Empowering Multiple Instance Histopathology Cancer Diagnosis by Cell Graphs

Melih Kandemir[1], Chong Zhang[2], and Fred A. Hamprecht[1]

[1] Heidelberg University HCI/IWR, Germany
[2] CellNetworks, Heidelberg University, Germany

Abstract. We introduce a probabilistic classifier that combines multiple instance learning and relational learning. While multiple instance learning allows automated cancer diagnosis from only image-level annotations, relational learning allows exploiting changes in cell formations due to cancer. Our method extends Gaussian process multiple instance learning with a relational likelihood that brings improved diagnostic performance on two tissue microarray data sets (breast and Barrett's cancer) when similarity of cell layouts in different tissue regions is used as relational side information.

1 Introduction

Developments in imaging techniques make increasingly large data sets of medical processes available, rendering annotation of data sets more difficult. Weakly supervised learning solves this problem by demanding labels only for large instance groups, while keeping prediction performance similar. A powerful weakly supervised learning framework is *multiple instance learning (MIL)* [9], which assumes a data set $\mathbf{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_B\}$ consisting of groups of observations called *bags*, and a vector of corresponding bag labels $\mathbf{y} = [y_1, \cdots, y_B]$. Every bag $\mathbf{X}_i$, in turn, consists of multiple instances $[\mathbf{x}_1^i, \cdots, \mathbf{x}_{N_b}^i]$. The specialty of MIL is to learn only from labels provided at the bag level, but not at the instance level: for binary classification, in a positive bag ($y_b = +1$), there exists at least one positively labeled instance in bag b, whereas in a negative bag ($y_b = -1$), all instances are known to have negative labels. MIL can easily be applied to classification of histology images by treating an image as a bag, and its regions (e.g. square patches) as an instance [4].

Relational learning [6] is a field of machine learning that proposes to exploit side information about relationships between data instances into the learning process. In addition to a set of instances and their ground-truth labels, relational methods leverage graphs indicating similarities of instance pairs for improved prediction accuracy. Relational side information is shown to be useful in many applications including web page categorization [3] and protein fold classification [12]. However, despite the possibility of constructing relational side information in many ways (e.g. spatial layout of cells, cross-existence of different cell types in different regions), their potential in tissue microarray (TMA) based cancer diagnosis applications has so far been largely ignored.

P. Golland et al. (Eds.): MICCAI 2014, Part II, LNCS 8674, pp. 228–235, 2014.

In this paper, we introduce a machine learning method that combines benefits of multiple instance learning and relational learning. Our method extends Gaussian process multiple instance learning (GPMIL) [8] with a likelihood function that explains the relations of inter-bag instances by forcing similar instance pairs to belong to the same class. We show in two computer-aided diagnosis (CAD) applications, malignant breast cancer diagnosis and Barrett's cancer diagnosis, that using the similarity of the spatial layout of cells between different image regions as side information brings a consistent increase in diagnostic accuracy. The source code of the proposed method is available under [1].

2 Related Work

The MIL framework has been shown to be useful in several digital pathology applications. Xu et al. [4] propose a colon cancer diagnosis and grading method that extends the boosting-based MIL approach [11] to multi-class classification based on clustering. GPMIL [8], which replaces the instance-level sigmoid likelihood of the standard Gaussian process classifier with a bag-level likelihood, has been used for detecting various skin diseases from biopsy images by [13].

Relational learning has been extensively studied within the Gaussian process framework. Mainstream approaches include calculating a relation dependent covariance matrix [10], and extending the likelihood by relational variables [3]. As will be clarified below, we follow the latter approach. No application of relational learning to cancer diagnosis from TMA images has been made prior to our work.

The only previous work that reconciles multiple instance learning and relational learning has been done by Zhang et al [12]. The model has been shown to perform protein fold classification with higher accuracy when alignment scores of protein pairs are used as relational side information.

