A Discriminative Data-Dependent Mixture-Model Approach for Multiple Instance Learning in Image Classification

Qifan Wang, Luo Si, and Dan Zhang

Department of Computer Science
Purdue University
West Lafayette, IN, USA, 47907-2107
{wang868,lsi,zhang168}@purdue.edu

Abstract. Multiple Instance Learning (MIL) has been widely used in various applications including image classification. However, existing MIL methods do not explicitly address the multi-target problem where the distributions of positive instances are likely to be multi-modal. This strongly limits the performance of multiple instance learning in many real world applications. To address this problem, this paper proposes a novel discriminative data-dependent mixture-model method for multiple instance learning (MM-MIL) approach in image classification. The new method explicitly handles the multi-target problem by introducing a data-dependent mixture model, which allows positive instances to come from different clusters in a flexible manner. Furthermore, the kernelized representation of the proposed model allows effective and efficient learning in high dimensional feature space. An extensive set of experimental results demonstrate that the proposed new MM-MIL approach substantially outperforms several state-of-art MIL algorithms on benchmark datasets.

1 Introduction

With the pervasion of digital images, automatic image classification has become increasingly important. Multiple-instance learning (MIL) [2] is a useful technique in machine learning that addresses the classification problem of a bag of data instances. In multiple instance learning, each bag is composed of multiple data instances associated with input features. The purpose of MIL is to accurately predict bag level labels based on all the instances in each bag with the assumption that a bag is labeled positive if at least one of its instances is positive, whereas a negative bag only contains negative instances. In the case of image classification, each image is treated as a bag and different regions inside the image are viewed as individual data instances [15].

The advantage of MIL ascribes to the fact that in training it only requires the label information of a bag instead of individual instances in the bag. However, due to the label ambiguity in the instances, traditional supervised classification methods may not be directly applied to MIL framework. Existing methods in solving MIL problem fall into two categories. The first category is generative model based algorithms, such as axis parallel hyper-rectangles [2], Diverse Density (DD) [9] and Expectation Maximization

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part IV, LNCS 7575, pp. 660–673, 2012.

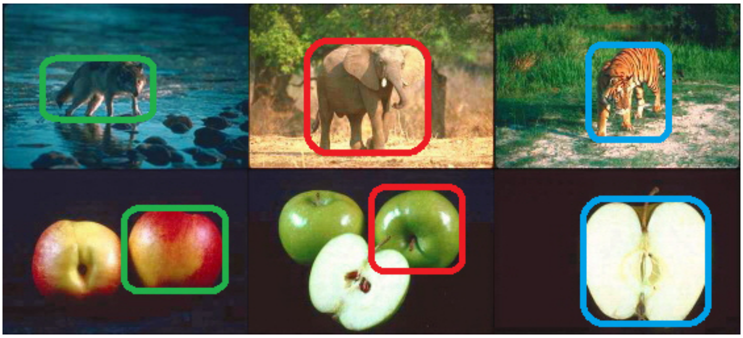

Fig. 1. Six images from COREL dataset. The top three images have a common concept 'animal'. The bottom three images form a concept 'apple'. Different colors represent different clusters the instances lie in.

DD (EM-DD) [10]. For example, EM-DD generates data instances in bags with their labels in a joint manner. The second category is discriminative model based methods including DD-SVM [7], MI-SVM [1], MILES [8], etc. These methods model the labels of bags and data instances by the input features of data instances or bags. For example, some methods based on SVM map features into a high dimensional feature space, with a non-linear function, and then apply the standard kernelized large-margin SVM framework to train a classifier from the constructed new features. These large margin discriminative methods often generate more robust results compared to the generative algorithms.

However, most existing multiple instance learning algorithms do not explicitly address the multi-target problem, where positive instances often tend to have multi-modal distributions or lie in different clusters in many real word applications. Two examples are provided as follows. In the first example the concept is 'animal'. There are various kinds of animals in the training samples like fox, elephant and tiger (top row in Fig.1). Different species have different characteristics in terms of color, size, shape, etc. Therefore, the positive instances come from distinct clusters and form a multi-modal distribution in the feature space. Even if the concept is relatively 'small', the instances could still form several compact clusters. In another scenario, the concept is 'apple'. The images in the bottom row in Fig.1 show three training examples. All the three images contain the concept 'apple'. However, the positive targets in the pictures are different as red apple, green apple and half-apple, which form different clusters. Please note that the multi-target problem of multiple instance learning is different from multi-class multiple instance learning since no specific class information is available for the diversified representation of positive instances and all positive bags are labeled in the same manner.

To address this problem, this paper proposes a novel data-dependent Mixture-Model MIL (MM-MIL) approach in the discriminative learning framework to handle the multi-modal distributions of positive data instances for image classification with multiple instance learning. In particular, a set of latent variables are introduced to represent the clusters associated with each data instance based on a multinomial logit model. Within each cluster, a logistic regression model is utilized to generate labels given the input features of individual data instances. These two models are integrated together for representing the assumption of multiple instance learning as each positive bag contains at least one positive data instance and each negative bag does not contain any positive instance. Furthermore, a kernelized presentation of the new method is proposed to allow effective and efficient learning in high-dimensional space. An efficient inference algorithm is derived for the proposed method based on a combination of Expectation and Maximization (EM) method and gradient descent optimization.

