Learning Frame Relevance for Video Classification*

Hua Wang
Department of Computer
Science and Engineering
University of Texas at
Arlington, TX 76019, USA
huawangcs@gmail.com

Feiping Nie
Department of Computer
Science and Engineering
University of Texas at
Arlington, TX 76019, USA
feipingnie@gmail.com

Heng Huang
Department of Computer
Science and Engineering
University of Texas at
Arlington, TX 76019, USA
heng@uta.edu

Yi Yang
School of Computer Science
Carnegie Mellon University,
Pittsburgh, PA 15213, USA
yiyang@cs.cmu.edu

ABSTRACT

Traditional video classification methods typically require a large number of labeled training video frames to achieve satisfactory performance. However, in the real world, we usually only have sufficient labeled video clips (such as tagged online videos) but lack labeled video frames. In this paper, we formalize the video classification problem as a Multi-Instance Learning (MIL) problem, an emerging topic in machine learning in recent years, which only needs bag (video clip) labels. To solve the problem, we propose a novel Parameterized Class-to-Bag (P-C2B) Distance method to learn the relative importance of a training instance with respect to its labeled classes, such that the instance level labeling ambiguity in MIL is tackled and the frame relevances of training video data with respect to the semantic concepts of interest are given. Promising experimental results have demonstrated the effectiveness of the proposed method.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video Analysis*

General Terms

Algorithms, Performance, Experimentation

Keywords

Video Classification, Multi-Instance Learning

1. INTRODUCTION

Many existing video annotation approaches label every frame of a video clip and traditional video classification methods are usually designed to use the frame labels. However, frame level labeling requires expensive human labors, which often makes training an accurate video classification model very cost prohibitive. On the other hand, due to the rocketing growth of online videos with user tags at video clip level, we have abundant inexpensive coarsely labeled video training data. Therefore, devising video classification methods relying on only video clip labels is of great practical interest. In this paper, we explore this challenging, yet important, multimedia content analysis problem.

Given a video clip, the associated semantic labels usually arise from only a few of its frames but not all. For example, for the video clip in Figure 1, the "Studio" concept only comes from the two left frames, while the "Outdoor" concept is only attached to the three right frames. Apparently, labeling all the frames of this video clip is neither cost effective nor necessary. Therefore, identifying the relevances of the frames of coarsely labeled video clips to the semantic concepts of interest could potentially reduce the labeling cost while maintaining satisfactory video classification performance. To this end, we propose a novel Parameterized Class-to-Bag (P-C2B) Distance method by placing video classification under the framework of Multi-Instance Learning (MIL) that only leverages video clip labels, such that we are able to predict labels for unseen video clips as well as to learn frame relevances for the training video clips.

MIL [2] is an emerging topic in machine learning to address the classifications of data bags, in which each *bag* is a collection of *instances* with features associated to the instance. The aim of MIL is to infer bag labels based on the assumption that a positive bag contains at least one positive instance, while a negative bag contains negative instances only. MIL has attracted a lot of attention in recent years, and has been applied to many real-world applications [4, 9, 10, 11]. In the scenario of video classification, as illustrated in Figure 1, a video clip is considered as a bag and its frames are considered as instances. Our goal is to predict labels for a new coming bag (video clip) using the classification model learned from training bags and their associated labels.

*Area chair: Lexing Xie

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0616-4/11/11 ...$10.00.

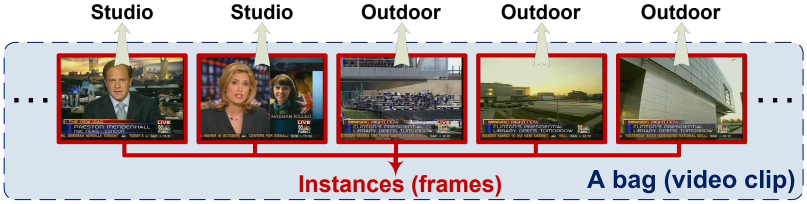

Figure 1: In multi-instance learning a video clip is represented as a bag and its frames are represented as instances. The semantic concepts associated with a video clip, *e.g.*, "Studio" and "Outdoor", usually arise only from a few of its frames but not all.

