Non-I.I.D. Multi-Instance Dimensionality Reduction by Learning a Maximum Bag Margin Subspace

Wei Ping
Tsinghua National Laboratory for Information Science and Technology(TNList), School of Software, Tsinghua University
weiping.thu@gmail.com

Ye Xu
National Key Laboratory for Novel Software Technology, Nanjing University
yexu@smail.nju.edu.cn

Kexin Ren
Dept. of Electronic Engineering & Automation, Nanjing Univ. of Aeronautics and Astronautics
renkexin@nuaa.edu.cn

Chi-Hung Chi
School of Software, Tsinghua University
chichihung@mail.tsinghua.edu.cn

Furao Shen
National Key Laboratory for Novel Software Technology, Nanjing University
frshen@nju.edu.cn

Abstract

Multi-instance learning, as other machine learning tasks, also suffers from the curse of dimensionality. Although dimensionality reduction methods have been investigated for many years, multi-instance dimensionality reduction methods remain untouched. On the other hand, most algorithms in multi-instance framework treat instances in each bag as independently and identically distributed (*i.i.d.*) samples, which fail to utilize the structure information conveyed by instances in a bag. In this paper, we propose a multi-instance dimensionality reduction method, which treats instances in each bag as *non-i.i.d.* samples. To capture the structure information conveyed by instances in a bag, we regard every bag as a whole entity. To utilize the bag label information, we maximize the bag margin between positive and negative bags. By maximizing the defined bag margin objective function, we learn a subspace to obtain salient representation of original data. Experiments demonstrate the effectiveness of the method.

Introduction

Multi-instance learning originated from investigating drug activity prediction (Dietterich, Lathrop, and Lozano-Perez 1997). In multi-instance learning, each training example is a bag containing many instances. A bag is positive if it contains at least one positive instance; otherwise it is labeled as negative bag. The labels of bags in training set are known. However, we do not know the labels of instances in the bags. The framework of multi-instance has attracted much attention in various domains such as object detection (Chen and Wang 2004), information retrieval (Settles, Graven, and Ray 2008), image classification (Qi et al. 2007), and biomedical informatics (Fung et al. 2007).

The "curse of dimensionality" is a serious problem for machine learning and pattern recognition tasks involving high-dimensional data. Reducing the dimensionality is an important tool for addressing this problem. Based on whether the label information is used or not, techniques for reducing dimensionality can be categorized into three classes, i.e. supervised (Fisher 1936), semi-supervised (Zhang and Yeung 2008), and unsupervised (Devijver and Kittler 1982) dimensionality reduction methods. Due to the ambiguity of not knowing which of the instances in a positive bag are the actual positive instances and which ones are not, either supervised or semi-supervised dimensionality reduction methods, which need to use the labels of instances, could not be directly applied to multi-instance task. On the other hand, unsupervised dimensionality reduction methods that ignore the labels of bags are not suitable to cope with this issue either. Although multi-instance learning often involves high-dimensional data problem, to the best of our knowledge, there is no dimensionality reduction method designed for solving this issue.

Another point worth mentioning is the distributions of instances in bags. Most recent multi-instance methods (Gartner et al. 2002) (Andrews, Tsochantaridis, and Hofmann 2003) (Zhou and Zhang 2003) treat instances in each bag as independently and identically distributed (*i.i.d.*), which ignore the valuable structure information conveyed by instances in a bag. Actually, as mentioned in (Zhou and Xu 2007) and (Qi et al. 2007), instances in a bag are hardly identical and independent. Simply treating different image segmentations as independent instances may lose important information of inter-correlations among instances. For example, in an object detection problem (Fig.1), monkeys are very likely to locate in trees. It means in an image bag (that is divided into several image segmentations/instances), the instances that contain monkeys are actually correlated with those instances that contain trees. Such correlation information among instances is helpful to the object detection applications (Chen and Wang 2004). Therefore, it is more desired to treat instances in each bags as *non-i.i.d.* samples in multi-instance learning. However, how to explore the dependency relations of instances in each bag is a thorny issue, and few multi-instance methods can cope with the challenge.