To our best knowledge, the MM-MIL model is the first concrete research work that explicitly addresses the multi-target problem in multiple instance learning. The main contributions of this paper are: First, the proposed MM-MIL model introduces a data-dependent mixture model that effectively captures the multi-modal distributions among the instances and formalizes the problem into a regularization framework. Second, we introduce an efficient inference algorithm to solve the optimization problem by combining the EM method and gradient descent scheme. Third, a kernelization framework is proposed to allow effective and efficient learning, especially for large scale image dataset.

The rest of the paper is organized as follows. Section 2 discusses the related work on MIL-based image classification. Section 3 proposes the novel MM-MIL method, which includes the problem formulation, the inference algorithm and the kernelization framework. We will also discuss the relationship between MM-MIL and some other existing MIL algorithms. Section 4 presents an extensive set of experimental results on different datasets for comparing the MM-MIL method with several state-of-the-art MIL algorithms. Section 5 concludes and points out some possible future research directions.

2 Related Work

Image classification algorithms based on multi-instance learning (MIL) model the relationship between labels and regions [2,10,7,8]. An image is treated as a bag consisting of multiple instances, ie, regions. Existing MIL algorithms can be divided into two categories, generative models and discriminative models. Generative model methods, like EM-DD [10], try to learn a single target distribution to generate instances/bags and their labels in a joint manner. Discriminative models focuses on modeling data/bag labels given features of data instances, which include MI-SVM [1] and MILES [8] based on kernelized support vector machine.

Many generative algorithms try to predict bag labels by first inferring the hidden labels of individual instances. The Diverse Density (DD) [9] approach uses a scaling and gradient search algorithm to find the prototype points in the instance space with the maximal DD value. Zhang and Goldman [10] combined the idea of Expectation-Maximization (EM) with DD and developed an algorithm, EM-DD, to search for the most likely concept. These methods are quite efficient in learning, but they are based on the assumption that that all positive instances form a tight cluster in the feature space [3], which is not realistic in applications with diversified positive instances. The research work in [9] briefly mentioned that it is possible to model multiple concepts within a generative model, but no concrete prior research work has been conducted for this. We also designed the first concrete generative multiple instance learning algorithm for multiple concepts in this paper. But the empirical results and discussions in section 4 show that our discriminative data-dependent mixture-model outperforms the generative model for multiple instance learning with multiple concepts.

Most discriminative methods attempt to directly predict bag labels in a large margin framework. DD-SVM [7] selects a set of instances using the DD function, and then a SVM is trained based on the bag-level features summarized by these selected instances. In MI-SVM [1], Andrews *et al* formulated MIL as a mixed integer quadratic programming problem. Integer variables are used to select a positive instance from each positive bag. A standard SVM framework is introduced to tune the variables. In the work of MILES [8], bags are embedded into a feature space defined by all the instances. 1-norm SVM is applied to train the bag-level classifiers. Some methods based on instance-level information were also proposed. Yang *et al* [11] proposed an Asymmetric Support Vector Machine-based MIL algorithm (ASVM-MIL) by defining an asymmetric loss function to exploit instance labels. Ray *et al* [17] extended the DD framework by using a Logistic Regression algorithm to estimate the equivalent probability for an instance and a $softmax$ function is used to combine the instance-level information to predict the bag label. Boosting methods such as MILBoost [13] translated MIL into an AdaBoost framework, where the combination function (eg, Integrated Segmentation and Recognition (ISR) or noisy-or) is applied to combine instance labels into bag label. Fu *et al* [3] proposed an instance selection MIL approach which aims to handle large scale data. A kernel density estimator is first learned from all the negative instances in negative bags to reduce the number of positive candidates. One instance per positive bag is selected to represent the concept. Standard SVM is then applied to train the classifier based on constructed bag-level features. Discriminative methods are often more robust and achieve improved performance compared to the generative approaches.

Recently, several MIL methods [23,27] has been used for online visual tracking. A discriminative classifier is trained in an online manner to separate the object from the background. Qi *et al* [6] explicitly modeled the inter-dependencies between instances by using concurrent tensors to better capture images' inherent semantics. Rank-1 tensor factorization is applied to obtain the label of each instance. A kernelization framework is then used for learning. In the work [25,32], Random Forest methods have been proposed to dealing with the multi-class/multi-label problem in MIL. Hidden class labels are defined inside bags as random variables. These random variables are optimized by training random forests and using a fast iterative homotopy method for solving the non-convex optimization problem. The multi-label issue is also addressed in work [5,26], where multi-label MIL algorithms are introduced to simultaneously captures both the connections between semantic labels and regions and the correlations among the labels based on hidden conditional random fields. Most recently, Dan *et al* [29,30] introduce the un-supervised learning methods under the maximum margin principle for multiple instance clustering, where bag labels are not utilized in training. A semi-supervised MIL approach [28] is also proposed by him in learning structured data. Multiple instance active learning for localized content based image retrieval is proposed in [32].