In the proposed method, we aim to address the two main challenges of MIL [9]: measuring set-to-set distance and weak label association. For the former, instead of computing the traditional Bag-to-Bag (B2B) distance that often does not truly reflect the semantic relationships between data objects [1], we consider to directly assess the relevance between classes and data objects, and propose to use *Class-to-Bag (C2B) distance*. For the latter, we take into account the relative importance of a training instance with respect to its labeled classes by assigning it with a weight for each of its labeled classes, called as *Significance Coefficient (SC)*. Ideally, the learned SCs of an instance, *i.e.*, frame relevance, with respect to its true associated classes should be large, whereas its SCs with respect other classes should be small. Parameterizing the C2B distance by the learned SCs, our P-C2B distance method for multi-instance data is proposed.

2. PARAMETERIZED CLASS-TO-BAG DISTANCE FOR MULTI-INSTANCE DATA

In this section, we first introduce a novel Parameterized Class-to-Bag (P-C2B) Distance to address the challenges of multi-instance data, followed by the objective and optimization procedures to learn it.

Problem formalization. Given a video classification task, we have N training video clips $\mathcal{X} = \{X_1, \ldots, X_N\}$ and K conceptual classes. Each video clip contains a number of frames represented by a bag of instances $X_i = \left[\mathbf{x}_i^1, \ldots, \mathbf{x}_i^{n_i}\right] \in \mathbb{R}^{d \times n_i}$, where n_i is the number of the frames (instances) in video clip X_i. Each instance is abstracted as a vector $\mathbf{x}_i^j \in \mathbb{R}^d$ of d dimensions. We are also given the class memberships of the input data, denoted as $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T \in \{0,1\}^{N \times K}$ where $\mathbf{y}_i$ is the label indicator of X_i. In the setting of MIL, if there exists $j \in \{1, \ldots, n_i\}$ such that $\mathbf{x}_i^j$ belongs to the k-th class, X_i is assigned to the k-th class and $Y_{ik} = 1$, otherwise $Y_{ik} = 0$. Yet the concrete value of the index j is unknown. More specifically, the following assumptions are held in the settings of MIL: (1) bag X is assigned to the k-th class $\Longleftrightarrow$ at least one instance of X belongs to the k-th class; (2) bag X is not assigned to the k-th class $\Longleftrightarrow$ no instance in X belongs to the k-th class. Our goal is to learn from the training data $\mathcal{D} = \{X_i, \mathbf{y}_i\}_{i=1}^N$ a classifier that is able to predict labels for a new query video clip X.

2.1 Parameterized Class-to-Bag Distance

To tackle the two major difficulties of MIL, the estimation of set-to-set distances and instance level labeling ambiguity [9], we first propose a novel P-C2B distance for multi-instance data.

Class-to-Bag (C2B) Distance. We first represent every class as a *super-bag* that comprises the instances of all its training bags: $C_k = \left\{\mathbf{x}_i^j \mid i \in \pi_k\right\}$ where $\pi_k = \{i \mid Y_{ik} = 1\}$ is the index set of all the training bags belonging to the k-th class. We denote the number of instances in C_k as m_k, *i.e.*, $|C_k| = m_k$. Note that, in single-label video data sets where each video clip belongs to exactly one class, *i.e.*, $\sum_{k=1}^K Y_{ik} = 1$, therefore $C_k \cap C_l = \varnothing \ (\forall\, k \neq l)$ and $\sum_{k=1}^K m_k = \sum_{i=1}^N n_i$. In multi-label video data sets where each video clip (thereby each instance) may belong to more than one class, *i.e.*, $\sum_{k=1}^K Y_{ik} \geq 1$, thus $C_k \cap C_l \neq \varnothing \ (\forall\, k \neq l)$ and $\sum_{k=1}^K m_k \geq \sum_{i=1}^N n_i$, *i.e.*, different super-bags may overlap and one instance $\mathbf{x}_i^j$ may appear in multiple super-bags.

