Robust and Discriminative Distance for Multi-Instance Learning

Hua Wang, Feiping Nie and Heng Huang
Department of Computer Science and Engineering
University of Texas at Arlington, Arlington, Texas 76019, USA
huawangcs@gmail.com, feipingnie@gmail.com, heng@uta.edu

Abstract

Multi-Instance Learning (MIL) is an emerging topic in machine learning, which has broad applications in computer vision. For example, by considering video classification as a MIL problem where we only need labeled video clips (such as tagged online videos) but not labeled video frames, one can lower down the labeling cost, which is typically very expensive. We propose a novel class specific distance Metrics enhanced Class-to-Bag distance (M-C2B) method to learn a robust and discriminative distance for multi-instance data, which employs the not-squared ℓ_2-norm distance to address the most difficult challenge in MIL, i.e.*, the outlier instances that abound in multi-instance data by nature. As a result, the formulated objective ends up to be a simultaneous $\ell_{2,1}$-norm minimization and maximization (minmax) problem, which is very hard to solve in general due to the non-smoothness of the $\ell_{2,1}$-norm. We thus present an efficient iterative algorithm to solve the general $\ell_{2,1}$-norm minmax problem with rigorously proved convergence. To the best of our knowledge, we are the first to solve a general $\ell_{2,1}$-norm minmax problem in literature. We have conducted extensive experiments to evaluate various aspects of the proposed method, in which promising results validate our new method in cost-effective video classification.*

1. Introduction

Multi-Instance Learning (MIL) [2, 16, 15] is an emerging topic in machine learning to address the classifications of data bags. In MIL, each *bag* is a collection of *instances* with features associated to the instance. The aim of MIL is to infer the bag level labels based on the assumption that a positive bag contains at least one positive instance, whereas a negative bag contains negative instances only. For example, placing video classification under the framework of MIL [17], we consider a video clip as a bag and its frames instances, where 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.

Although multi-instance representation has a number of advantages, such as capturing the inherent data structures [16, 15] and reducing the labeling cost [17], it also brings new challenges for statistical learning. First, because in MIL a data object is represented as a bag of instances, the distance between data objects turns out to be a set-to-set distance. Compared to single-instance data that use the vector distance such as Euclidian distance, the distance estimation in MIL is more complicated. Second, in MIL labels are assigned to bags, but not instances. Thus, the instance level labels by nature are ambiguous. As a result, although a bag belongs to a class, some, or even most, of its instances may not be truly related to the same class, which are considered as outlier instances with respect to the concerned class and by design largely exist in a multi-instance data set.

In this paper we explore the challenges in MIL, as well as the opportunities, to improve the multi-instance classification. Following the same idea in our earlier works [16, 15, 17], we use the *Class-to-Bag (C2B)* distance to deal with multi-instance data. Specifically, we consider each class as a "super-bag", which comprises all the instances in the data bags belonging to the concerned class. The elementary distance from an instance in a super-bag to a data bag is first estimated, then the C2B distance from the class to the data object is computed as the sum of the elementary distances from all the instances in the super-bag to the data bag of interest. The main difference between this study and our earlier works, as well as most, if not all, existing MIL methods, lies in that we use *not-squared* ℓ_2-norm distance to design the multi-instance distance, which is of particular use to address the abundant outlier instances due to the labeling ambiguity. Although it is generally accepted that [9] the not-squared ℓ_2-norm distance is more robust against noises and outlier samples, it is more challenging to use it because of its non-smoothness. Specifically, because we aim to learn distance metrics to improve the data discriminativity, our learning objective ends up to be a simultaneous $\ell_{2,1}$-norm minimization and maximization (minmax) problem, which is obviously much harder to solve than the widely studied

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

$\ell_{2,1}$-norm minimization problem in literature. As an important theoretical contribution of this work, we present an efficient iterative algorithm to solve the general $\ell_{2,1}$-norm minmax problem, as well as our objective, and rigorously prove its convergence.