However, none of existing works in multiple instance learning addresses the multi-target problem where positive instances may lie in different clusters in the feature space.

To address this issue, we propose the MM-MIL algorithm, which will be described in the next section.

3 Mixture Model Multiple Instance Learning

This section presents the novel MM-MIL model that explicitly addresses the multi-target problem in multiple instance learning. We first introduce some notations. Let bag set $B = \{B_i\}, i = 1, 2, \ldots, N$. Let $L = \{l_i\}$ denotes the bag labels. $l_i = 1$ or 0 indicates B_i is a positive or negative bag. Let $B_i = \{B_{ij}\}, j = 1, 2, \ldots, N_i$ where B_{ij} is the j^{th} instance in bag B_i. Let $y_i = P(+|B_i)$ denotes the probability of B_i being a positive bag and $y_{ij} = P(+|B_{ij})$ denotes the probability of B_{ij} being a positive instance.

3.1 Problem Formulation

Given B and L, our goal is to maximize the following conditional probability:

$$P(L|B) = \prod_{i=1}^{N} P(l_i|B_i) = \prod_{i=1}^{N} P(+|B_i)^{l_i}(1 - P(+|B_i))^{1-l_i} \tag{1}$$

In our method, we make a similar choice like many existing multiple instance learning works, eg, IS-MIL [3], for modeling $P(+|B_i)$ as follows:

$$P(+|B_i) = \max_j P(+|B_{ij}) \tag{2}$$

which means we select the instance with the maximum probability to be positive to represent the bag. This is also consistent with the MIL assumption. It is also possible to make other choices like a $softmax$ [17] to combine instance labels.

As we discussed in section 2, traditional MIL algorithms do not explicitly address the multi-target problem when modeling the probability of an instance being positive, ie, $P(+|B_{ij})$. For example, say the concept is 'animal' (Fig.1), the positive instance could lie in a cluster that stands for 'tiger' where the bag should be labeled as positive. It is also possible that the instance comes from an 'elephant' cluster which also indicates the bag positive. In order to capture the multi-modal distribution, we encode a data-dependent mixture model on $P(+|B_{ij})$ assuming that there are M clusters that represent the M targets in the feature space. A latent variable z_m is introduced to denote the m^{th} cluster that the instance lies in. Then the probability of an instance to be positive can be written as:

$$P(+|B_{ij}) = \sum_{m=1}^{M} P(+|z_m, B_{ij})P(z_m|B_{ij}) \tag{3}$$

The first term $P(+|z_m, B_{ij})$ indicates the probability of B_{ij} being positive within cluster z_m. We use a logistic regression model for the purpose, which is similar with the logistic function chosen in [13] and [19]:

$$P(+|z_m, B_{ij}) = \frac{1}{1 + \exp(-t_m^T B_{ij})} \tag{4}$$

where t_m is the model parameter in the m^{th} cluster. The second term, $P(z_m|B_{ij})$, in Eqn.3 indicates the probability that instance B_{ij} lies in the cluster z_m , which is actually a multi-class distribution and we apply a multinomial logit model to capture the underlying probability:

$$P(z_m|B_{ij}) = \frac{\exp(w_m^T B_{ij})}{\sum_{r=1}^{M} \exp(w_r^T B_{ij})} \tag{5}$$

where w_m is the model parameter. Both two parts in the mixture model are dependent on the data instance B_{ij}, which is more flexible to capture the dependencies among instances. Let $y_{ijm} = P(+|z_m, B_{ij})$, $\theta_{ijm} = P(z_m|B_{ij})$. Note that $\sum_m \theta_{ijm} = 1$ for every instance. Substituting Eqn. 2,3,4 and 5 into Eqn.1 and taking the negative logarithm on both sides we have:

$$E = -\sum_{i=1}^{N} \left((l_i \ln(\max_j \sum_{m=1}^{M} y_{ijm}\theta_{ijm}) + (1-l_i)\ln(1 - \max_j \sum_{m=1}^{M} y_{ijm}\theta_{ijm}) \right) \tag{6}$$

Maximizing the probability in Eqn.1 is equivalent to minimize Eqn.6. In order to avoid overfitting, a regularizer is introduced on the model parameters, w_m and t_m. Then we obtain the following optimization problem:

$$\begin{aligned} \min_{w,t} \quad & -\sum_{i=1}^{N} \left((l_i \ln(\max_j \sum_{m=1}^{M} y_{ijm}\theta_{ijm}) + (1-l_i)\ln(1 - \max_j \sum_{m=1}^{M} y_{ijm}\theta_{ijm}) \right) \\ & \qquad + \lambda \sum_{m=1}^{M} ||w_m||^2 + \beta \sum_{m=1}^{M} ||t_m||^2 \end{aligned} \tag{7}$$

where λ and β are weight parameters. We now describe an iterative EM and gradient descent algorithm for solving the above optimization problem.