In this paper, we propose a multi-instance dimensionality reduction algorithm named as MidLABS (Multi-Instance Dimensionality reduction by Learning a mAximum Bag margin Subspace), which does not regard instances in each bag as *i.i.d.* samples. To capture the structure information conveyed by instances, every bag is treated as a whole entity

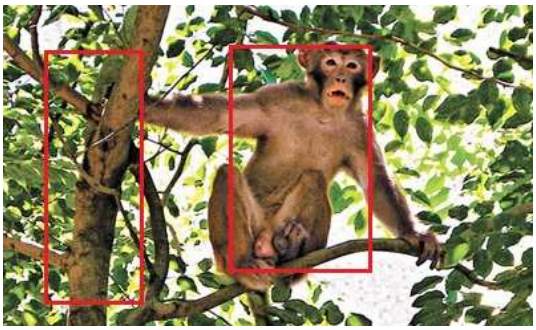

Figure 1: In the image bag, instances(segmentations) containing monkeys are correlated with those instances containing trees.

and instances in the bag are treated as inter-dependent components of the entity. According to this basic idea, we define a bag margin objective function that involves the discriminative information of bags. Then, based on the criteria of maximizing the margin of positive and negative bags, a subspace is learned for multi-instance dimensionality reduction task.

We organize the rest of this paper as follows. In section 2, we briefly introduce some related works. MidLABS is proposed in section 3. In section 4, we report on experimental results. Finally in Section 5, we conclude the paper and discuss some future directions.

Related Work

As a powerful framework, multi-instance learning (Dietterich, Lathrop, and Lozano-Perez 1997) has been investigated for many years, and many algorithms have been developed, such as Diverse Density (Maron and Lozano-Perez 1998), Bayesian-KNN (Wang and Zucker 2000), MI kernels (Gartner et al. 2002), MI SVMs (Andrews, Tsochantaridis, and Hofmann 2003), MI Ensembles (Zhou and Zhang 2003) et al. However, as indicated in (Zhou and Xu 2007), all such methods treat instances in each bag as independently and identically distributed samples, which ignore the important structure information conveyed by inter-correlated instances in each bag. Nevertheless, encoding structure information of instances into multi-instance learner is a difficult issue. As far as we know, (Zhou, Sun, and Li 2009) is the only multi-instance learner that treats instances as *non-i.i.d.* samples. By defining a graph kernel among bags, it explores the structure information of instances in each bag which significantly improves the classification performance in multi-instance framework.

Our work also relates to Dimensionality Reduction (DR), which is an effective tool to cope with "the curse of dimensionality". Many dimensionality reduction algorithms have been proposed during past few years. According to whether the label information is needed, Dimensionality Reduction algorithms can be categorized into supervised DR, semi-supervised DR, and unsupervised DR methods. Linear Discriminant Analysis (LDA) (Fisher 1936), along with some LDA based methods (Friedman 1989) (Howland, Jeon, and Park 2003) (Ye et al. 2004), achieves maximum discrimination by finding an transformation that maximizes between-class distance and minimizes within-class distance simultaneously. Maximum Margin Criterion (MMC) (Li and Jiang 2003) and some variants (Yan et al. 2004) optimize a different objective function from LDA, but achieve the goal of maximum discrimination as well. Semi-supervised Discriminant Analysis (SSDA) (Zhang and Yeung 2008) aims at fulfilling dimensionality reduction task when supervisory information is available for some but not all training samples. However, those supervised and semi-supervised DR algorithms need to avail the labels of instances in training set, which are impossible to be applied in multi-instance framework. Principal Component Analysis (PCA) (Devijver and Kittler 1982) which aims at finding the projection that maximize the data variance, along with some other similar algorithms (Weng, Zhang, and Hwang 2003) (Bartelmaos and Abed-Meraim 2008) (Xu et al. 2009) are a typical class of unsupervised DR methods. Some other unsupervised nonlinear dimensionality reduction schemes such as Locally Linear Embedding (LLE) (Roweis and Saul 2000), ISOMAP (Tenenbaum, Silva, and Langford 2000), and Locality Preserving Projections (LPP) (He and Niyogi 2003) employ local symmetries to learn the global structure of original data. Such methods can compute a low-dimensional embedding of a set of high-dimensional original data. However, simply using those unsupervised DR methods to address the high-dimensional data in multi-instance learning framework will miss the label information of bags. Therefore, a DR method that can take advantage of bag label information in multi-instance learning framework is very desired. However, to the best of our knowledge, there is no such method.

In this paper, we propose a dimensionality reduction method for multi-instance learning framework. The label information of bags is taken advantage of to ensure a better discriminant performance. Moreover, we consider the instances in each bag as *non-i.i.d.* samples, which takes the structure information of instances in each bag into account.

The Proposed MidLABS Method

In this section, we introduce our MidLABS algorithm, which takes advantage of both discriminant information of bags and geometrical structures of instances in each bag.