Then we define the elementary distance from an instance $\mathbf{x}_i^j$ of a super-bag C_k to a data bag $X_{i'}$ by the distance between $\mathbf{x}_i^j$ and its nearest neighbor instance in $X_{i'}$:

$$dist_k\left(\mathbf{x}_i^j, X_{i'}\right) = \left\|\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right\| \quad , \quad \forall i \in \pi_k \quad , \tag{1}$$

where $\mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)$ denotes the nearest neighbor of $\mathbf{x}_i^j$ in $X_{i'}$.

Finally, the C2B distance from C_k to $X_{i'}$ is computed as:

$$Dist\left(C_k, X_{i'}\right) = \sum_{i \in \pi_k} \sum_{j=1}^{n_i} dist_k\left(\mathbf{x}_i^j, X_{i'}\right) = \sum_{i \in \pi_k} \mathbf{e}^T \mathbf{d}_{ii'k} \quad , \tag{2}$$

where $\mathbf{e} = [1, \ldots, 1]^T$ is a constant vector, and $\mathbf{d}_{ii'k} \in \mathbb{R}^{n_i}$ in which the j-th element is $\left\|\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right\|$.

Parameterized Class-to-Bag (P-C2B) Distance. Because the C2B distance defined in Eq. (2) does not take into account the the instance level labeling ambiguity in MIL, we further develop it by weighting the instances in a super-bag upon their relevance to a concerned class.

Due to the ambiguous associations between instances and labels, not all the instances in a super-bag really characterize the corresponding class. For example, in Figure 1 the rightmost instance (frame) is in the super-bag of "Studio" class, because the entire video clip is labeled with both "Studio" and "Outdoor". Intuitively, we should give it a smaller, or even no, weight when determining whether to assign "Studio" label to a query video clip; and give it a higher weight when deciding "Outdoor" label. To be more precise, let w_{ik}^j be the weight for $\mathbf{x}_i^j$ with respect to the k-th class, we compute the parameterized C2B distance from C_k to $X_{i'}$ as:

$$Dist\left(C_k, X_{i'}\right) = \sum_{i \in \pi_k} \mathbf{w}_{ik}^T \mathbf{d}_{ii'k} \quad , \tag{3}$$

where $\mathbf{w}_{ik} = \left[w_{ik}^1, \ldots, w_{ik}^{n_i}\right]^T$. As can be seen, different from Eq. (2) in which all the elementary distances are equally weighted by $\mathbf{e}$, the C2B distance defined in Eq. (3) gives different weights to the instances in a super-bag upon their relevances by $\mathbf{w}_{ik}$. Because w_{ik}^{j} reflects the relative importance of instance $\mathbf{x}_i^j$ when determining the label for the k-th class, we call it as the Significance Coefficient (SC) of $\mathbf{x}_i^j$ with respect to the k-th class, and the resulted C2B distance computed by Eq. (3) as the proposed Parameterized Class-to-Bag (P-C2B) Distance.

2.2 Objective and Optimization Algorithm

Armed with the P-C2B distance defined in Eq. (3), we learn w_{ik}^{j} by maximizing the data separability, *i.e.*, we minimize the overall P-C2B distance from a class to all its belonging bags, whilst maximizing the overall P-C2B distance from the same class to all the bags not belonging to it. Formally, for a given class C_k, we solve the following optimization problem:

$$\min_{\mathbf{w}_{ik}\geq 0,\ \mathbf{w}_{ik}^T\mathbf{e}=1} \frac{\sum_{i'\in\pi_k}\sum_{i\in\pi_k}\mathbf{w}_{ik}^T\mathbf{d}_{ii'k}}{\sum_{i'\notin\pi_k}\sum_{i\in\pi_k}\mathbf{w}_{ik}^T\mathbf{d}_{ii'k}} \ . \tag{4}$$

Let $\mathbf{d}_{ik}^{w}=\sum_{i'\in\pi_k}\mathbf{d}_{ii'k}\in\mathbb{R}^{n_i}$ and $\mathbf{d}_{ik}^{b}=\sum_{i'\notin\pi_k}\mathbf{d}_{ii'k}\in\mathbb{R}^{n_i}$, we rewrite the problem in Eq. (4) as:

$$\min_{\mathbf{w}_{ik}\geq 0,\ \mathbf{w}_{ik}^T\mathbf{e}=1} \frac{\sum_{i\in\pi_k}\left(\mathbf{w}_{ik}^T\mathbf{d}_{ik}^{w}+\alpha\mathbf{w}_{ik}^T\mathbf{w}_{ik}\right)}{\sum_{i\in\pi_k}\mathbf{w}_{ik}^T\mathbf{d}_{ik}^{b}} \ , \tag{5}$$

where the second term in the numerator is added to avoid the degenerate solution.