3 Relational Gaussian Process Multiple Instance Learning

We are given a set of N observed data instances $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ partitioned into bags as $\mathbf{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_B\}$, where $\mathbf{X}_b = [\mathbf{x}_1^b, \cdots, \mathbf{x}_{N_b}^b]$ is the set of N_b instances belonging to bag b. Hence, $N = \sum_{b=1}^{B} N_b$. We are also given the corresponding observed binary bag labels $\mathbf{y} = [y_1, \cdots, y_B]$ such that $y_b \in \{-1, +1\}$, and a set $\mathcal{R}$ containing an observed set of triples (i, j, r_{ij}) indicating a relation r_{ij} between instances $\mathbf{x}_i$ and $\mathbf{x}_j$. A positive relation $r_{ij} = +1$ implies that instances i and j belong to the same class (i.e. f_i and f_j have the same sign), and a negative relation $r_{ij} = -1$ implies the opposite. The generative process of the proposed model, Relational Gaussian process multiple instance learning

[1] http://hci.iwr.uni-heidelberg.de/Staff/mkandemi/rgpmil.tar.gz

(RGPMIL), is as follows:

$$\mathbf{f}|\mathbf{X},\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K}), \tag{1}$$

$$y_b|\mathbf{f}_b \sim \frac{1}{1+(\sum_{i=1}^{N_b} e^{f_i^b})^{-y_b}}, \qquad \forall b \tag{2}$$

$$r_{ij}|f_i,f_j \sim \frac{1}{1+e^{-r_{ij}f_if_j}}. \qquad \forall (i,j,r_{ij}) \in \mathcal{R} \tag{3}$$

Here, $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\mathbf{K}$ is a Gram matrix constructed by evaluating any valid kernel function $k(\mathbf{x}_i,\mathbf{x}_j|\boldsymbol{\theta})$ on each pair of instances $\mathbf{x}_i$ and $\mathbf{x}_j$ in the data set, and $\boldsymbol{\theta}$ is the vector of kernel parameters. Equation 1 is a Gaussian process prior defined on the latent decision outputs f_i for each data instance. The vector $\mathbf{f} = [f_1,\cdots,f_N]$ contains the decision outputs of all instances, and $\mathbf{f}_b = [f_1^b,\cdots,f_{N_b}^b]$ contains those of instances of bag b. The sign of f_i determines the class the instance $\mathbf{x}_i$ belongs. Equation 2 is the multiple-instance likelihood approximating $\max(f_1^b,\cdots,f_{N_b}^b)$ by $\log\sum_i^{N_b}\exp(f_i^b)$, where $\sigma(z) = 1/(1+\exp(-z))$ is a sigmoid function. Combination of Equations 1 and 2 corresponds to GPMIL [8], which we extend with the relational likelihood in Equation 3. A variable r_{ij} is added for all pairs of instances with an observed relation. Equation 3 corresponds to the softened version of the link-likelihood suggested in [3]. The variable r_{ij} incorporates similarity of instances i and j into the learning process via coupling their latent decision output values f_i and f_j in the denominator. If $r_{ij} = +1$, the link likelihood is maximized when f_i and f_j have large values with the same sign, and the situation is opposite if $r_{ij} = -1$.

3.1 Inference

Within the Bayesian paradigm, a model is fit to data by inferring the posterior of model parameters given a prior distribution over model parameters and a data likelihood. The parameter set of RGPMIL is the latent decision output vector $\mathbf{f}$. Due to the non-conjugate likelihood functions in Equations 2 and 3, the posterior distribution $p(\mathbf{f}|\mathbf{y},\mathbf{R},\mathbf{X})$ is not available in closed form, where $\mathbf{R}$ is the $N \times N$ relation matrix having r_{ij} on its entries: $\mathbf{R}_{ij} = r_{ij}$. Thus, we approximate the posterior by Laplace approximation $p(\mathbf{f}|\mathbf{y},\mathbf{R},\mathbf{X}) \simeq \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}},\mathbf{H}^{-1})$, where $\hat{\mathbf{f}}$ is the estimated mode of the posterior and $\mathbf{H}$ is the negative Hessian of the logaritm of the posterior at its mode. We estimate the posterior mode by gradient search, for which we need the logarithm of the posterior:

$$\begin{aligned}
\Psi(\mathbf{f}) =& \log\ p(\mathbf{f}|\mathbf{y},\mathbf{R},\mathbf{X}) = \log p(\mathbf{f}|\mathbf{X}) + \sum_{b=1}^{B}\log p(y_b|\mathbf{f}_b) + \sum_{r_{ij}\in\mathcal{R}}\log p(r_{ij}|f_i,f_j) + const \\
=& -\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\log|\mathbf{K}| - \sum_{b}^{B}\log\left(1+\left(\sum_{i}^{N_b}e^{f_i^b}\right)^{-y_b}\right) \\
& - \sum_{r_{ij}\in\mathcal{R}}\log(1+e^{-r_{ij}f_if_j}) + const,
\end{aligned}$$

and its gradient with respect to all entries of $\mathbf{f}$:

$$\frac{\partial \Psi(\mathbf{f})}{\partial f_i^b} = - [\mathbf{K}^{-1}]_{(i_b,:)}\mathbf{f} + \sum_{r_{ij}=-1,+1} \frac{r_{ij} f_j}{1+e^{r_{ij} f_i^b f_j}} + \frac{e^{f_i^b} y_b \left(\sum_{j=1}^{N_b} e^{f_j^b}\right)^{-y_b-1}}{1+\left(\sum_{j=1}^{N_b} e^{f_j^b}\right)^{-y_b}},$$

where $[\mathbf{K}^{-1}]_{(i_b,:)}$ denotes the row of the inverse kernel matrix corresponding to ith instance of bag b. The Hessian of the negative log-posterior at the mode is $\mathbf{H} = -\mathbf{K}^{-1} - \mathbf{W}$, where

$$\mathbf{W}_{ij} = \left(r_{ij}(1+e^{r_{ij}\hat{f}_i\hat{f}_j}) - \hat{f}_i\hat{f}_j e^{r_{ij}\hat{f}_i\hat{f}_j}\right) \Big/ \left(1+e^{r_{ij}\hat{f}_i\hat{f}_j}\right)^2 + \mathbb{I}(i=j)\sigma(f_i)(1-\sigma(f_i)).$$

and $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the argument is true and 0 otherwise.

3.2 Marginal Likelihood and Hyperparameter Learning

The choice of kernel hyperparameters is known to have a significant effect on prediction performance. Thanks to the probabilistic nature of GPs, model hyperparameters can be tuned in a principled way using Type II Maximum Likelihood (empirical Bayes), where the marginal likelihood is maximized with respect to hyperparameters. Based on the Laplace approximated posterior, the marginal likelihood of our model is $p(\mathbf{y},\mathbf{R}|\mathbf{X}) = \int p(\mathbf{y},\mathbf{R}|\mathbf{f})p(\mathbf{f}|\mathbf{X})d\mathbf{f} = \int \exp(\Psi(\mathbf{f}))d\mathbf{f}$. When we perform second order Taylor expansion of $\Psi(\mathbf{f})$ around $\hat{\mathbf{f}}$, we have $\Psi(\mathbf{f}) \simeq \Psi(\hat{\mathbf{f}}) - \frac{1}{2}(\mathbf{f}-\hat{\mathbf{f}})^T\mathbf{H}_{\hat{\mathbf{f}}}(\mathbf{f}-\hat{\mathbf{f}})$. where $\mathbf{H}_{\hat{\mathbf{f}}}$ is the Hessian at the mode. Thus,

$$p(\mathbf{y},\mathbf{R}|\mathbf{X}) \simeq q(\mathbf{y},\mathbf{R}|\mathbf{X}) = \exp(\Psi(\hat{\mathbf{f}})) \int \exp\left(-\frac{1}{2}(\mathbf{f}-\hat{\mathbf{f}})^T\mathbf{H}_{\hat{\mathbf{f}}}(\mathbf{f}-\hat{\mathbf{f}})\right) d\mathbf{f}.$$

Solving the integral analytically and taking the logarithm of the result, we get:

$$\log q(\mathbf{y},\mathbf{R}|\mathbf{X}) = -\frac{1}{2}\hat{\mathbf{f}}^T\mathbf{K}^{-1}\hat{\mathbf{f}} + \log p(\mathbf{Y},\mathbf{R}|\hat{\mathbf{f}}) - \frac{1}{2}\log|\mathbf{I}+\mathbf{K}\mathbf{W}|.$$