We summarize our contributions as following.

- To address the major difficulty of MIL, *i.e.*, the huge amount of outlier instances due to the labeling ambiguity, we propose a novel class specific distance Metrics enhanced Class-to-Bag distance (M-C2B), which is defined by not-squared ℓ_2-norm distances thereby is more robust against the outlier instances.
- To improve the data discriminativity, we formulate an objective that simultaneously minimizes and maximizes $\ell_{2,1}$-norm terms, which is hard to solve in general due to the non-smoothness of the component ℓ_2-norms (without square). Thus we present an efficient iterative algorithm to solve the proposed objective with rigorously proved convergence. To the best of our knowledge, our work is the *first* attempt to solve a general $\ell_{2,1}$-norm minmax problem.
- We apply MIL in video classification, by which we only need video clip (bag) level labels but not frame (instance) level labels. Compared to existing video classification models that rely on frame level labels, the training cost of our model is much lower.
- Promising results in comprehensive experiments to evaluate a variety of aspects of the proposed method validate its effectiveness.

2. MIL via M-C2B distance

In this section we first describe the M-C2B distance for multi-instance data, where we introduce the class specific distance metrics. Different to existing MIL studies, our new distance is not squared, which thereby is robust against outlier instances that by nature abound in multi-instance data. Then we develop our optimization objective to learn the distance metrics, followed by an efficient iterative algorithm to solve the objective with rigorously proved convergence.

Problem formalization of MIL. 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 frames (instances) in the video clip. 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 class membership indication 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 MIL settings: (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. C2B distance for multi-instance data

We first introduce a robust C2B distance, by which we tackle the main difficulties in MIL, *i.e.*, the estimation of set-to-set distances and instance level label ambiguity.

Definition of the C2B distance. First, we 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 where each video clip belongs to one and only one class, *i.e.*, $\sum_{k=1}^K Y_{ik} = 1$, we have $C_k \cap C_l = \varnothing \ \ (\forall\, k \neq l)$ and $\sum_{k=1}^K m_k = \sum_{i=1}^N n_i$. In multi-label video data, each video clip (thereby each instance) may belong to more than one class, *i.e.*, $\sum_{k=1}^K Y_{ik} \geq 1$, therefore $C_k \cap C_l \neq \varnothing \ \ (\forall\, k \neq l)$ 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 $X_{i'}$ using the distance between $\mathbf{x}_i^j$ and its nearest neighbor instance in $X_{i'}$ as:

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

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

Finally we compute the C2B distance from C_k to $X_{i'}$ as:

$$\begin{aligned} D\left(C_k, X_{i'}\right) &= \sum_{i \in \pi_k} \sum_{j=1}^{n_i} d_k\left(\mathbf{x}_i^j, X_{i'}\right) \\ &= \sum_{i \in \pi_k} \sum_{j=1}^{n_i} \left\|\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right\| \ . \end{aligned} \tag{2}$$

Note that, a crucial difference between the multi-instance distances defined in Eqs. (1–2) and that in related works in [4, 3] as well as our earlier works in [16, 15] lies in that they are ***not*** squared distance. It is well known in statistical learning and machine vision community [9] that the squared distance can be easily dominated and biased by outlier elements, such that it cannot reflect the true relationships between the objects. However, as discussed before, outlier instances by nature are abundant in multi-instance data. As a result, the robustness of a classification model against these outlier instances is of primary significance in multi-instance

learning. Therefore, although the squared distance could simply the mathematical formulations of subsequent learning processes as in traditional metric learning [18, 7], as the first contribution of this work, we propose to use *not-squared* distance defined in Eqs. (1–2) to measure the relationships for multi-instance data.

