows some results of test data. Standard Boosting tends to detect non-cancer tissues as cancer tissues since it considers all the instances in positive bags as positive. Even explicitly formulated as an MIL scheme, MIL-Boost is based on a single class/model/classifier assumption and can not explain all the clusters among positive bags, which limits its application on multi-cluster multi-instance tasks like cancer image recognition.

4.3. Patch-level Clustering

MCIL obtains the clustering results at the same time. On the same test data mentioned in pixel-level segmentation, we demonstrate the advantage of MCIL for exploring unknown patterns of cancer images in this section. Here we build two baselines: MIL-Boost [24] + K-means and standard Boosting [17] + K-means. Particularly, we first run MIL-Boost or standard Boosting to perform instance-level segmentation and then use K-means to obtain K clusters among positive instances (cancer tissues). Since we mainly focus on clustering performance here, we only include true positive instances as measured data by removing the influence of poor segmentation results of MIL-Boost and standard Boosting. Purity is used as evaluation measure. The purity of MCIL is 99.70% while the purities of MIL + K-means and Boosting + K-means are only 86.45% and 85.68% respectively. The experiment shows an integrated learning framework of MCIL is better than two separate steps of instance-level segmentation and clustering.

MCIL is able to successfully discriminate cancer types since different types of cancer images are mapped to different clusters (See Figure 4). The SRC cancer image is mapped to red; the MTA cancer images are mapped to green and yellow; and the LTA cancer image is mapped to blue. Both MIL-Boost + K-means and standard Boosting + K-means divide one type of cancer images into several clusters and the results are not consistent between multiple images. The reason why MTA cancer images are divided into two separate clusters is that lymphocytes (green area) are strongly related to cancer cells (yellow area). Lymphocytes