When the learned $\hat{\mathbf{f}}$ vector is fixed, the derivative of the marginal likelihood with respect to kernel hyperparameters $\boldsymbol{\theta}$ becomes:

$$\frac{\log q(\mathbf{y},\mathbf{R}|\mathbf{X})}{\partial\theta_r} = \frac{1}{2}\mathbf{a}^T\left(\frac{\partial\mathbf{K}}{\partial\theta_r}\right)\mathbf{a} - \frac{1}{2}\mathrm{tr}\left((\mathbf{W}^{-1}+\mathbf{K})^{-1}\left(\frac{\partial\mathbf{K}}{\partial\theta_r}\right)\right),$$

where $\mathbf{a} = \mathbf{K}^{-1}\hat{\mathbf{f}}$. A local optimum can be found if $\hat{\mathbf{f}}$ and $\boldsymbol{\theta}$ are updated in turns by fixing one and optimizing the other using any gradient-based technique.

3.3 Prediction

For a trained model, we have the learned decision outputs $\hat{\mathbf{f}}$ and learned kernel hyperparameters $\hat{\boldsymbol{\theta}}$. For a newly-seen bag $\mathbf{X}^* = \{\mathbf{x}_1^*, \cdots, \mathbf{x}_{N_b}^*\}$, the corresponding latent decision outputs can be predicted in the same way as in Gaussian process classification: $\mathbb{E}(\mathbf{f}^*|\mathbf{y},\mathbf{X},\mathbf{X}^*) = \mathbf{k}^*\mathbf{K}^{-1}\hat{\mathbf{f}}$, where $\mathbf{k}^*$ is the vector of kernel values between training and test instances. Finally, the bag class probability can simply be calculated by plugging the point estimate of $\mathbf{f}^*$ into Equation 2.

4 Data Sets

We evaluate our method on two hematoxylin and eosin (H&E) stained TMA image data sets:

- **Malignant breast cancer data set:** This public data set [2] consists of 58 TMA image excerpts of 896×768 pixel size taken from breast cancer patients, 32 of which are at benign status, and 26 are malignant. The learning task is to classify images as benign and malignant [5]. We split each image into an equal-sized 7 × 7 grid.
- **Barrett's cancer data set:** This private data set consists of 214 whole-core images taken from the biopsy samples of the esophagus tissue of 97 Barrett's cancer patients. 145 of the images include tumorous regions, and 69 are normal. Average resolution of the images is 2179 × 1970. Each image is split into patches of 200 × 200 pixels.

For both data sets, each image is treated as a bag, and each patch as an instance. A bag is assigned a positive label if its corresponding image includes a diseased region, and a negative label otherwise. Each patch is represented by a 657-dimensional feature vector including the following features: mean SIFT descriptors, local binary patterns with 20 × 20-pixel cells, intensity histogram of 26 bins for each of the RGB channels, and mean feature vector of the cells lying in that patch. We extract 81 morphological and intensity-based features from each cell within a patch, as described in [7]. We avoid duplication of highly-correlated features and increase variance by reducing the feature dimensionality to 100 using principal component analysis.

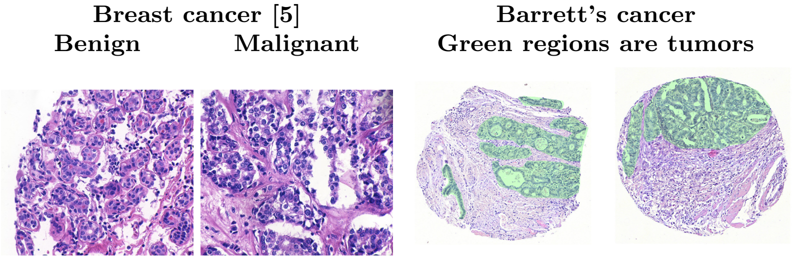


Fig. 1. Sample images from breast and Barrett's cancer data sets

We segment the breast cells using morphological filtering. We first perform a colour deconvolution to separate the H-stain channel from the color image [2], followed by a morphological opening filtering to reduce subtle textures in the H-channel. Breast cell nuclei are then segmented by detecting extended regional maximas. These filtering steps are done using Fiji and its plugins [3].