Class specific distance metrics enhanced C2B (M-C2B) distance. The C2B distance defined in Eq. (2) essentially is a Euclidean distance, which thereby is independent of input data. Similar to many other computer vision applications, using Mahalanobis distance with an appropriate distance metric to capture the second-order statistics of input data is also desirable for video classification. Taking into account the high heterogeneity of video data, instead of learning one single global distance metric as in existing works [4, 3], following our earlier work [15] we propose to learn K different class specific distance metrics $\{M_k \succ 0\}_{k=1}^{K} \subset \mathbb{R}^{d\times d}$, one for each class. Note that, using class specific distance metrics is only feasible with the distance defined between classes and bags such as the one used in this work, because we are only concerned with intra-class distance. However, traditional B2B distance needs to compute distances between bags belonging to different classes that involve inter-class distance metrics, which inevitably complicates the problem.

Specifically, instead of using Eq. (2), we compute the C2B distance using the Mahalanobis distance as follows:

$$D\left(C_k, X_{i'}\right) = \sum_{i\in\pi_k}\sum_{j=1}^{n_i}\sqrt{\left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right]^T M_k \left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right]} \quad . \tag{3}$$

We refer to $D\left(C_k, X_{i'}\right)$ computed in Eq. (3) as the proposed class specific distance Metrics enhanced C2B (M-C2B) distance in the sequel of this paper.

2.2. Objective to learn class specific distance metrics

Armed with the M-C2B distance defined in Eq. (3) for a multi-instance data set, following standard learning strategy, we learn the class specific distance metrics $M_k\,(1 \le k \le K)$ by maximizing the data separability. To be more precise, we minimize the overall M-C2B distance from a class to all its belonging bags, whilst maximizing the overall M-C2B distance from the same class to all the bags not belonging to it. Formally, for a given class, say C_k, we solve the following optimization problem:

$$\min_{M_k\succ 0} \frac{\sum_{i'\in\pi_k} D\left(C_k, X_{i'}\right)}{\sum_{i'\notin\pi_k} D\left(C_k, X_{i'}\right)} \quad . \tag{4}$$

A potential problem that impedes us to directly solve the problem in Eq. (4) is its huge variable space, with a total of $d \times d$ variables for the k-th class. Thus, following many prior metric learning works [7], because M_k is positive definite, we can reasonably write it as $M_k = U_k U_k^T$ where $U_k \in \mathbb{R}^{d\times r}$. Thus, the optimization problem in Eq. (4) is transformed as the following problem:

$$\min_{U_k} \frac{\sum_{i'\in\pi_k}\sum_{i\in\pi_k}\sum_{j=1}^{n_i}\left\|\left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right]^T U_k\right\|}{\sum_{i'\notin\pi_k}\sum_{i\in\pi_k}\sum_{j=1}^{n_i}\left\|\left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right]^T U_k\right\|} \quad . \tag{5}$$

Because in practice we often select $r \ll d$, the variable space of the problem is considerably compressed.

Despite the compact formulation, solving Eq. (5) is very challenging, because it simultaneously minimizes and maximizes a number of non-smooth terms of ℓ_2-norm (without square). In the next subsection, we will present an efficient algorithm to seek the optimal solution of Eq. (5).

Upon the solved distance metrics $M_k \;\; (1 \le k \le K)$, given a query video clip X, we can compute $D\left(C_k, X\right) \;\; (1 \le k \le K)$ by Eq. (3). Sorting $D\left(C_k, X\right)$, we can assign labels to the query video clip following the same classification rules as introduced in [16, 15].