Dimensionality Reduction for Multi-Instance Learning Problem

Before presenting the dimensionality reduction problem for multi-instance learning, we give the formal description of multi-instance learning as follows. Let $\mathcal{X}$ denote the instance space. Given a data set $T = \{(\mathbf{X_1}, L_1), ..., (\mathbf{X_i}, L_i), ..., (\mathbf{X_N}, L_N)\}$, where $\mathbf{X_i} = \{\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{ij}, ..., \boldsymbol{x}_{i,n_i}\} \subset \mathcal{X}$ is called a bag, $L_i \in \mathcal{L} = \{-1, +1\}$ is the label of $\mathbf{X_i}$, and N is the number of training bags, the goal is to learn some concept from the training set for correctly labeling unseen bags. Here $\boldsymbol{x}_{ij} \in \mathbf{X}$ is an instance $[x_{ij1}, ..., x_{ijk}, ..., x_{ijD}]^\top$, where x_{ijk} is the value of $\boldsymbol{x}_{ij}$ at the k^{th} attribute, n_i is the number of instances in $\mathbf{X_i}$,

and D is the dimension of original space $\mathcal{X}$. If there exists $p \in \{1, ..., n_i\}$ such that $\boldsymbol{x}_{ip}$ is a positive instance, then $\mathbf{X_i}$ is a positive bag and thus $L_i = +1$, but the concrete value of the index p is usually unknown; otherwise $L_i = -1$.

Then, the problem of dimensionality reduction for multi-instance learning is explained as follows: given a data set $T = \{(\mathbf{X_1}, L_1), ..., (\mathbf{X_i}, L_i), ..., (\mathbf{X_N}, L_N)\}$ as above, finding a transformation matrix $\mathbf{W} = [\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_d}]$, which maps every $\boldsymbol{x}_{ij} \in \mathcal{R}^D$ in each bag $\mathbf{X_i}$ to $\boldsymbol{y}_{ij} \in \mathcal{R}^d$ in new bag $\mathbf{Y_i}$ i.e. $\boldsymbol{y}_{ij} = \mathbf{W}^\top \boldsymbol{x}_{ij}$, such that $\boldsymbol{y}_{ij}$ "represents" $\boldsymbol{x_{ij}}$ and $\mathbf{Y_i}$ "represents" $\mathbf{X_i}$, then we could get data set $\{(\mathbf{Y_1}, L_1), ..., (\mathbf{Y_i}, L_i), ..., (\mathbf{Y_N}, L_N)\}$ in feature space.

Maximize the Bag Margin Objective Function for Non-i.i.d. Multi-instance Dimensionality Reduction

We propose a Dimensionality Reduction (DR) method to not only utilize bag label information, but also capture structure information of instances in each bag. To capture the structure information, we treat instances in each bag as a whole entity and establish the local geometric structure. To utilize bag label information, we maximize the bag margin between positive and negative bags after mapping. Specifically, we define a bag margin objective function involving bag labels based on a novel distance metric among bags which takes geometrical structure into account. Then the question of dimensionality reduction can be fulfilled by learning a subspace which maximizes the bag margin objective function. In the rest of this subsection, the proposed MidLABS will be described in details.

First, we consider the particular problem of mapping the instances in each bag to a line, i.e. $y_{ij} = \mathbf{w}^\top \boldsymbol{x_{ij}}, i = 1 : N; j = 1 : n_i$, so that different classes of bags stay as distant as possible, whereas same class of bags stay as close as possible. In other words, our goal is choosing a "good" vector $\mathbf{w}$, $\|\mathbf{w}\| = 1$, which maximizes the margin of between-class bags and minimizes the margin of within-class bags. Before formally introducing the criterion function of this goal, we must define a distance metric of bags in this line. A reasonable one is:

$$Dis(\mathbf{X_i}, \mathbf{X_j}) = \sum_{a=1}^{n_i} \sum_{b=1}^{n_j} (y_{ia} - y_{jb})^2 \tag{1}$$

where $y_{ia} = \mathbf{w}^\top \boldsymbol{x_{ia}}$ is the mapped point of $\boldsymbol{x_{ia}}$ on this line from bag $\mathbf{X_i}$, and $y_{jb} = \mathbf{w}^\top \boldsymbol{x_{jb}}$ is the mapped point of $\boldsymbol{x_{jb}}$ on this line from bag $\mathbf{X_j}$. This definition means that the distance between two bags is measured by the sum of pairwise instances from different bags. Similar metric is employed in (Gartner et al. 2002) to setup kernels between multi-instance bags by treating each bag as a set. This pairwise metric methodology has been proven to be effective to measure the similarity among bags.