3.2 Inference Algorithm

Directly minimizing Eqn.7 is intractable, as many terms are coupled together and a max function makes it non-differentiable. The EM framework is a powerful tool in learning mixture models [16]. In this section, we first derive an upper bound for Eqn.7 and then an iterative EM scheme is developed to solve the optimization problem.

Inspired by IS-MIL [3] and MI regression [21], in the E-step of each iteration, we remove the max function in Eqn.6 by choosing one instance per bag which has the maximum probability to be positive based on the previous w and t as follows:

$$j^* = arg \max_j \sum_{m=1}^{M} y_{ijm}\theta_{ijm} \tag{8}$$

Denote $y_{im} = y_{ij^*m}$, $\theta_{im} = \theta_{ij^*m}$ since j^* is fixed during the current iteration. Using the fact $\sum_m \theta_{im} = 1$, we can obtain $1 - \sum_{m=1}^{M} y_{im}\theta_{im} = \sum_{m=1}^{M} \theta_{im}(1 - y_{im})$. Then Eqn.6 can be written as:

$$\begin{aligned} E &= -\sum_{i=1}^{N}\left(l_i \ln(\sum_{m=1}^{M} y_{im}\theta_{im}) + (1-l_i)\ln(1-\sum_{m=1}^{M} y_{im}\theta_{im})\right) \\ &= -\sum_{i=1}^{N}\left(l_i \ln(\sum_{m=1}^{M} \theta_{im}y_{im}) + (1-l_i)\ln(\sum_{m=1}^{M} \theta_{im}(1-y_{im}))\right) \end{aligned} \tag{9}$$

We now establish an upper bound of Eqn. 9 with Jensen's inequality by observing that logarithm function is a concave function and $\sum_m \theta_{im} = 1$.

$$E \leq -\sum_{i=1}^{N}\sum_{m=1}^{M} \theta_{im}(l_i \ln y_{im} + (1-l_i)\ln(1-y_{im})) \tag{10}$$

Denote $\gamma_{im} = l_i \ln y_{im} + (1-l_i)\ln(1-y_{im})$. In M-step, using a similar divide-and-conquer strategy in [24], we minimize the above upper bound plus regularization terms by splitting it into two slightly simpler sub-problems. The idea is that we first fix $\theta_{im} = \theta_{im}^p$ that is obtained from the previous iteration, and then find t which optimize the following sub-problem:

$$SP1: \ -\sum_{i=1}^{N}\sum_{m=1}^{M} \theta_{im}^p \gamma_{im} + \beta\sum_{m=1}^{M} ||t_m||^2 \tag{11}$$

Furthermore, we can fix $y_{im} = y_{im}^p$ that gives us γ_{im}^p and solve for the following optimization problem for γ:

$$SP2: \ -\sum_{i=1}^{N}\sum_{m=1}^{M} \theta_{im}\gamma_{im}^p + \lambda\sum_{m=1}^{M} ||w_m||^2 \tag{12}$$

$SP1$ is essentially a combination of weighted logistic regression and $SP2$ can be viewed as a multi-class logistic regression. A direct gradient descent scheme could be applied for solving these two sub-problems. We refer to chapter 4.3 in [22] for full details. By solving $SP1$ and $SP2$ iteratively in the M-step, the obtained optimal solutions of w_m^* and t_m^* are then substituted into Eqn.8 to update the instance chosen from each bag.

3.3 Kernelization Framework

In this section, we will seek for optimal functions defined over the feature space on the basis of a kernelized representation of two sub-problems, $SP1$ and $SP2$. Consider $SP1$ first, since the objective function is point-wise, which only defines on the value of $t_m^T B_{ij}$ at the instances $\{B_{ij^*} : 1 <= i <= N\}$, based on the generalized representer theorem [20], the minimizer exists and has a representation of the form:

$$t_m^T B_{i'r} = \sum_{i=1}^{N} \alpha_{mi}^t k(B_{i'r}, B_{ij^*}) = \boldsymbol{k}_{B_{i'r}}^T \boldsymbol{\alpha_m^t} \tag{13}$$

where $k(B_{i'r}, B_{ij})$ is a kernel function defined on the feature space of instance. A Gaussian Kernel is defined as $k(B_{i'r}, B_{ij}) = exp(-\frac{||B_{i'r} - B_{ij}||^2}{2\sigma^2})$, σ^2 is the radius parameter. Substituting Eqn.13 into $SP1$, we obtain:

$$SP1_{ker} : -\sum_{m=1}^{M}((\boldsymbol{\theta l})_{\boldsymbol{m}}^{\boldsymbol{T}}\boldsymbol{K}\boldsymbol{\alpha_m^t} - \boldsymbol{\theta_m^T}\ln(1+\exp(\boldsymbol{K}\boldsymbol{\alpha_m^t}))) + \beta\sum_{m=1}^{M}(\boldsymbol{\alpha_m^t})^{\boldsymbol{T}}\boldsymbol{K}\boldsymbol{\alpha_m^t} \quad (14)$$