2.3. Algorithm and its analysis

First, for the k-th class, we rewrite Eq. (4) in an equivalent yet more compact form as followings:

$$\min_{U^TU=I} \frac{\|AU\|_{2,1}}{\|BU\|_{2,1}} = \frac{\sum_i \left\|\mathbf{a}^i U\right\|}{\sum_i \left\|\mathbf{b}^i U\right\|}, \tag{6}$$

where $\|M\|_{2,1}$ denotes the $\ell_{2,1}$-norm of the matrix M, which is a non-smooth norm and broadly used the studies of multi-task learning [1]. For a general matrix $M = [m_{ij}]$, we denote its rows as $\mathbf{m}^i$, thus the $\ell_{2,1}$-norm of the matrix M is defined as $\|M\|_{2,1} = \sum_i \sqrt{\sum_j m_{ij}^2} = \sum_i \left\|\mathbf{m}^i\right\|_2$ [1, 9]. Here, we call A as the within-class distance matrix for the k-th class, whose row $\mathbf{a}^i$ is one $\left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right]^T$ that satisfies $i' \in \pi_k$, $i \in \pi_k$, and $1 \le j \le n_i$. Similarly, we call B as the between-class distance matrix for the k-th class, whose row is one $\left[\mathbf{x}_i^j - \mathcal{N}_{i'}\left(\mathbf{x}_i^j\right)\right]^T$ that satisfies $i' \notin \pi_k$, $i \in \pi_k$, and $1 \le j \le n_i$. Without ambiguity, we drop the subscript k for notation brevity.

Although there exist in literature [1, 9] a plethora of algorithms that minimize the objectives involving $\ell_{2,1}$-norm terms, to the best of our knowledge, no algorithm exists to solve the objectives that simultaneously minimize and maximize (minmax) $\ell_{2,1}$-norm terms, which is definitely much harder than $\ell_{2,1}$-norm minimization problems. As an important theoretical contribution of this work, we present the following algorithm. to solve the general $\ell_{2,1}$-norm minmax problem in Eq. (6).

Convergence analysis of the algorithm. The following theorem guarantees the convergence of Algorithm 1.

Theorem 1 *Algorithm 1 decreases the objective value of the problem of Eq.* (6) *in each iteration till converges.*

Algorithm 1: An efficient iterative algorithm to solve the general $\ell_{2,1}$-norm minmax problem in Eq. (6).

Input: Within-class distance matrix A and between-class distance matrix B.

begin

1. Initialize U by random guess.

while *not converge* **do**

2. Compute $\lambda = \frac{\sum_i \|\mathbf{a}^i U\|}{\sum_i \|\mathbf{b}^i U\|}$.

3. Compute $d_{ii} = \frac{1}{2\|\mathbf{a}^i U\|}$ and construct the diagonal matrix D with its diagonal entries as d_{ii}.

4. Compute $\mathbf{s}^i = \frac{\mathbf{b}^i U}{\|\mathbf{b}^i U\|}$ and construct the matrix S with its rows as $\mathbf{s}^i$.

5. Compute $U = \lambda \left(2A^T DA\right)^{-1} B^T S$.

Output: The learned distance metric matrix $M = UU^T$ for the k-th class.

Proof. First, it is obvious that step 5 in Algorithm 1 computes the optimal solution of the following problem:

$$\min_U \ \mathbf{tr}\left(U^T A^T DAU\right) - \lambda \mathbf{tr}\left(U^T B^T S\right) \ . \tag{7}$$

For an iteration, we denote the updated U by Algorithm 1 by $\widetilde{U}$. Then according to step 5 and Eq. (7), we have:

$$\begin{aligned} &\mathbf{tr}\left(\widetilde{U} A^T DA\widetilde{U}\right) - \lambda \mathbf{tr}\left(\widetilde{U}^T B^T S\right) \leq \\ &\mathbf{tr}\left(UA^T DAU\right) - \lambda \mathbf{tr}\left(U^T B^T S\right) \ , \end{aligned} \tag{8}$$

from which by the definitions in step 3 and step 4 we have:

$$\begin{aligned} &\sum_i \frac{\left\|\mathbf{a}^i \widetilde{U}\right\|^2}{2\|\mathbf{a}^i U\|} - \lambda \sum_i \frac{\mathbf{b}^i \widetilde{U} U^T \left(\mathbf{b}^i\right)^T}{\|\mathbf{b}^i U\|} \leq \\ &\sum_i \frac{\|\mathbf{a}^i U\|^2}{2\|\mathbf{a}^i U\|} - \lambda \sum_i \frac{\mathbf{b}^i U U^T \left(\mathbf{b}^i\right)^T}{\|\mathbf{b}^i U\|} \ . \end{aligned} \tag{9}$$