However, as we mentioned before, in order to capture the structure information conveyed by instances, the instances in a bag should not be simply treated as *i.i.d.* samples. As Fig.2 illustrates, the instances(denoted as small squares) inside each bag(denoted as ellipse) are correlated in fact, and the two bags are obviously dissimilar. But if we treat these instances as *i.i.d.* samples as indicated in Fig.3, the two

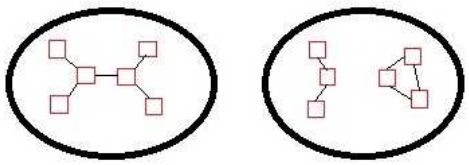

Figure 2: The two bags are obviously dissimilar according to their different inter-correlations among instances.

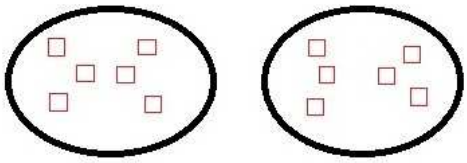

Figure 3: If the instances in each bag are treated as *i.i.d.* samples, the structure information conveyed by instances is ignored. In this case the two bags which are actually different would be regarded as similar to each other.

bags are regarded as similar. Therefore, the inter-instance information that provides critical structure information is worth considering when defining the metric among bags. As (Tenenbaum, Silva, and Langford 2000) shows that ϵ-graph is useful for discovering the underlying manifold structure of data, we establish an ϵ-graph for every bag to discover the geometrically dependent information among instances inside the bag. This method is first employed in (Zhou, Sun, and Li 2009), and appears to be effective in practical applications. The procedure is very straightforward. For a bag $\mathbf{X_i}$, we view every instance of it as a node. Then, we compute the distance of every pair of nodes, e.g. $\boldsymbol{x_{iu}}$ and $\boldsymbol{x_{iv}}$. If the distance between them is smaller than a pre-set threshold ϵ, then an edge is established between those two nodes. To reflect the "texture" information of each edge in bag, we define the edge as the vector difference of two associated nodes. We choose the node which has the first larger attribute as the start node. For example, we have two nodes $\boldsymbol{x_{iu}}$ and $\boldsymbol{x_{iv}}$, and the distance between them is smaller than ϵ. If there exists a $k \in [1, D]$, such that $x_{iuk} > x_{ivk}$ and $x_{ium} = x_{ivm}$ for all $m \in [1, k-1]$, we choose $\boldsymbol{x_{iu}}$ as the start node. Hence, the edge associated with $\boldsymbol{x_{iu}}$ and $\boldsymbol{x_{iv}}$ is $\boldsymbol{e} = \boldsymbol{x_{iu}} - \boldsymbol{x_{iv}}$. From this method, we could extract the "texture" or *non-i.i.d.* information among instances inside the bag. Now, we redefine the distance metric of bags (graphs) in the line $\mathbf{w}$. We use Dis_{node} to incorporate the information conveyed by the nodes, and use Dis_{edge} to incorporate the *non-i.i.d.* information conveyed by the edges. Formally,

$$\begin{aligned} Dis_G(\mathbf{X_i}, \mathbf{X_j}) &= Dis_{node}(\mathbf{X_i}, \mathbf{X_j}) + C \cdot Dis_{edge}(\mathbf{X_i}, \mathbf{X_j}) \\ &= \sum_{a=1}^{n_i} \sum_{b=1}^{n_j} (\mathbf{w}^\top \boldsymbol{x_{ia}} - \mathbf{w}^\top \boldsymbol{x_{jb}})^2 + \\ &\quad C \cdot \sum_{c=1}^{m_i} \sum_{d=1}^{m_j} (\mathbf{w}^\top \boldsymbol{e_{ic}} - \mathbf{w}^\top \boldsymbol{e_{jd}})^2 \end{aligned} \tag{2}$$

Where $\mathbf{w}^\top \boldsymbol{x_{ia}}$ and $\mathbf{w}^\top \boldsymbol{x_{jb}}$ are declared as before, $\mathbf{w}^\top \boldsymbol{e_{ic}}$ is the projection of the edge $\boldsymbol{e_{ic}}$ from bag $\mathbf{X_i}$ in the line, $\mathbf{w}^\top \boldsymbol{e_{jd}}$ is the projection of the edge $\boldsymbol{e_{jd}}$ from bag $\mathbf{X_j}$ in the

line, m_i is the number of edges in $\mathbf{X_i}$, m_j is the number of edges in $\mathbf{X_j}$, and C is the weight ratio of node and edge. (In our paper, we set $C = 1$ for the purpose of convenience.) To avoid numerical problem or any preference for the bag which has large number of instances, Dis_G is normalized to

$$Dis_G(\mathbf{X_i}, \mathbf{X_j}) = \frac{Dis_{node}(\mathbf{X_i}, \mathbf{X_j})}{n_i n_j} + C\frac{Dis_{edge}(\mathbf{X_i}, \mathbf{X_j})}{n_i^2 n_j^2} \quad (3)$$