where $\boldsymbol{\theta_m^T} = [\theta_{1m}^p, \ldots, \theta_{Nm}^p], (\boldsymbol{\alpha_m^t})^{\boldsymbol{T}} = [\alpha_{m1}^t, \ldots, \alpha_{mN}^t], (\boldsymbol{\theta l})_{\boldsymbol{m}}^{\boldsymbol{T}} = [\theta_{1m}^p l_1, \ldots, \theta_{Nm}^p l_N]$ and $\boldsymbol{K}$ is the Gram matrix with the kernel function defined above. To solve $SP1_{ker}$, we derive the partial derivative w.r.t. $\boldsymbol{\alpha_m^t}$:

$$\frac{\partial SP1_{ker}}{\boldsymbol{\alpha_m^t}} = -(\boldsymbol{\theta l})_{\boldsymbol{m}}^{\boldsymbol{T}}\boldsymbol{K} + \boldsymbol{\theta_m}^{\boldsymbol{T}}\frac{\exp(\boldsymbol{K}\boldsymbol{\alpha_m^t})}{1+\exp(\boldsymbol{K}\boldsymbol{\alpha_m^t})}\boldsymbol{K} + 2\beta(\boldsymbol{\alpha_m^t})^{\boldsymbol{T}}\boldsymbol{K} \quad (15)$$

With this obtained gradient, L-BFGS quasi-Newton method [18] is applied to solve this optimization problem. Similar to the work [12] and [4], the minimizer of $SP2$ has a form:

$$w_m^T B_{i'r} = \sum_{i=1}^{N}\alpha_{mi}^w k(B_{i'r}, B_{ij^*}) = \boldsymbol{k}_{B_{i'r}}^T\boldsymbol{\alpha_m^w} \quad (16)$$

Substituting Eqn.16 into $SP2$, we obtain:

$$SP2_{ker} : -\sum_{i=1}^{N}\sum_{m=1}^{M}\gamma_{im}^p\frac{\exp(\boldsymbol{k}_{\boldsymbol{B_{ij*}^T}}\boldsymbol{\alpha_m^w})}{\sum_r\exp(\boldsymbol{k}_{\boldsymbol{B_{ij*}^T}}\boldsymbol{\alpha_r^w})} + \lambda\sum_{m=1}^{M}(\boldsymbol{\alpha_m^w})^{\boldsymbol{T}}\boldsymbol{K}\boldsymbol{\alpha_m^w} \quad (17)$$

The scheme for solve $SP2_{ker}$ is contained in [12], we refer to section 5 in [12] for details on the optimization algorithm of the above multi-class kernel logistic regression. The complete kernelization framework for MM-MIL is shown in Table 1. Note that in the kernelization framework, the parameters are $\boldsymbol{\alpha^t}$ and $\boldsymbol{\alpha^w}$, which are updated in the M-step and are fixed and utilized to calculate y_{ijm} and θ_{ijm} in the E-step.

3.4 Discussion

In the novel MM-MIL model, M is the number of latent clusters formed by the instances. Different M will have different behavior. When M equals 1, which means we assume all instance comes from one cluster, then Eqn.9 becomes:

$$E = -\sum_{i=1}^{N}(l_i\ln y_i + (1-l_i)\ln(1-y_i)) \quad (18)$$

Now we discuss the relationship between our MM-MIL and some previous methods when $M = 1$. If choosing $\ln y_i$ to be a quadratic loss function, Eqn.18 is exactly the EM-DD model. When modeling $\ln y_i$ by a logistic loss function, the above model turns out to be MI-Regression in work [21]. If putting a hinge loss function on $\ln y_i$, then Eqn.18 could be optimized using a standard SVM framework in a similar way to MI-SVM [1] and MILES [8]. With a value of M larger than 1, ie, the latent number of

Table 1. Our full kernelized MM-MIL inference framework

Initialize M,λ,β,σ and $\boldsymbol{K}$
Initialize parameters $\boldsymbol{\alpha^t}$ and $\boldsymbol{\alpha^w}$
Start EM iterations
E-step:
Calculate y_{ijm} based on Eqns.4 and 13
Calculate θ_{ijm} based on Eqns.5 and 16
Select one instance per bag from Eqn.8
M-step:
Obtain $\boldsymbol{\alpha^t}$ by solving $SP1_{ker}$
Obtain $\boldsymbol{\alpha^w}$ by solving $SP2_{ker}$
Update θ_{im} and γ_{im} by Eqns.4,5,13 and 16
Repeat the above three steps until convergence
Update $\boldsymbol{\alpha^t}$ and $\boldsymbol{\alpha^w}$ repeat EM iteration until convergence

clusters increase, which makes our model more flexible in modeling the dependencies between the instances. The desired value of M can be obtained by cross-validation or utilizing some model selection criterions like the Bayesian Information Criterion. This work uses cross validation and the empirical studies in section 4 show that robust classification results can often be obtained with a reasonably wide range of M values.