Because it can verified that for function $g(x) = x - \frac{x^2}{2\alpha}$, given any $x \neq \alpha \in \mathbb{R}^n$, $g(x) \leq g(\alpha)$ holds, we have:

$$\left\|\mathbf{a}^i \widetilde{U}\right\| - \frac{\left\|\mathbf{a}^i \widetilde{U}\right\|^2}{2\|\mathbf{a}^i U\|} \leq \|\mathbf{a}^i U\| - \frac{\|\mathbf{a}^i U\|^2}{2\|\mathbf{a}^i U\|} \ . \tag{10}$$

According to Cauchy-Schwarz inequality, we can derive:

$$\begin{aligned} &\left\|\mathbf{b}^i \widetilde{U}\right\| \|\mathbf{b}^i U\| \geq \mathbf{b}^i \widetilde{U} U^T \left(\mathbf{b}^i\right)^T \Longrightarrow \\ &\frac{\mathbf{b}^i \widetilde{U} U^T \left(\mathbf{b}^i\right)^T}{\|\mathbf{b}^i U\|} - \left\|\mathbf{b}^i \widetilde{U}\right\| \leq 0 = \frac{\mathbf{b}^i U U^T \left(\mathbf{b}^i\right)^T}{\|\mathbf{b}^i U\|} - \|\mathbf{b}^i U\| \ . \end{aligned} \tag{11}$$

Then by adding the three inequalities in Eqs. (9–11) in the both sides, we obtain the following inequality:

$$\sum_i \left\|\mathbf{a}^i \widetilde{U}\right\| - \lambda \sum_i \left\|\mathbf{b}^i \widetilde{U}\right\| \leq \sum_i \|\mathbf{a}^i U\| - \lambda \sum_i \|\mathbf{b}^i U\| = 0$$

$$\Longrightarrow \frac{\sum_i \left\|\mathbf{a}^i \widetilde{U}\right\|}{\sum_i \left\|\mathbf{b}^i \widetilde{U}\right\|} \leq \lambda = \frac{\sum_i \|\mathbf{a}^i U\|}{\sum_i \|\mathbf{b}^i U\|} \ . \tag{12}$$

Note that, the equalities in Eqs. (9–12) hold if and only if the objective value converges. Therefore, the objective value of the problem of Eq. (6) is decreased in each iteration till converges, which completes the proof of Theorem 1. ■

3. Experimental results

In this section, we experimentally evaluate the proposed M-C2B method in the task of cost-effective video classification using the following data set.

TRECVID 2005 video data set [11] is a benchmark multi-label data set for video analysis with 39 visual concepts, which contains 277 video clips with 61,901 shots. Each shot is represented by a key frame, which is considered as an instance in our study. For each instance frame, following [13, 14, 12] 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.

3.1. Results of standard video classification

We first evaluate the proposed M-C2B method in standard video classification tasks, where we assume the labels of every frame of the training video clips are known and the labels of a video clip is assigned by the labels of all its component frames. This corresponds to the most costly case in real video classification tasks. We perform standard 5-fold cross-validation for evaluation and report the average performance over the 5 trials.

Experimental settings. We first compare our method to two baseline classification methods including support vector machine (SVM) method and transductive support vector machine (TSVM) [5] method. The former is the most widely used supervised classification method in statistical learning, while the latter is a semi-supervised extension of the former. Because the both methods are designed for single-instance data, they are not able to handle data objects with varied sizes. Therefore we train and classify the video data at frame level. Specifically, 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.*, $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\beta \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$, 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} software package.