where n_i and n_j are the number of instances in $\mathbf{X_i}$ and $\mathbf{X_j}$ respectively. Note that, Dis_{edge} is divided by $n_i^2 n_j^2$, because the number of edges in a bag is usually proportional to the square of nodes number. Based on above distance metric among bags, we could formally introduce the criterion for choosing $\mathbf{w}$ which maximizes the margin of between-class bags and minimizes the margin of within-class bags after mapping. The objective function which needs to be optimized is

$$\mathbf{J}(\mathbf{w}) = \frac{\sum_{\mathbf{L_i}\neq\mathbf{L_j}} Dis_G(\mathbf{X_i}, \mathbf{X_j})}{\sum_{\mathbf{L_i}=\mathbf{L_j}} Dis_G(\mathbf{X_i}, \mathbf{X_j})} \quad (4)$$

where numerator represents the margin of between-class mapped bags, and denominator represents the margin of within-class mapped bags. Therefore, maximizing (4) is an attempt to ensure that: if $\mathbf{X_i}$ and $\mathbf{X_j}$ share the same label, they stay as close as possible after mapping; if $\mathbf{X_i}$ and $\mathbf{X_j}$ have different labels, they stay as distant as possible after mapping.

Optimal Linear Subspace

We solve the objective functions (4) in closed form. Following some simple algebraic steps, the $Dis_G(\mathbf{X_i}, \mathbf{X_j})$ in (3) can be written as follows:

$$\begin{aligned} Dis_G(\mathbf{X_i}, \mathbf{X_j}) &= \frac{\sum_{a=1}^{n_i}\sum_{b=1}^{n_j}(\mathbf{w}^\top \boldsymbol{x_{ia}} - \mathbf{w}^\top \boldsymbol{x_{jb}})^2}{n_i n_j} \\ &\quad + C\frac{\sum_{c=1}^{m_i}\sum_{d=1}^{m_j}(\mathbf{w}^\top \boldsymbol{e_{ic}} - \mathbf{w}^\top \boldsymbol{e_{jd}})^2}{n_i^2 n_j^2} \\ &= \mathbf{w}^\top \frac{\sum_{a=1}^{n_i}\sum_{b=1}^{n_j}(\boldsymbol{x_{ia}} - \boldsymbol{x_{jb}})(\boldsymbol{x_{ia}} - \boldsymbol{x_{jb}})^\top}{n_i n_j}\mathbf{w} \\ &\quad + \mathbf{w}^\top C\frac{\sum_{c=1}^{m_i}\sum_{d=1}^{m_j}(\boldsymbol{e_{ic}} - \boldsymbol{e_{jd}})(\boldsymbol{e_{ic}} - \boldsymbol{e_{jd}})^\top}{n_i^2 n_j^2}\mathbf{w} \quad (5) \end{aligned}$$

We denote

$$\begin{aligned} \mathbf{K_{ij}} &= \frac{\sum_{a=1}^{n_i}\sum_{b=1}^{n_j}(\boldsymbol{x_{ia}} - \boldsymbol{x_{jb}})(\boldsymbol{x_{ia}} - \boldsymbol{x_{jb}})^\top}{n_i n_j} \\ &\quad + C\frac{\sum_{c=1}^{m_i}\sum_{d=1}^{m_j}(\boldsymbol{e_{ic}} - \boldsymbol{e_{jd}})(\boldsymbol{e_{ic}} - \boldsymbol{e_{jd}})^\top}{n_i^2 n_j^2} \quad (6) \end{aligned}$$

Then

$$Dis_G(\mathbf{X_i}, \mathbf{X_j}) = \mathbf{w}^\top \mathbf{K_{ij}} \mathbf{w} \quad (7)$$

Now, the objective function (4) can be reduced to

$$\mathbf{J}(\mathbf{w}) = \frac{\mathbf{w}^\top(\sum_{\mathbf{L_i}\neq\mathbf{L_j}} \mathbf{K_{ij}})\mathbf{w}}{\mathbf{w}^\top(\sum_{\mathbf{L_i}=\mathbf{L_j}} \mathbf{K_{ij}})\mathbf{w}} \quad (8)$$

