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

Yan Xu, Jun-Yan Zhu, Eric Chang and Zhuowen Tu
Microsoft Research Asia

Microsoft® Research 微软亚洲研究院

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.

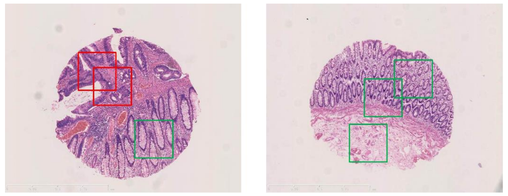
▪Supervised approaches: Require a large amount of accurately annotated data, which is not only labor-intensive and time consuming to obtain, but also intrinsically ambiguous.

▪Unsupervised learning methods: Ease the burden of manual annotation, but often at the cost of inferior results.

▪Weakly supervised learning scenario: Use coarse-grained annotations to aid automatic exploration of fine-grained information.

▪Clustering: Cluster different cancer cells to different subclasses. Fig. 4 shows how different cancer cells are mapped to different colors.

▪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) in an integrated learning framework under weakly supervised scenario.



(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).

Advantage

▪Develop an integrated system to perform pixel-level segmentation (cancer vs. non-cancer) and image-level classification.

▪Discover/identify the subclasses of various cancer tissue types as a universal protocol for cancer tissue classification is not yet available; this results in patch-level clustering of the cancer tissues.

▪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.

Method

- Given:

  a training dataset containing bags $X_i = \{x_{i1}, \ldots, x_{im}\}$

  bag labels $y_i \in \mathrm{Y} = \{-1, 1\}$

- Assuming:

  Hidden variable k^{th} which denotes whether the instance x_{ij} belongs to the k^{th} cluster.

- MCIL assumption:

  If one instance belongs to one of K clusters, this instance could be considered as a positive instance; only if at least one instance in a bag is labeled as positive, the bag is considered as positive.

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

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

- Loss function :

$$\mathrm{L}(h) = -\sum_{i=1}^{n} w_i (1(y_i = 1) \log p_i + 1(y_i = -1) \log(1 - p_i)).$$

- Softmax function, a differentiable approximation of max, is then introduced as follows:

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

- Bag probability:

$$p_i = g_j(g_k(p_{ij}^k)) = g_{jk}(p_{ij}^k) = g_{jk}(\sigma(2h_{ij}^k)).$$

- The weights w_{ij}^k and derivatives $-\frac{\partial \mathrm{L}}{\partial h_{ij}^k}$ could be given as:

$$w_{ij}^k = -\frac{\partial \mathrm{L}}{\partial h_{ij}^k} = -\frac{\partial \mathrm{L}}{\partial p_i} \frac{\partial p_i}{\partial p_{ij}^k} \frac{\partial p_{ij}^k}{\partial h_{ij}^k}.$$

$$\frac{\partial \mathrm{L}}{\partial p_i} = \begin{cases} -\frac{1}{p_i}, y = 1; \\ \frac{1}{1 - p_i}, y = -1. \end{cases} \qquad \frac{\partial p_i}{\partial p_{ij}^k} = \begin{cases} \frac{1 - p_i}{1 - p_{ij}^k}, \text{NOR}; & \frac{\exp(rp_{ij}^k)}{\sum_{j,k} \exp(rp_{ij}^k)}, \text{LSE}; \\ p_i \frac{(p_{ij}^k)^{r-1}}{\sum_{j,k} (p_{ij}^k)^r}, \text{GM}; & \frac{1 - p_i}{1 - p_{ij}^k})^2, \text{ISR}. \end{cases}$$

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

NOR, LSE, GM and ISR are the four models of softmax function.

- Get classifier:

$$h_t^k = \arg\min_h \sum_{ij} 1(h(x_{ij}^k) \neq y_i) \mid w_{ij}^k \mid$$

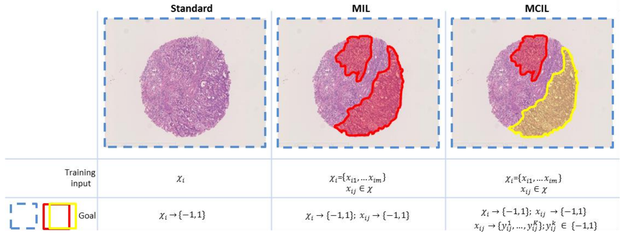


Figure 2: Distinct learning goals of supervised learning, MIL and MCIL.

Experiments

In the experiments, we apply our method on several cancer image datasets. Specific dataset component is shown in Table 1.

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

Table 1: Number of images in the datasets. Bianry, multi1 and multi2 are the three datasets. NC is noncancer image while MTA, LTA, MA and SRC are the four cancer types.

▪Image types:

Non-cancer (NC), Middle tubular adenocarcinoma (MTA), Low tubular adenocarcinoma (LTA), Mucinous adenocarcinoma (MA), and Signet-ring carcinoma (SRC).

▪Generic features:

L∗a∗b∗ Color Histogram, Local Binary Pattern, and SIFT.

Results

▪Image-level Classification

(a) classification results *(b) different softmax models*

Figure 3: Comparisons of image-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 the apparent advantages. (b) demonstrates the effect of using different soft-max functions.

▪Pixel-level Segmentation:

The F-measures of MCIL, MIL-Boost, and standard Boosting are 0.588, 0.231, and 0.297 respectively. Figure 4 shows some results of test data.

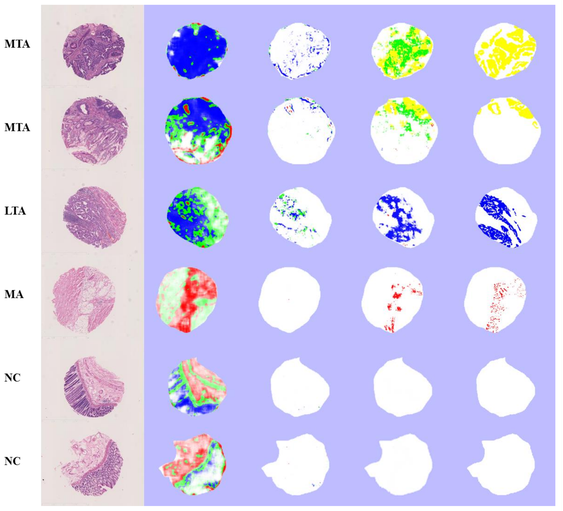


Figure 4: Image Types: (a): The original images. (b), (c), (d): The instance-level results (pixel-level segmentation and patch-level clustering) for standard Boosting + K-means, MIL + K-means, and MCIL. (e): The instance-level ground truth. Different colors stand for different types of cancer tissues.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

▪Patch-level Clustering:

MCIL obtains the clustering results at the same time of segmentation. 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.

MCIL is able to successfully discriminate cancer types since different types of cancer images are mapped to different clusters (See Figure 4).

An integrated learning framework of MCIL is better than separate two steps of instance-level segmentation and clustering.

Our method shows the promising potential of discovering a new classification standard for cancer research.

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 cancer.