We also compare our method to two most related methods, *i.e*., two recent MIL methods including miGraph [19] method and MIMLSVM+ [6] method. Because miGraph method is a single-label classification method, one-vs-others strategy is used to conduct classification, one class at a time. We implement these two methods using the codes published by the authors. Note that, because both of these two method and the proposed M-C2B method are multi-instance classification methods, we perform classification at bag (video clip) level. Although in the current experimental settings, we know the instance labels in a priori, 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) [8] method which has demonstrated state-of-the-art video classification performance. Because this method is designed to work at frame level, we employ the same strategy as that used for SVM to conduct classification. We implement the method following its original work and set the parameters as optimal.

We implement three versions of our method to evaluate its various aspects that could impact the classification performance. First, we implement a non-robust version of the proposed method, in which the squared distance is used. Specifically, we square the $d_k\left(\mathbf{x}_i^j, X_{i'}\right)$ in Eq. (1) and stay with the rest definitions. We denote this version as S-C2B method. Second, we compute the C2B distance without using the class specific distance metrics, *i.e*., we compute the C2B distance by Eq. (2), which is a non-parametric method and no learning process is involved. We denote this version as C2B method. Finally, the full version of the proposed method is implemented and denoted as M-C2B method.

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 "$\downarrow$" indicates "the smaller the better" while "$\uparrow$" indicates "the bigger the better". We refer readers to [10] for details of these evaluation metrics.

The average classification performances (mean $\pm$ standard deviation) of the compared methods over the 5 experimental trials are reported in the top half Table 1, from which we can see a number of interesting observations as follows.

First, the proposed method is consistently better than the other compared methods, which demonstrate its effectiveness in video classification with standard yet costly settings.

Second, the MIL methods are generally better than the two baseline classification methods and SML method. This observation is reasonable in that these three methods are all single-instance classification methods, which learn and classify using every frame independently. As a result, the important structural information contained in video clips are not exploited, which leads to inferior performance.

Third, the S-C2B method is worse than the other two versions of the proposed method, which is consistent with our theoretical analysis in that it has no robustness against the outlier instances that abound in multi-instance data. In contrast, although the C2B method is a non-parametric method, it still performs better due to the usage of not-squared distance measures. The proposed M-C2B distance is always better than its two degenerate versions, which confirms the correctness and advantages of our new method for video classification, as well as for multi-instance learning.

Finally, although the proposed method is better than traditional single-instance methods, the improvements are mediocre. This can be attributed to the experimental settings that the labels of all the frames are known, therefore existing methods can still perform well, which, however, is at the price of expensive human labeling cost for training the classification models.

3.2. Study of low cost video classification

Because the main purpose of the proposed M-C2B method is to deal with cases of low cost video classification, in this subsection we evaluate it in the conditions where only video clips labels are given while frame labels are not available. We still employ 5-fold cross-validation for evaluation, and emulate the low cost conditions as following. 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 the condition when the amount of selected frames are 50%, and the corresponding results are reported in the bottom half of Table 1. For the three multi-instance methods, we still conduct classification as before, because they are able to deal with the low cost multi-instance representation by nature. For the three single-instance methods, including the two baseline classification methods and SML method, 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 but not all.

From the results in the bottom half of Table 1 we can see that the proposed method still performs the best, which again demonstrate its effectiveness in video classification. Compared to the results in the top half of Table 1, when labeling cost is reduced, the classification performance degradations of the multi-instance methods, including the proposed M-C2B method, are not very significant, whereas those of the single-instance methods are considerably large.

Table 1. Comparison of video classification performances (mean $\pm$ std). Top: all frames are labeled; bottom: only 50% frames are labeled.