Let

$$\mathbf{S_b} = \sum_{\mathbf{L_i}\neq\mathbf{L_j}} \mathbf{K_{ij}} \quad (9)$$

and

$$\mathbf{S_w} = \sum_{\mathbf{L_i}=\mathbf{L_j}} \mathbf{K_{ij}} \quad (10)$$

Obviously, $\mathbf{S_b}$ and $\mathbf{S_w}$ are symmetric. In a sense, $\mathbf{S_b}$ describes the between-class scatter information of bags; $\mathbf{S_w}$ describes the within-class scatter information of bags. Then, $\mathbf{J}(\mathbf{w})$ can be rewritten as the following form:

$$\mathbf{J}(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S_b} \mathbf{w}}{\mathbf{w}^\top \mathbf{S_w} \mathbf{w}} \quad (11)$$

It is a generalized Rayleigh quotient, and could be maximized through Lagrange Multipliers method. Let denominator $\mathbf{w}^\top \mathbf{S_w} \mathbf{w} = \alpha \neq 0$ as a constraint, we can turn (11) as a Lagrange multiplication problem with λ being the Lagrange multiplier and noting that the constraint should be rewritten to equal 0. The Lagrange multiplier function is

$$\mathbf{L}(\mathbf{w}, \lambda) = \mathbf{w}^\top \mathbf{S_b} \mathbf{w} - \lambda(\mathbf{w}^\top \mathbf{S_w} \mathbf{w} - \alpha) \quad (12)$$

Taking the first order derivative with respect to $\mathbf{w}$ yields

$$\frac{\partial \mathbf{L}(\mathbf{w}, \lambda)}{\partial \mathbf{w}} = 2\mathbf{S_b}\mathbf{w} - 2\lambda \mathbf{S_w}\mathbf{w} \quad (13)$$

Let the derivative equal zero, then

$$\mathbf{S_b}\mathbf{w} = \lambda \mathbf{S_w}\mathbf{w} \quad (14)$$

We note that equation (14) is precisely in the form of generalized eigenvalue problem with $\mathbf{w}$ being the eigenvector. That means the projection vector w that maximizes (11) is given by the maximum eigenvalue solution to this generalized eigenvalue problem[1]. Let the column vector $\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_d}$ be the solution of (14) ordered according to their eigenvalues $\lambda_1 > \lambda_2 > \lambda_d$. Thus, the dimensionality reduction is given as follows:

$$\begin{aligned} &\boldsymbol{x_{ij}} \rightarrow \boldsymbol{y_{ij}} = \mathbf{W}^\top \boldsymbol{x_{ij}}; \quad (15) \\ &\mathbf{W} = [\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_d}] \end{aligned}$$

Where $\boldsymbol{x_{ij}}$ is a D-dimensional vector from bag $\mathbf{X_i}$, $\boldsymbol{y_{ij}}$ is the d-dimensional vector represented $\boldsymbol{x_{ij}}$, and $\mathbf{W}$ is an $D \times d$ matrix.

Summary of the Proposed Algorithm

According to the above analysis, the detailed algorithm is given in Algorithm 1. First, we treat instances in each bag as a whole entity and build a ϵ-graph for every bag. The structure information conveyed by instances in each bag is taken into account. Then, we compute $\mathbf{S_b}$ and $\mathbf{S_w}$ that reflect the discriminant information of bags. Finally, we solve the generalized eigenvalue equation, and get the transformation matrix $\mathbf{W}$.

[1]Discussions on numerical stability of generalized eigenproblems can be found in (Ye et al. 2004). But it is out of the scope of this paper.

Input: Data set $\{(\mathbf{X_1}, L_1), , (\mathbf{X_i}, L_i), , (\mathbf{X_N}, L_N)\}$ and the target dimension d

1: Construct the ϵ-graph for every bag $\mathbf{X_i}$, establish edges inside bag.
2: Compute $\mathbf{S_b}$ and $\mathbf{S_w}$ according to (9) and (10).
3: Solve the generalized eigenvalue equation (14).
4: Construct the $D \times d$ matrix $\mathbf{W}$ whose columns are composed by the eigenvectors corresponding to largest d eigenvalues.

Output: $\mathbf{W}$: the projection from $\mathcal{R}^D$ to $\mathcal{R}^d$.

Algorithm 1: Pseudo-code of the MidLABS method

Experiments

We compare MidLABS with two typical dimensionality reduction methods, including a linear dimensionality reduction method PCA (Devijver and Kittler 1982) and a nonlinear dimensionality reduction method LLE (Roweis and Saul 2000). The MI-Kernel method (Gartner et al. 2002) is used for classification after dimensionality reduction. As a baseline, we also evaluate the performance of MI-Kernel in the original feature space (denoted by ORI).

In our experiments, the classification accuracy (CA) and dimension ratio (DR) are used to evaluate the proposed method. Here, dimension ratio is defined as the target dimension d dividing the original dimension D.