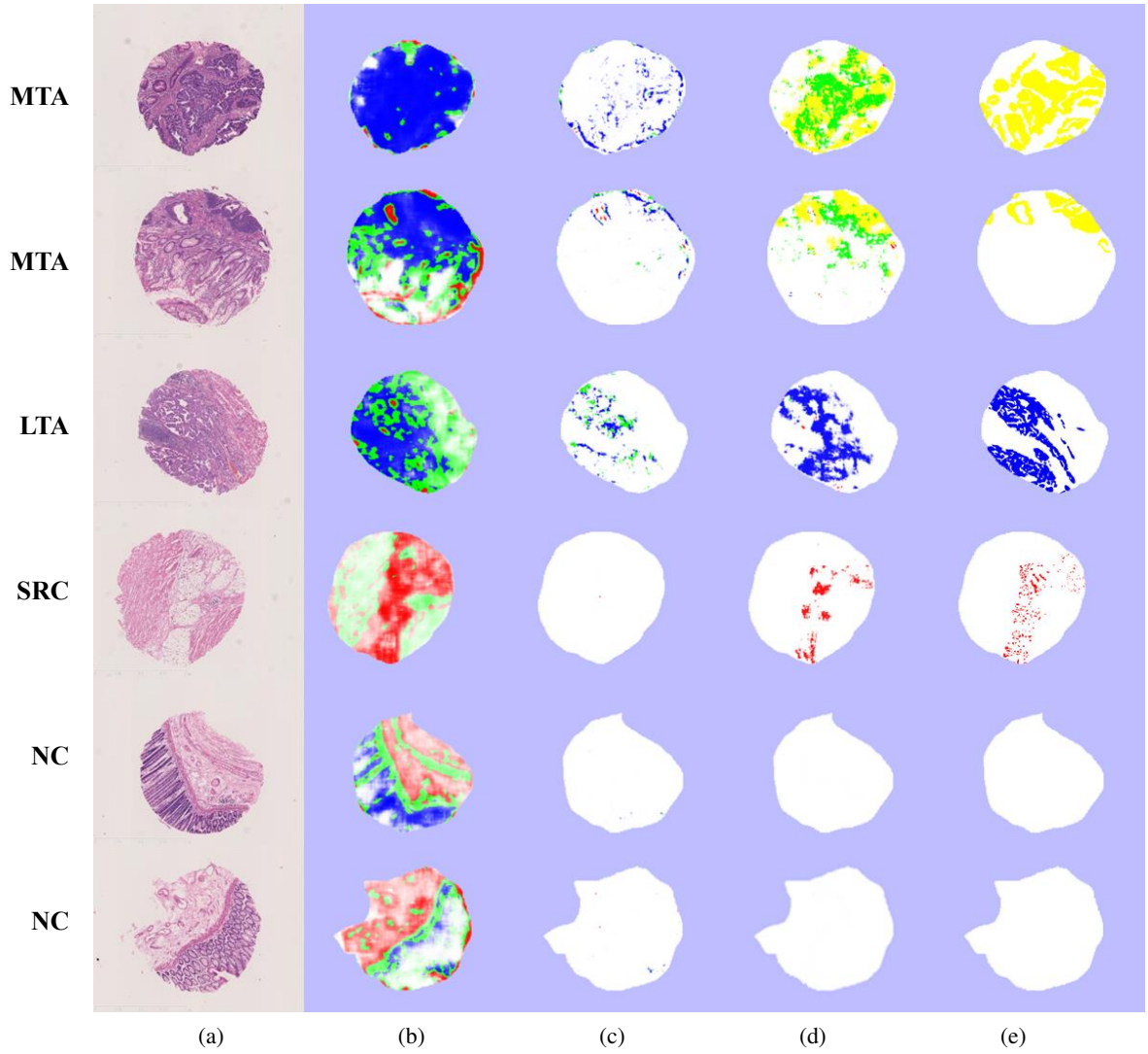


Figure 4: Image Types: from left to right: (a): The original images. (b), (c), (d): The instance-level results (pixel-level segmentation and patch-level clustering) for standard Boosting + K-means, MIL + K-means, and our MCIL. (e): The instance-level ground truth labeled by three pathologists. Different colors stand for different types of cancer tissues. Cancer Types: from top to bottom: MTA, MTA, LTA, SRC, NC, and NC.

have the ability to be resistant to cancer cells. When cancer cells appear, lymphocytes can quickly gather together to defend against cancer cells. In cancer images, the purple regions around cancer are lymphocytes. For some patients, it is common that lymphocytes occur around the cancer cells and seldom appear around non-cancer tissues (in our dataset, no lymphocytes appear in non-cancer images) although lymphocytes itself are not considered as cancer tissues in medical research.

The main reason we set $K = 4$ clusters rather than 3 (the number of cancer types) is to show the MCIL's potential for exploring new subclasses from a vision perspective. Our method divides MTA cancer images into two clusters (green and yellow area) owning to different vision patterns. Since a clear definition of all subclasses is still not available, our method shows the promising potential of discovering a new classification standard for cancer research.

5. Conclusion

In this paper, we have introduced an integrated learning framework for classifying histopathology cancer images, performing segmentation, and obtaining cancer clusters via weakly supervised learning. The advantage of MCIL is evident over the state-of-the-art methods that perform the individual tasks. Experimental results demonstrate the efficiency and effectiveness of MCIL in detecting colon can-

cers.

Acknowledgments: This work was supported by Microsoft Research Asia. The work was also supported by ONR N000140910099, NSF CAREER award IIS-0844566, MSRA eHealth grant, Grant 61073077 from National Science Foundation of China and Grant SKLSDE-2011ZX-13 from State Key Laboratory of Software Development Environment in Beihang University in China. We would like to thank Lab of Pathology and Pathophysiology, Zhejiang University in China to provide data and help.

A. Verification for Remark 1

We verify Remark 1 (eqn. (9)): $g_j(g_k(p_{ij}^k)) = g_{jk}(p_{ij}^k) = g_k(g_j(p_{ij}^k))$ for each model. Given the number of clusters K and the number of instances m in each bag, we develop derivations for four models respectively:

For the NOR model:

$$\begin{aligned} g_k g_j(p_{ij}^k) &= 1 - \prod_k (1 - (1 - \prod_j p_{ij}^k)) \\ &= 1 - \prod_k (\prod_j p_{ij}^k) = 1 - \prod_{j,k} p_{ij}^k = g_{jk}(p_{ij}^k) \end{aligned} \tag{15}$$

For the GM model:

$$\begin{aligned} g_k g_j(p_{ij}^k) &= (\frac{1}{K}\sum_k (p_i^k)^r)^{\frac{1}{r}} = (\frac{1}{K}\sum_k ((\frac{1}{m}\sum_j (p_{ij}^k)^r)^{\frac{1}{r}})^r)^{\frac{1}{r}} \\ &= (\frac{1}{Km}\sum_{j,k} (p_{ij}^k)^r)^{\frac{1}{r}} = g_{jk}(p_{ij}^k) \end{aligned} \tag{16}$$

For the LSE model:

$$\begin{aligned} g_k g_j(p_{ij}^k) &= \frac{1}{r}\ln(\frac{1}{K}\sum_k \exp(r p_i^k)) \\ &= \frac{1}{r}\ln(\frac{1}{K}\sum_k \exp(r\frac{1}{r}\ln(\frac{1}{m}\sum_j \exp(r p_{ij}^k)))) \\ &= \frac{1}{r}\frac{1}{Km}\sum_{j,k}\exp(r p_{ij}^k) = g_{jk}(p_{ij}^k) \end{aligned} \tag{17}$$