4 Experimental Results

In this section, the MM-MIL is evaluated with three configurations of experiments. First, MM-MIL is evaluated on several multi-target datasets to show the advantage of data-dependent mixture model against several existing algorithms in this setting. Second, MM-MIL is compared with existing MIL approaches in image classification on the commonly used COREL and SIVAL benchmark datasets. Third, we provide more experimental results to study the choice of M in terms of classification accuracy.

Each image is a bag and segments are instances. A set of low-level features is extracted from each segment to represent an instance, including color correlogram, color moment, region size, wavelet texture and shape. Some model parameters in our experiment are Gaussian Kernel radius σ^2, and the weight parameters λ and β. We apply a twofold cross-validation on the training set to obtain the optimal values. σ^2 is chosen from 1 to 15 where λ and β are selected from 0.01, 0.1, 1,10,100. The number of hidden clusters M is picked in the same manner from 1 to 15. During each experiment, images are randomly partitioned into two halves to form the training and the testing sets. Each experiment is repeated 10 times and the average results are calculated.

4.1 Evaluation on Multi-Target Datasets

In order to illustrate the ability of MM-MIL in capturing the multi-modal concepts, we merge several categories that form similar concepts together into a lager dataset. Within our experiment, we construct three such merged data sets. The first merged data set, we refer to MergeData1, is collected from the *Tiger*, *Fox* and *Elephant* data set [1] which

Fig. 2. Examples of positive instances selected from different clusters (three different colors). The left three columns are from MergeData2 and the right three columns are from MergeData3. The first and third row contains twelve original images and the second and fourth row shows the corresponding segmented regions.

form a general concept ‘animal’. The are 600 images in MergeData1 with 300 positive images and 300 negative ones. The second data set, MergeData2, is mixed from three SIVAL categories, ie, *DataMiningBook*, *RapBook* and *StripedNotebook*, containing a common concept ‘book’. MergeData3 is combined by another three classes, *CardboardBox*, *FabricSoftenerBox* and *GreenTeaBox*, from SIVAL data set, where ‘box’ is the ideal concept. Both MergeData2 and MergeData3 contain 360 images with half positive images and half negative images, where the negative ones are randomly chosen from other categories.

Various measurements can be applied for evaluating the performance. In our experiments we will use AUC (area under the ROC curve), which is a widely used metric in multi-instance learning tasks. The ROC curve shows the relationship between the true positive rate and the false positive rate, and AUC measures the probability that a randomly chosen positive image will be ranked higher than a randomly chosen negative image [6].

We compare our MM-MIL with EM-DD, MI-SVM, mi-SVM, DD-SVM, MILES, IS-MIL and MIForest. In order to obtain a full comparison, we also implement a generative multiple instance learning algorithm MC-EMDD for multiple concepts within the EM-DD framework as we mentioned in section 2. In MC-EMDD, y_{ij} is modeled by $P(+|B_{ij}) = \max_t P(+_t|B_{ij})$ where $+_t$ is the t^{th} disjunctive concept [9]. The results are given in Table 2, which show that MM-MIL achieves the best results among the key MIL methods on all three merged datasets. This is because all these merged data sets strongly reflect the multi-target problem, and MM-MIL can effectively model this underlying pattern with a data-dependent mixture. Although MC-EMDD also considers multi-modal concepts, the results of MM-MIL are substantially better. Our hypothesis is that MM-MIL benefits from both the smaller asymptotic error rate as a discriminative model and the data-dependent mixture modeling, while MC-EMDD is a generative model and can be shown to use data-independent mixtures. Different from previous methods, the proposed MM-MIL can not only label the regions (instances), but also tell which cluster a positive instance lies in by computing the *posterior* probability $P(z_m|B_{ij}, +)$. Figures 1 and 2 show several examples of positive instances selected from different bags. As illustrated, the new MM-MIL algorithm successfully localizes the target regions from each image and explicitly identifies the latent cluster the target belongs to.

Table 2. Average AUC for merged datasets and benchmark datasets by different algorithms

Algorithms	MergeData1	MergeData2	MergeData3	COREL	SIVAL
EM-DD [10]	0.543	0.643	0.661	0.564	0.687
MC-EMDD	0.602	0.694	0.718	0.616	0.691
MI-SVM [1]	0.536	0.628	0.652	0.535	0.698
mi-SVM [1]	0.542	0.614	0.674	0.557	0.683
DD-SVM [7]	0.568	0.671	0.704	0.675	0.762
MILES [8]	0.574	0.682	0.726	0.683	0.814
MIForest [25]	0.669	0.675	0.731	0.671	0.784
IS-MIL [3]	0.661	0.745	0.768	0.697	0.805
MM-MIL	**0.713**	**0.815**	**0.854**	**0.790**	**0.819**

4.2 Evaluation on Benchmark Datasets

The COREL dataset contains 2000 images from 20 different categories, with 100 images in each category and the SIVAL benchmark includes 25 different image categories with 60 images in each. COREL images contain various scenes and objects, eg, building, bus and elephant, where the target is typically close-ups and centered in the image. SIVAL consists of images of single objects photographed under different backgrounds, where objects may occur anywhere spatially in the image and also may be photographed at a wide-angle or close up. These two benchmarks were used extensively in the previous MIL researches [7,8,6,5,14]. The COREL dataset contains diversified positive instances while SIVAL dataset generally contains images with a single object in each category.