Musk Data Set

In this subsection, we evaluate the proposed MidLABS on the benchmark data sets popularly used in multi-instance learning, including Musk1 and Musk2. Each instance inside the bag is represented by a 166-dimensional vector in both Musk1 and Musk2 (i.e. $D = 166$). More detailed information about the datasets can be available in (Dietterich, Lathrop, and Lozano-Perez 1997).

Table 1: Classification Accuracy (CA) and Dimension Ratio (DR) of MidLABS, PCA, LLE, and ORI under Musk1 and Musk2.

Algorithm	MidLABS	PCA	LLE	ORI
Musk1 CA	$\mathbf{90.0}\%$	87.5%	85.9%	86.4%
	$\pm\mathbf{2.7}\%$	$\pm4.1\%$	$\pm5.1\%$	$\pm3.1\%$
Musk1 DR	$\mathbf{18.1}\%$	33.1%	36.1%	100.0%
Musk2 CA	85.3%	86.2%	85.2%	$\mathbf{88.0}\%$
	$\pm1.8\%$	$\pm2.8\%$	$\pm2.5\%$	$\pm\mathbf{1.5}\%$
Musk2 DR	$\mathbf{12.1}\%$	39.2%	48.2%	100.0%

In this experiment, we obtain the optimal dimension in feature space via ten times 10-fold cross validation (i.e., we repeat 10-fold cross validation for ten times with different random data partitions). After dimensionality reduction, we employ MI-Kernel to evaluate classification performance by 10-fold cross validation. For each method, Table 1 lists the average classification accuracy (CA) and the dimension ratio (DR), i.e. the target dimension dividing by the original dimension. The best performances are highlighted with figures in bold typeface. The results dedicate that the proposed method has significant improvement in classification performance than ORI under Musk1. Treating instances in each bag as *non-i.i.d.* samples, MidLABS successfully explores structure information. Therefore, it is no wonder that MidLABS can capture salient features and obtain a higher classification accuracy. Compared with PCA and LLE, the classification performance of MidLABS is higher under Musk1. Besides that MidLABS is able to obtain higher classification accuracy, the dimension ratio of MidLABS is better than other methods as well, as indicated in Table 1. Under Musk2, the classification accuracy is a slightly poorer. It is because that in MidLABS, we use pairwise metric to evaluate the distance among bags and there are a relatively large number of instances in each bag. Under this distance metric, too many negative instances in the positive bag will overwhelm the positive instance.

Automatic Image Annotation

In this part, we evaluate MidLABS method on three data sets for image annotation tasks, including Elephant, Fox, and Tiger. In each case, the data sets have 100 positive and 100 negative example bags. Each instance inside the bag is represented by a 230-dimensional vector (i.e. $D = 230$). More details can be found in (Andrews, Tsochantaridis, and Hofmann 2003).

As in last subsection, the optimal dimension in feature space is tuned by ten times 10-fold cross validation. We employ MI-Kernel to evaluate classification accuracy (CA), which is recorded in Table 2. The results show that the proposed method is better than ORI in classification accuracy. Compared with PCA and LLE, the classification performance of MidLABS is higher under all the three image databases. To the aspects of dimension ratio (DR) that is indicated in Table 2, MidLABS is averagely better than other methods.

Table 2: Classification Accuracy (CA) and Dimension Ratio (DR) of MidLABS, PCA, LLE, and ORI under datasets of Elephant, Fox, and Tiger.

Algorithm	MidLABS	PCA	LLE	ORI
Elephant CA	$\mathbf{86.5}\%$	86.0%	84.0%	84.3%
	$\pm\mathbf{1.4}\%$	$\pm1.2\%$	$\pm1.1\%$	$\pm1.6\%$
Elephant DR	$\mathbf{13.0}\%$	17.4%	26.1%	100.0%
Fox CA	$\mathbf{67.0}\%$	64.0%	66.0%	60.3%
	$\pm\mathbf{2.1}\%$	$\pm2.4\%$	$\pm2.2\%$	$\pm1.9\%$
Fox DR	$\mathbf{17.4}\%$	21.3%	20.4%	100.0%
Tiger CA	$\mathbf{87.5}\%$	84.5%	86.0%	84.2%
	$\pm\mathbf{1.6}\%$	$\pm1.9\%$	$\pm1.9\%$	$\pm1.0\%$
Tiger DR	21.7%	21.7%	$\mathbf{17.4}\%$	100.0%

Conclusion and Future Works

In this paper, we propose a multi-instance dimensionality reduction algorithm named as MidLABS. Treating instances in each bag as *non-i.i.d.* samples, MidLABS effectively captures important structure information conveyed by instances in each bag, which plays an important role in obtaining

salient features from original data space. Meanwhile, MidLABS takes advantage of the label information of bags to guarantee a powerful discriminant ability. The results of experiments validate the effectiveness of MidLABS.