For the ISR model:

$$g_k g_j(p_{ij}^k) = \sum_k \frac{p_i^k}{1-p_i^k} / (1 + \sum_k \frac{p_i^k}{1-p_i^k}) \tag{18}$$

$$\sum_k \frac{p_i^k}{1-p_i^k} = \sum_k \frac{\sum_j \frac{p_{ij}^k}{1-p_{ij}^k} / (1 + \sum_j \frac{p_{ij}^k}{1-p_{ij}^k})}{1 - \sum_j \frac{p_{ij}^k}{1-p_{ij}^k} / (1 + \sum_j \frac{p_{ij}^k}{1-p_{ij}^k})} = \sum_{j,k} \frac{p_{ij}^k}{1-p_{ij}^k} \tag{19}$$

$$g_k g_j(p_{ij}^k) = \frac{\sum_k \frac{p_i^k}{1-p_i^k}}{1 + \sum_k \frac{p_i^k}{1-p_i^k}} = \frac{\sum_{j,k} \frac{p_{ij}^k}{1-p_{ij}^k}}{1 + \sum_{j,k} \frac{p_{ij}^k}{1-p_{ij}^k}} = g_{jk}(p_{ij}^k) \tag{20}$$

Now we show $g_{jk}(p_{ij}^k) = g_k g_j(p_{ij}^k)$ for each *softmax* models. $g_{jk}(p_{ij}^k) = g_j g_k(pij^k)$ could also be given in the same way. Thus Remark 1 (eqn. (9)) could be verified.

References

[1] D. Altunbay, C. Cigir, C. Sokmensuer, and C. Gunduz-Demir. Color graphs for automated cancer diagnosis and grading. *IEEE Transaction on Biomedical Engineering*, 57(3):665–674, 2010. 2

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002. 1, 4, 5

[3] B. Babenko, P. Dollár, Z. Tu, and S. Belongie. Simultaneous learning and alignment: Multi-instance and multi-pose learning. In *ECCV workshop on Faces in Real-Life Images*, 2008. 1, 2, 3

[4] J. Bi and J. Liang. Multiple instance learning of pulmonary embolism detection with geodesic distance along vascular structure. In *CVPR*, 2007. 2

[5] L. E. Boucheron. *Object- and Spatial-Level Quantitative Analysis of Multispectral Histopathology Images for Detection and Characterization of Cancer*. PhD thesis, University of California, Santa Barbara, Mar 2008. 1, 2

[6] P. Dollár, B. Babenko, S. Belongie, P. Perona, and Z. Tu. Multiple component learning for object detection. In *ECCV*, 2008. 1, 2

[7] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2nd edition, Nov. 2001. 1, 4

[8] M. Dundar, S. Badve, V. Raykar, R. Jain, O. Sertel, and M. Gurcan. A multiple instance learning approach toward optimal classification of pathology slides. In *ICPR*, 2010. 2

[9] A. Esgiar, R. Naguib, B. Sharif, M. Bennett, and A. Murray. Fractal analysis in the detection of colonic cancer images. *IEEE Transaction on Information Technology in Biomedicine*, 6(1):54–58, 2002. 1, 5, 6

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

[11] P.-W. Huang and C.-H. Lee. Automatic classification for pathological prostate images based on fractal analysis. *IEEE Trans. Medical Imaging*, 28(7):1037–1050, 2009. 1, 2, 5

[12] 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. 1, 2

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, 2004. 5

[14] L. Lu, J. Bi, M. Wolf, and M. Salganicoff. Effective 3d object detection and regression using probabilistic segmentation features in ct images. In *CVPR*, 2011. 2

[15] A. Madabhushi. Digital pathology image analysis: opportunities and challenges. *Imaging in Medicine*, 1(1):7–10, 2009. 1

[16] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *NIPS*, 1997. 1

[17] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting algorithms as gradient descent. In *NIPS*, 2000. 3, 4, 5, 6

[18] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 5

[19] S. Park, D. Sargent, R. Lieberman, and U. Gustafsson. Domain-specific image analysis for cervical neoplasia detection based on conditional random fields. *IEEE Trans. Medical Imaging*, 30(3):867 –78, 2011. 1

[20] A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Trans. PAMI*, 29(10):1848–1852, 2007. 2

[21] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. Kumar, D. Verbel, A. Kotsianti, and O. Saidi. Multifeature prostate cancer diagnosis and gleason grading of histological images. *IEEE Trans. Medical Imaging*, 26(10):1366–78, 2007. 1, 2, 5, 6

[22] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE Trans. PAMI*, 21(10):1744–1757, 2010. 1

[23] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009. 1, 4, 5

[24] P. A. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. In *NIPS*, 2005. 1, 2, 3, 4, 5, 6

[25] Z.-J. Zha, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *CVPR*, 2008. 2

[26] D. Zhang, F. Wang, L. Si, and T. Li. M^{3}IC:maximum margin multiple instance clustering. In *IJCAI*, 2009. 1, 2

[27] Z.-H. Zhou and M.-L. Zhang. Multi-instance multilabel learning with application to scene classification. In *NIPS*, 2007. 2

[28] X. Zhu. Semi-supervised learning literature survey. *Computer Science TR 1530, University of Wisconsin-Madison*, 2008. 1