MM-MIL is compared with EM-DD, MI-SVM, mi-SVM, DD-SVM, MILES, IS-MIL and MIForest on these two benchmark datasets. M is chosen by cross-validation as in section 4.1. The average AUC results are reported in Table 2 and it shows that MM-MIL outperforms other methods on both COREL and SIVAL datasets. The AUC difference between MM-MIL and previous methods on SIVAL is relative small, whereas the difference on COREL is larger. The reason is that for one category, the targets from COREL images have very different features. For example, the 'Dinosaur' category consists of various kinds of dinosaurs. While in SIVAL dataset, each category contains one identical object with different backgrounds. Therefore, the AUC gap between MM-MIL and existing method is larger on COREL than that on SIVAL images. The superior performance of our method against existing discriminative MIL methods is mainly because: traditional MIL approaches are trying to learn one classifier for all instances/bags based on SVM framework, while our method first learn to separate instances into different clusters, and then a classifier is trained inside each cluster. Therefore, our MM-MIL method is more powerful in capturing the underlying patterns of the distribution of instances.

4.3 Experiments with Different Number of Hidden Mixtures

Figure 3 illustrates how the performance of MM-MIL varies with different values of M as the number of clusters. We plot the average AUC of MergeData1, MergeData2, MergeData3, COREL and SIVAL against the number of clusters from 1 to 10. When M equals 1, our proposed method degrades to a logistic regression model and has almost the same power as existing discriminative algorithms. With increases of M, up to a certain value, the performance saturates, which represents the true underlying pattern in the dataset. As illustrated in Figure 3, the saturated M in MergeData1, MergeData2 and MergeData3 is around 3 which capture the true clusters in these datasets. The AUC curve of COREL keeps increasing till M approaches 6, while the SIVAL curve is almost flat since there is a single target in SIVAL dataset from each category. It can be seen from Figure 3 that MM-MIL generates accurate results with a reasonably wide range of M values.

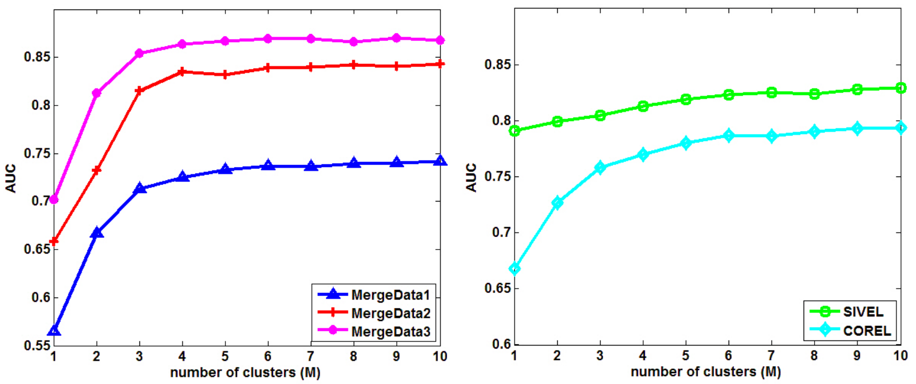

Fig. 3. AUC curves on different number of clusters for MergeData, COREL and SIVAL

5 Conclusions

Multiple instance learning is an important research topic with many applications such as image classification. Existing MIL methods do not explicitly address the multi-target problem where the distributions of positive instances are likely to be multi-modal in many practical applications. This paper presents a novel data-dependent mixture-model approach in the discriminative framework for multiple instance learning, which explicitly addresses the multi-target problem. Furthermore, a kernelized framework is proposed to allow efficient modeling within high dimensional feature space. Empirical results in image classification have shown that the new method outperforms several existing MIL algorithms on several datasets with multi-target positive instances and is consistently better than existing algorithms on benchmark datasets.

There are several possibilities to extend the research in this paper. For example, we plan to investigate different methods of combining instance labels to bag labels. We also plan to study the behavior of different types of kernels used in the classification. Furthermore, we plan to explore a non-parametric Bayesian method for modeling mixtures.