[2] `http://www.bioimage.ucsb.edu/research/biosegmentation`

[3] `http://fiji.sc/Fiji`

For the large esophagus cohort, we apply the segmentation scheme described in [7]. We classify pixels as tumor, non-tumor, stromal cells, and background from annotations provided on small regions (5.8 % of pixels) of two tumorous and two normal tissue cores. We then segment the cell nuclei by applying watershed transform on the probabilistic classifier output. We discard the four tissue cores used in supervised cell segmentation from further analysis. We finally keep tumor and non-tumor cells which are detected with high prediction probability (> 0.5), and discard the rest. Even though this policy causes many false negatives on the cell level, it captures higher-quality information of a subset of cells, which suffices to make better diagnostic decisions.

5 Cell Graphs as Relational Side Information

One effect of cancer common to all tissues is an increase in cell population, and one tissue-dependent effect is gland formation. While a normal colon tissue has glands which disappear during cancer, the situation in Barrett's cancer is vice versa: cancer causes cells to form glands. Motivated by this widespread effect of cancer on cell layout, we propose to construct relational side information from the spatial positions of segmented cells in order to capture the differences in cell formations caused by the disease status.

For each image patch we construct a graph that contains a node for each cell and an edge between each pair of cells whose centroid distance is smaller than τ pixels (roughly chosen to connect only two adjacent cells, being 15 for the breast cancer and 20 for the Barrett's cancer data sets). Let $\mathbf{A}_i^b$ and $\mathbf{A}_j^b$ be the adjacency matrices of cell graphs of patches i and j of bag b, and $\boldsymbol{\lambda}_i^b$ and $\boldsymbol{\lambda}_j^b$ be vectors containing the largest C_{min} eigenvalues of these matrices in decreasing order, where C_{min} is the minimum of cell counts of the two patches. We define the similarity of these graphs as $\mathrm{S}_{ij} = \exp(-||\boldsymbol{\lambda}_i^b - \boldsymbol{\lambda}_j^b||_2)$. Finally, we draw a positive link $r_{ij} = +1$ between each pair of patches i and j whose similarity is larger than the mean of all pairs of patches in the data set. The resultant relation matrix $\mathbf{R}$ constructed from the training bags are then added to our model as side information.

Figure 2 shows cell graphs constructed for sample tumorous and normal patches taken from one core of the Barrett's cancer data set. Cancer causes longer connected components of adjacent cells due to gland formation (Tumor 1), and uncontrolled proliferation (Tumors 2 and 3). The aforementioned graph similarity metric gives larger values for the majority of the patches sharing the same disease status than the ones with opposite disease status, as illustrated by the Hinton diagram shown on the right.

6 Results and Discussion

For both data sets, we evaluate our method using 4-fold cross validation. For the breast cancer data set, we train on three data splits and evaluate on one, due

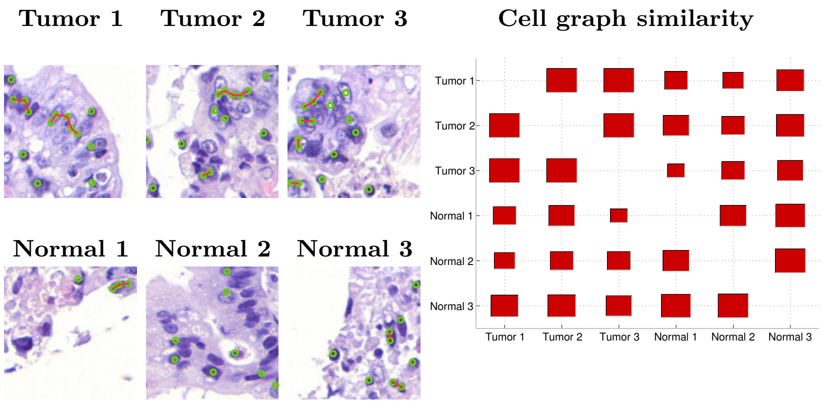


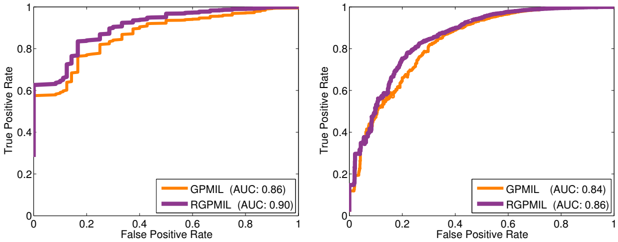
Fig. 2. Left: Cell graphs of tumorous and normal patches taken from the same core of the Barrett's cancer data set. Green dots are tissue cells detected with high confidence and red lines connect cells closer to each other than 20 pixels. Tumorous patches tend to contain larger connected components. **Right:** Hinton diagram of patch similarity. Higher similarity is shown by a larger square.

to its small scale. On the other hand, for the Barrett's cancer data set, we train on one split and evaluate on three. We use the radial basis function (RBF) as kernel, and tune its length scale as described in Section 3.2.