In the future, we intend to design a framework to simultaneously reduce the dimensionality of data and the number of instances in each bag because it is often the case that in multi-instance learning tasks the number of instances is huge and the problem of reducing the instances number is at least as important as the one tackled in this paper.

Acknowledgments

We thank the anonymous reviewers, Yida Wang from Princeton University and Xinjian Guo from Shandong University for their invaluable inputs. This work was supported in part by the China NSF grant(#60723003, #60975047 and #90604028) and 863 project #2008AA01Z129.

References

Andrews, S.; Tsochantaridis, I.; and Hofmann, T. 2003. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*, 561–568.

Bartelmaos, S., and Abed-Meraim, K. 2008. Fast principal component extraction using givens rotations. *IEEE Signal Processing Letters* 15:369–372.

Chen, Y., and Wang, J. Z. 2004. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research* 5:913–939.

Devijver, P. A., and Kittler, J. 1982. *Pattern Recognition: A Statistical Approach*. Prince-Hall.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Perez, T. 1997. Solving the multiple-instance problem with axis-parallel retangles. 89(1):31–71.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(2):179–188.

Friedman, J. 1989. Regularized discriminant analysis. *Journal of American Statistical Association* 84(405):165–175.

Fung, G.; Dundar, M.; Krishnappuram, B.; and Rao, R. B. 2007. Multiple instance learning for computer aided diagnosis. In *Advances in Neural Information Processing Systems*, 425–432.

Gartner, T.; Flach, P. A.; Kowalczyk, A.; and Smola, A. 2002. Multi-instance kernels. In *Proceedings of the International Conference on Machine Learning*, 179–186.

He, X., and Niyogi, P. 2003. Locality preserving projections. In *Advances in Neural Information Processing Systems*.

Howland, P.; Jeon, M.; and Park, H. 2003. Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. *SIAM Journal of Matrix Analysis and Applications* 25(1):165–179.

Li, H. F., and Jiang, T. 2003. Efficient and robust feature extraction by maximum margin criterion. In *Advances in Neural Information Processing Systems*, 97–104.

Maron, O., and Lozano-Perez, T. 1998. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems*, 570–576.

Qi, G.-J.; Hua, X.-S.; Rui, Y.; Mei, T.; Tang, J.; and Zhang, H.-J. 2007. Concurrent multiple instance learning for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290:2323–2326.

Settles, B.; Graven, M.; and Ray, S. 2008. Multiple-instance active learning. In *Advances in Neural Information Processing Systems*, 1289–1296.

Tenenbaum, J. B.; Silva, V.; and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290:2319–2323.

Wang, J., and Zucker, J.-D. 2000. Solving the multiple-instance problem: a lazy learning approach. In *Proceedings of the International Conference on Machine Learning*, 1119–1125.

Weng, J.; Zhang, Y.; and Hwang, W. S. 2003. Candid covariance-free incremental principal component analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence* 25(8):1034–1040.

Xu, Y.; Shen, F.; Zhao, J.; and Hasegawa, O. 2009. To obtain orthogonal feature extraction using training data selection. In *ACM Conference on Information and Knowledge Management*, 1819–1822.

Yan, J.; Zhang, B.; Yan, S.; Yang, Q.; Li, H.; Chen, Z.; Xi, W.; Fan, W.; Ma, W.-Y.; and Cheng, Q. 2004. Immc: incremental maximum margin criterion. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 725–730.

Ye, J.; Janardan, R.; Park, C.; and Park, H. 2004. An optimization criterion for generalized discriminant analysis on undersampled problems. *IEEE Trans. Pattern Analysis and Machine Intelligence* 26(8):982–994.

Zhang, Y., and Yeung, D.-Y. 2008. Semi-supervised discriminant analysis via cccp. In *Proceedings of the European Conference on Machine Learning*, 644–659.

Zhou, Z.-H., and Xu, J.-M. 2007. On the relation between multi-instance learning and semi-supervised learning. In *Proceedings of the International Conference on Machine Learning*, 1167–1174.

Zhou, Z.-H., and Zhang, M.-L. 2003. Ensembles of multi-instance learners. In *Proceedings of the European Conference on Machine Learning*, 492–502.

Zhou, Z.-H.; Sun, Y.-Y.; and Li, Y.-F. 2009. Multi-instance learning by treating instances as non-i.i.d. samples. In *Proceedings of the International Conference on Machine Learning*, 1249–1256.