To solve the optimization objective in Eq. (5), we derive an iterative algorithm as following.

Algorithm 1: The algorithm to solve the problem (5).

```
$t=0$. Randomly initialize $\mathbf{w}_{ik}^{(0)}$ satisfying $\mathbf{w}_{ik}^{(0)}\geq 0$, $\left(\mathbf{w}_{ik}^{(0)}\right)^T\mathbf{e}=1$.
while not converge do
    1. Calculate $\lambda^{(t)}=\frac{\sum_{i\in\pi_k}\left[\left(\mathbf{w}_{ik}^{(t)}\right)^T\mathbf{d}_{ik}^{w}+\alpha\left(\mathbf{w}_{ik}^{(t)}\right)^T\mathbf{w}_{ik}^{(t)}\right]}{\sum_{i\in\pi_k}\left(\mathbf{w}_{ik}^{(t)}\right)^T\mathbf{d}_{ik}^{b}}$.
    2. Calculate
       $\mathbf{w}_{ik}^{(t+1)}=\arg\min_{\mathbf{w}_{ik}\geq 0,\ \mathbf{w}_{ik}^T\mathbf{e}=1}\sum_{i\in\pi_k}\left(\mathbf{w}_{ik}^T\mathbf{d}_{ik}^{w}+\alpha\mathbf{w}_{ik}^T\mathbf{w}_{ik}\right)-\lambda^{(t)}\sum_{i\in\pi_k}\mathbf{w}_{ik}^T\mathbf{d}_{ik}^{b}$.   (6)
    3. $t=t+1$.
end
```

Let $\mathbf{d}_{ik}=\mathbf{d}_{ik}^{w}-\lambda^{(t)}\mathbf{d}_{ik}^{b}$, we may rewrite the optimization problem in Eq. (6) as following:

$$\min_{\mathbf{w}_{ik}\geq 0,\ \mathbf{w}_{ik}^T\mathbf{e}=1} \sum_{i\in\pi_k}\left(\mathbf{w}_{ik}^T\mathbf{d}_{ik}+\alpha\mathbf{w}_{ik}^T\mathbf{w}_{ik}\right) \ . \tag{7}$$

We can see that the problem in Eq. (7) can be decoupled to solve the following subproblems separately for each $i\in\pi_k$:

$$\min_{\mathbf{w}_{ik}\geq 0,\ \mathbf{w}_{ik}^T\mathbf{e}=1} \mathbf{w}_{ik}^T\mathbf{d}_{ik}+\alpha\mathbf{w}_{ik}^T\mathbf{w}_{ik} \ , \tag{8}$$

which are convex quadratic programming (QP) problems, and can be efficiently solved because $\mathbf{w}_{ik}\in\mathbb{R}^{n_i}$ and the value of n_i, *i.e.*, the bag size of X_i, is usually not large.

Given the learned w_{ik}^{j} $(1\leq k\leq K, 1\leq i\leq N, 1\leq j\leq n_i)$, we can compute $D\left(C_k, X\right)$ $(1\leq k\leq K)$ for a query video clip X using Eq. (3), based on which the classification can be conducted following the same rules as in [9].

3. EXPERIMENTAL RESULTS

In this section, we experimentally evaluate the proposed P-C2B method in automatic video classification tasks, where we emphasize its effectiveness in low-cost conditions.