Figure 3 shows the receiver operating characteristics (ROC) curves of GPMIL and its relational extension (RGPMIL). Relational side information improves the area under ROC curve of GPMIL 4 percentage points in breast cancer and 2 percentage points in Barrett's cancer. According to paired t-test, improvement gained by using cell graphs as relational side information is statistically significant for breast cancer data set ($p < 0.027$), and is not significant for Barrett's cancer data set ($p < 0.071$). As a reference, we compare our results with three existing MIL methods. Area Under ROC Curve of these methods averaged over two data sets are: EMDD [14] 0.67, MILBoost [11,4] 0.83 , and MI-SVM [1] 0.82, compared to 0.85 reached by GPMIL and 0.88 reached by RGPMIL. MILBoost and MI-SVM give competitive performance for Barrett's cancer, but are more than 10 percentage points below RGPMIL for breast cancer.

Average training time of RGPMIL and GPMIL are as follows: **Breast cancer:** RGPMIL=10.3 seconds versus GPMIL=8.4 seconds, **Barrett's cancer:** RGPMIL=43.0 seconds versus GPMIL=33.9 seconds on one 3.20GHz CPU and 64 GB memory. Hence, training RGPMIL is slightly but not drastically slower than GPMIL.

The proposed method effectively incorporates cell graphs as relational side information into the diagnosis process, and provides a consistent performance increase over its non-relational counterpart in both applications. For the Barrett's cancer data set, we observe that in 75.4 % of the cores, patches with the same ground-truth label to have a higher average similarity score than the ones



(a) Breast cancer data set (b) Barrett's cancer data set

Fig. 3. ROC curve for GPMIL [8], and our method RGPMIL for two tissue types. Incorporating cell graph similarity as relational information brings accuracy gain in both diagnostic applications.

with opposite labels. We attribute the superior performance of RGPMIL over GPMIL to its ability to exploit this high correlation between label and cell graph similarity for diagnosis.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS (2003)
2. Ruifrok, A.C., et al.: Quantification of histochemical staining by color deconvolution. Anal. Quant. Cytol. Histol. 23, 291–299 (2001)
3. Sindhwani, W., et al.: Relational learning with Gaussian processes. In: NIPS (2007)
4. Xu, Y., et al.: Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In: CVPR (2012)
5. Gelasca, E.D., et al.: Evaluation and benchmark for biological image segmentation. In: ICIP (2008)
6. Getoor, L., Taskar, B.: Introduction to statistical relational learning. MIT Press (2007)
7. Kandemir, M., Feuchtinger, A., Walch, A., Hamprecht, F.A.: Digital Pathology: Multiple instance learning can detect Barrett's cancer. In: ISBI (2014)
8. Kim, M., Torre, F.: Gaussian processes multiple instance learning. In: ICML (2010)
9. Maron, O., et al.: A framework for multiple-instance learning. In: NIPS (1998)
10. Silva, R., Chu, W., Ghahramani, Z.: Hidden common cause relations in relational learning. In: NIPS (2007)
11. Viola, P., et al.: Multiple instance boosting for object detection. In: NIPS (2005)
12. Zhang, D., Liu, Y., Si, L., Zhang, J., Lawrence, R.D.: Multiple instance learning on structured data. In: NIPS (2011)
13. Zhang, G., Yin, J., Li, Z., Su, X., Li, G., Zhang, H.: Automated skin biopsy histopathological image annotation using multi-instance representation and learning. BMC Medical Genomics 6(suppl. 3), S10 (2013)
14. Zhang, Q., Goldman, S.A.: EM-DD: An improved multiple-instance learning technique. In: NIPS (2001)