Method	Hamming loss $\downarrow$	One-error $\downarrow$	Coverage $\downarrow$	Rank loss $\downarrow$	Average precision $\uparrow$
SVM	0.183 ± 0.016	0.336 ± 0.018	1.025 ± 0.014	0.186 ± 0.015	0.476 ± 0.022
TSVM	0.180 ± 0.015	0.331 ± 0.016	1.022 ± 0.012	0.183 ± 0.016	0.478 ± 0.025
SML	0.177 ± 0.013	0.327 ± 0.013	1.020 ± 0.019	0.180 ± 0.015	0.480 ± 0.021
miGraph	0.173 ± 0.011	0.306 ± 0.018	1.013 ± 0.018	0.178 ± 0.013	0.483 ± 0.023
MIMLSVM+	0.176 ± 0.014	0.323 ± 0.024	0.999 ± 0.015	0.177 ± 0.010	0.485 ± 0.022
S-C2B	0.176 ± 0.015	0.311 ± 0.014	1.003 ± 0.011	0.174 ± 0.010	0.481 ± 0.022
C2B	0.168 ± 0.016	0.301 ± 0.020	0.986 ± 0.015	0.169 ± 0.014	0.498 ± 0.023
M-C2B	0.161 ± 0.009	0.291 ± 0.011	0.972 ± 0.010	0.154 ± 0.002	0.485 ± 0.013
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
S-C2B	0.220 ± 0.015	0.394 ± 0.014	1.211 ± 0.010	0.221 ± 0.011	0.335 ± 0.012
C2B	0.210 ± 0.011	0.374 ± 0.018	1.199 ± 0.013	0.214 ± 0.013	0.363 ± 0.020
M-C2B	0.196 ± 0.012	0.350 ± 0.013	1.181 ± 0.020	0.198 ± 0.013	0.416 ± 0.018

This important observation provides a concrete evidence to support the usefulness of multi-instance learning in cost effective video classification.

4. Conclusions

In this paper, we proposed a novel M-C2B method to solve the video classification problem with low cost via the MIL model. The class specific distance metrics are introduced to the C2B distance to narrow down the gap between high-level semantic concepts and low-level visual features. To address the overwhelmed outlier instances in multi-instance data caused by instance level labeling ambiguity, we employed a new not-squared distance to learn the distance metrics, which, however, lead a hard objective to solve a highly non-smooth $\ell_{2,1}$-norm minmax problem. We presented an efficient iterative solution algorithm and proved its convergence. The promising experimental results are demonstrated in comprehensive empirical evaluations.

Acknowledgments. This research was partially supported by NSF-CNS 0923494, NSF-IIS 1041637, NSF-CNS 1035913, NSF-IIS 1117965.

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 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] M. Guillaumin, J. Verbeek, and C. Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *ECCV*, 2010.

[4] R. Jin, S. Wang, and Z. Zhou. Learning a distance metric from multi-instance multi-label data. In *IEEE CVPR*, 2009.

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

[6] 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.

[7] W. Liu, S. Ma, D. Tao, J. Liu, and P. Liu. Semi-supervised sparse metric learning using alternating linearization optimization. In *SIGKDD*, 2010.

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

[9] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and Robust Feature Selection via Joint l2,1-Norms Minimization. In *NIPS*, 2010.

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

[11] A. Smeaton and P. Over. TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video. *Image and Video Retrieval*, pages 451–456, 2003.

[12] H. Wang, C. Ding, and H. Huang. Multi-label linear discriminant analysis. In *ECCV*, 2010.

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

[14] H. Wang, H. Huang, and C. Ding. Multi-label feature transform for image classifications. In *ECCV*, 2010.

[15] H. Wang, H. Huang, F. Kamangar, F. Nie, and C. Ding. Maximum margin multi-instance learning. In *NIPS*, 2011.

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

[17] H. Wang, F. Nie, H. Huang, and Y. Yang. Learning frame relevance for video classification. In *Proceedings of the 19th ACM international conference on Multimedia (ACM MM)*, pages 1345–1348, 2011.

[18] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *NIPS*, 2002.

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