3.1 Data Preparation

We conduct our experiments using **TRECVID 2005** video data set[1], which contains 277 video clips with 61,901 shots labeled by 39 LSCOM-Lite concepts. Each shot (key frame) is considered as an instance in our study. For each instance, following [8] we extract a 384-dimensional low-level visual feature vector by dividing the corresponding key frame into 64 blocks by a 8×8 grid and computing the first and second moments (mean and variance) of each color band. We split each video clip into 5 consecutive parts and end up with 1385 bags. Different bags have different numbers of instances. In average, each bag comprises 44.7 instances. Because in TRECVID 2005 data set, each video clip is annotated with more than one semantic label, it is a multi-label data set.

3.2 Experimental Settings

Because the main purpose of the proposed P-C2B distance method is to deal with low-cost video classification, we evaluate it in the conditions where we only have video clips labels but not frame labels. We employ standard 5-fold cross-validation for evaluation and report the average performance over the 5 trials. For each video clip, we randomly select a fraction of its frames and assign their labels to the video clip, while the labels of both selected and not-selected frames are assumed to be unknown. We emulate two different conditions when the amount of selected frames are 80% and 50%, and the corresponding results are reported in the top and bottom halves of Table 1 respectively.

The proposed P-C2B method has only one parameter, *i.e.*, the regularization parameter α in Eq. (5). Empirically, we set $\alpha=0.01$ throughout our experiments.

We first compare our method to two baseline classification methods including support vector machine (SVM) method and transductive support vector machine (TSVM) [3] method. The former is one of the most widely used supervised classification method in statistical learning, while the latter is an extension of the former and is a semi-supervised classification method. Because both of these two methods are designed for single-instance data, they are not able to handle data objects with varied sizes. To this end, we assign the bag labels to all its component frames. This brings instance level labeling ambiguity, which, on the other hand, significantly reduces required labeling cost for training, because we only need human experts to label either the whole video clips or a fraction of their frames instead of all. For each class we train a one-vs-others classier using the frames in the training video clips, and classify the frames in the test video clips. Gaussian kernel is used for the both methods, *i.e.*, $\mathcal{K}\left(\mathbf{x}_i,\mathbf{x}_j\right)=\exp\left(-\beta\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2\right)$, where β and the regularization parameter C are fine tuned by searching the grid of $\left\{10^{-5},\ldots,10^{-1},1,10,\ldots,10^{5}\right\}$ via an internal 5-fold cross-validation using the training data of each of the 5 trails. The both methods are implemented using SVM-light[2].

We also compare our method to two most related methods, *i.e.*, two recent MIL methods including miGraph [11]

[1] `http://www-nlpir.nist.gov/projects/trecvid/`

[2] `http://svmlight.joachims.org/`

Table 1: Video classification performances (mean ± standard deviation) comparison when only 80% (top half) and 50% (bottom half) of the frames are labeled.

Method	Hamming loss ↓	One-error ↓	Coverage ↓	Rank loss ↓	Average precision ↑
SVM	0.239 ± 0.014	0.448 ± 0.010	1.323 ± 0.010	0.240 ± 0.012	0.335 ± 0.020
TSVM	0.235 ± 0.012	0.444 ± 0.012	1.312 ± 0.016	0.233 ± 0.014	0.338 ± 0.022
SML	0.228 ± 0.013	0.441 ± 0.015	1.308 ± 0.011	0.229 ± 0.016	0.342 ± 0.025
miGraph	0.215 ± 0.012	0.366 ± 0.016	1.202 ± 0.015	0.213 ± 0.010	0.386 ± 0.021
MIMLSVM+	0.208 ± 0.010	0.357 ± 0.015	1.116 ± 0.018	0.205 ± 0.012	0.398 ± 0.021
P-C2B	**0.191 ± 0.011**	**0.341 ± 0.010**	**1.093 ± 0.007**	**0.188 ± 0.010**	**0.420 ± 0.012**
SVM	0.282 ± 0.012	0.515 ± 0.013	1.585 ± 0.019	0.289 ± 0.017	0.240 ± 0.019
TSVM	0.280 ± 0.013	0.511 ± 0.015	1.580 ± 0.020	0.283 ± 0.015	0.245 ± 0.020
SML	0.276 ± 0.010	0.505 ± 0.013	1.574 ± 0.015	0.277 ± 0.013	0.249 ± 0.018
miGraph	0.233 ± 0.012	0.418 ± 0.015	1.310 ± 0.019	0.244 ± 0.012	0.306 ± 0.020
MIMLSVM+	0.227 ± 0.014	0.404 ± 0.015	1.293 ± 0.016	0.237 ± 0.013	0.318 ± 0.018
P-C2B	**0.209 ± 0.011**	**0.373 ± 0.012**	**1.183 ± 0.014**	**0.214 ± 0.011**	**0.354 ± 0.013**