Acknowledgments. This work is partially supported by NSF research grants IIS-0746830, CNS-1012208 and IIS-1017837. This work also partially supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support Vector Machines for Multiple-Instance Learning. In: NIPS, pp. 561–568 (2002)
2. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the Multiple Instance Problem with Axis-Parallel Rectangles. Artif. Intell. 89(1-2), 31–71 (1997)
3. Fu, Z., Robles-Kelly, A.: An Instance Selection Approach to Multiple Instance Learning. In: CVPR (2009)
4. Hu, Y., Li, M., Yu, N.: Multiple-Instance Ranking: Learning to Rank Images for Image Retrieval. In: CVPR (2008)
5. Zha, Z., Hua, X., Mei, T., Wang, J., Qi, G., Wang, Z.: Joint Multi-Label Multi-Instance Learning for Image Classification. In: CVPR (2008)
6. Qi, G., Hua, X., Rui, Y., Mei, T., Tang, J., Zhang, H.: Concurrent Multiple Instance Learning for Image Categorization. In: CVPR (2007)
7. Chen, Y., Wang, J.Z.: Image Categorization by Learning and Reasoning with Regions. Journal of Machine Learning Research (5), 913–939 (2004)
8. Chen, Y., Bi, J., Wang, J.Z.: MILES: Multiple-Instance Learning via Embedded Instance Selection. IEEE Trans. Pattern Anal. Mach. Intell. 28(12), 1931–1947 (2006)
9. Maron, O., Lozano-Pérez, T.: A Framework for Multiple-Instance Learning. In: NIPS (1997)
10. Zhang, Q., Goldman, S.A.: EM-DD: An Improved Multiple-Instance Learning Technique. In: NIPS (2001)
11. Yang, C., Dong, M., Hua, J.: Region-based Image Annotation using Asymmetrical Support Vector Machine-based Multiple-Instance Learning. In: CVPR (2006)
12. Zhu, J., Hastie, T.: Kernel logistic regression and the import vector machine. Journal of Computational and Graphical Statistics 14(1), 185–205 (2005)
13. Viola, P.A., Platt, J.C., Zhang, C.: Multiple Instance Boosting for Object Detection. In: NIPS (2005)
14. Rahmani, R., Goldman, S.A.: MISSL: Multiple-Instance Semi-supervised Learning. In: ICML (2006)
15. Maron, O., Ratan, A.L.: Multiple-Instance Learning for Natural Scene Classification. In: ICML (1998)
16. Si, L., Jin, R.: Flexible Mixture Model for Collaborative Filtering. In: ICML, pp. 704–711 (2003)
17. Ray, S., Craven, M.: Supervised Versus Multiple Instance Learning: An Empirical Comparison. In: ICML (2005)

18. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Mathematical Programming 45(1-3), 503–528 (1989)
19. Lin, Z., Hua, G., Davis, L.S.: Multiple Instance Feature for Robust Part-based Object Detection. In: CVPR (2009)
20. Schölkopf, B., Herbrich, R., Smola, A.J.: A Generalized Representer Theorem. In: Helmbold, D.P., Williamson, B. (eds.) COLT/ EuroCOLT 2001. LNCS (LNAI), vol. 2111, pp. 416–426. Springer, Heidelberg (2001)
21. Ray, S., Page, D.: Multiple Instance Regression. In: ICML, pp. 425–432 (2001)
22. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer Science, Business Media, LLC (2006)
23. Babenko, B., Yang, M.-H., Belongie, S.J.: Visual tracking with online Multiple Instance Learning. In: CVPR (2009)
24. Wang, Q., Tao, L., Di, H.: A Globally Optimal Approach for 3D Elastic Motion Estimation from Stereo Sequences. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 525–538. Springer, Heidelberg (2010)
25. Leistner, C., Saffari, A., Bischof, H.: MIForests: Multiple-Instance Learning with Randomized Trees. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 29–42. Springer, Heidelberg (2010)
26. Xue, X., Zhang, W., Zhang, J., Wu, B., Fan, J., Lu, Y.: Correlative Multi-Label Multi-Instance Image Annotation. In: ICCV (2011)
27. Zeisl, B., Leistner, C., Saffari, A., Bischof, H.: On-line Semi-supervised mMultiple-instance Boosting. In: CVPR (2010)
28. Zhang, D., Liu, Y., Si, L., Zhang, J., Lawrence, R.D.: Multiple Instance Learning on Structred Data. In: NIPS (2011)
29. Zhang, D., Wang, F., Si, L., Li, T.: M3IC: Maximum Margin Multiple Instance Clustering. In: IJCAI (2009)
30. Zhang, D., Wang, F., Si, L., Li, T.: Maximum Margin Multiple Instance Clustering With Applications to Image and Text Clustering. IEEE Transactions on Neural Networks 22(5), 739–751 (2011)
31. Vezhnevets, A., Buhmann, J.M.: Towards Weakly Supervised Semantic Segmentation by Means of Multiple Instance and Multitask learning. In: CVPR (2010)
32. Zhang, D., Wang, F., Shi, Z., Zhang, C.: Interactive Localized Content-Based Image Retrieval with Multiple Instance Active Learning. Pattern Recognition 43(2), 478–484 (2010)