method and MIMLSVM+ [4] method. Because miGraph method is a single-label classification method, one-vs-others strategy is used to conduct classification, one class at a time. Note that, because both of these two method and the proposed P-C2B method are multi-instance classification methods, thus we perform classification at bag, *i.e.*, video clip, level. Namely, although we know the ground truth instance labels, these three methods do not use them. They only use bag labels following standard MIL settings.

Finally, we report the performance of a most recent video classification method, *i.e.*, supervised manifold learning (SML) [5] method which has demonstrated state-of-the-art classification performance. Since this method is designed to work at frame level, we employ the same strategy as that for SVM to conduct classification. We implement the method following its original work and set the parameters as optimal.

3.3 Experimental Results

Because TRECVID 2005 data set is a multi-label data set, we evaluate the classification performances of the compared methods using five widely used multi-label evaluation metrics, as shown in Table 1, where "↓" indicates "the smaller the better" while "↑" indicates "the bigger the better". We refer readers to [6] for details of these evaluation metrics.

The average classification performances (mean ± standard deviation) of the compared methods over the 5 trials of the experiments are reported in Table 1, from which we can see that the proposed method is consistently better than the other compared methods, sometimes very significantly. Moreover, when labeling cost is reduced, *i.e.*, the amount of labeled frames are reduced, the classification performance degradations of the multi-instance methods, including the proposed P-C2B distance method, are not very significant, whereas those of the single-instance methods are considerably large. These results concretely demonstrate the usefulness of multi-instance learning in cost effective video classification, as well as the effectiveness of the proposed method.

4. CONCLUSION

In this paper, we proposed a novel Parameterized Class-to-Bag (P-C2B) Distance method to solve the video classification problem under the framework of Multi-Instance Learning (MIL). The Significance Coefficients (SCs) are learned to assess the relative importance of a training instance with respect to its labeled classes, which thus solves the notorious instance label assignment ambiguity in MIL. Moreover, through the learned SCs, the frame relevances to the concerned semantic concepts of the coarsely labeled video clips are explicitly given to reveal the insight of the input data set. The promising experimental results were reported in empirical evaluations, which validated the proposed method.

Acknowledgment.

This research was supported by NSF-IIS 1117965, NSFCCF-0830780, NSF-DMS-0915228, NSF-CCF-0917274.

5. REFERENCES

[1] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *IEEE CVPR*, 2008.

[2] T. Dietterich, R. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[3] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209.

[4] Y. Li, S. Ji, S. Kumar, J. Ye, and Z. Zhou. Drosophila Gene Expression Pattern Annotation through Multi-Instance Multi-Label Learning. *ACM/IEEE TCBB*, 2011.

[5] Y. Liu, Y. Liu, and K. Chan. Supervised manifold learning for image and video classification. In *ACM Multimedia*, 2010.

[6] R. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine learning*, 39(2):135–168, 2000.

[7] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Jouranl of Royal Statistics Society B.*, 58:267–288, 1996.

[8] H. Wang, H. Huang, and C. Ding. Image annotation using multi-label correlated green's function. In *IEEE ICCV*, 2009.

[9] H. Wang, F. Nie, and H. Huang. Learning instance specific distance for multi-instance classification. In *AAAI*, 2011.

[10] C. Zhang and P. Viola. Multiple-instance pruning for learning efficient cascade detectors. In *NIPS*, 2008.

[11] Z. Zhou, Y. Sun, and Y. Li. Multi-instance learning by treating instances as non-I.I.D. samples. In *ICML*, 2009.