Computers and Mathematics with Applications 64 (2012) 500–510

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Latent topic based multi-instance learning method for localized content-based image retrieval

Da-xiang Li *, Jiu-lun Fan, Dian-wei Wang, Ying Liu

School of Telecommunication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an, 710121, China

ARTICLE INFO

Article history:
Received 7 March 2010
Received in revised form 7 September 2011
Accepted 9 December 2011

Keywords:
Multi-instance learning (MIL)
Image retrieval
Probabilistic latent semantic analysis (PLSA)
Transductive support vector machine (TSVM)

ABSTRACT

Focusing on the problem of localized content-based image retrieval, based on probabilistic latent semantic analysis (PLSA) and transductive support vector machine (TSVM), a novel semi-supervised multi-instance learning (SSMIL) algorithm is proposed, where a bag corresponds to an image and an instance corresponds to the low-level visual features of a segmented region. In order to convert an MIL problem into a standard supervised learning problem, first, all the instances in training bags be clustered by K-Means method, and regards each cluster center as "*visual-word*" to build a visual vocabulary. Second, according to the distance between "*visual-word*" and instance, a fuzzy membership function is defined to establish a fuzzy term-document matrix, then use PLSA method to obtain bag's (image's) latent topic models, which can convert every bag to a single sample. Finally, in order to use the unlabeled images to improve retrieval accuracy, using semi-supervised TSVM to train classifiers. Experimental results on the COREL data sets show that the proposed method, named PLSA-SSMIL, is robust, and its performance is superior to other key existing MIL algorithms.

1. Introduction

Traditional content-based image retrieval (CBIR) [1] usually measures the overall visual similarity between images, while the users often judge the similarity of images based on "semantic similarity" rather than "visual similarity". Because of the existence of "semantic gap", it is difficult for traditional CBIR techniques to obtain satisfactory retrieval results. In general, in order to establish a bridge between the low-level visual features of images and the high-level semantic, some supervised machine learning methods (i.e. SVM) must be used to train classifiers and training samples need to be accurately labeled. When carries on semantic image retrieval under the framework of supervised learning, there are three major problems: (1) The problem of representing the semantic of an image, which is the study of representing the variety of high-level semantics of an image; (2) The problem of training sample annotation. Because manually collecting (and possibly further annotating, aligning, cropping, etc.) training examples is a labor intensive endeavor, and the process is both time consuming and error-prone [2]; (3) The problem of small sample learning. Because hand-labeled training samples are difficult, user wishes to provide as little training samples as possible in semantic image retrieval.

Aiming at above problems, we propose a novel semi-supervised multi-instance learning (SSMIL) algorithm based on probabilistic latent semantic analysis (PLSA) and transductive support vector machine (TSVM), where each image as a bag, the low-level visual feature vectors of the segmented regions as instances. Moreover, an image is labeled as a positive bag if the user regards it as semantically relevant (otherwise it will be labeled as a negative bag), so semantic-based image retrieval

* Corresponding author. Tel.: +86 15353707981; fax: +86 29 88787778.
E-mail address: www_ldx@163.com (D.-x. Li).

0898-1221/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2011.12.030

problem can be transformed into an MIL problem [3]. Because MIL allows for coarse labeling at the image level, instead of fine labeling at region level, it can significantly improve the efficiency of labeling training samples.

The organization of this paper is as follows: Section 2, we introduce recent work related to MIL. Section 3 gives detailed description of the proposed new feature representation scheme, including the method of constructing visual vocabulary and the PLSA based method to extract bag's latent topic feature. Section 4 introduces PLSA-SSMIL method. Experimental results and analysis are presented in Section 5. Section 6 contains our conclusion and future work.

2. Related work

MIL has become an active area of investigation in machine learning since it was first put forward for drug activity prediction [4]. In the MIL problems, the training samples are regarded as bags, where each bag consists of multiple instances. Training labels are associated with bags rather than instances. In other word, the labels of the training instances are unknown, and only the labels of the bags are known. According to the original MIL definition, a bag is labeled as positive if at least one of its instances is positive, and it is labeled as negative if all of its instances are negative. The goal of an MIL algorithm is to generate a classifier that will classify unseen bags correctly.

During the past decade, many MIL algorithms have been presented, including axis parallel hyper-rectangles [4], Citation-kNN [5], Diverse Density (DD) [6], DD with Expectation Maximization (EM-DD) [7], Neural Network [8], etc. It is difficult to list all existing MIL algorithms. Here, we mainly focus on methods based on the support vector machines (SVM), which have been successfully used in many machine-learning problems. Andrews et al. [9] first modified the SVM formulation, and presented mi-SVM and MI-SVM algorithms. However, unlike the standard SVM, they lead to non-convex optimization problems, which suffer from local minima. Therefore, Gehler and Chapelle [10] applied deterministic annealing to solve this non-convex optimization problem, and proposed AL-SVM method, which could find better local minima of the objective function. Gartner et al. [11] designed kernel functions (i.e. set kernel and statistics kernel) directly on the bags, and uses a standard SVM to solve MIL problem. Because of the instance labels were unavailable, these multi-instance kernels implicitly assumed that all instances in a bag be equally important. This is a very crude assumption. Therefore, Kwok and Cheung [12] designed marginalized kernel and Zhou et al. [13] designed graph kernel for MIL by considering the different contributions of different instances. Recently, some meta-data based methods have been proposed, which convert each bag in the MIL problem into a single meta-data so that standard single instance learning (SIL) methods (i.e. SVM method) can be used to solver MIL problem. For example, DD-SVM [14], Multi-Instance Learning via Embedded Instance Selection (MILES) [15], CCE [16], MICCLLR [17] and EC-SVM [3], etc. Several techniques for mapping bags into single meta-data are discussed in the following. In this paper, we use the term, single instance learning (SIL) to refer to the traditional supervised learning paradigm in which each individual example has a class label.

Converting every bag in the MIL problem into a single representation vector, and then using a standard SIL method to solve the MIL problem, is a kind of very effective MIL algorithms. However, few existing feature representation methods are effective to describe the bags. This makes it difficult to adapt some well-known SIL methods for MIL problems. Inspired by DD-SVM [14] and MILES [15] methods, in this paper, we present a probabilistic latent semantic analysis (PLSA) [18] based method to convert the MIL problem into a SIL problem. The main contributions of this paper are:

- PLSA based feature representation method is proposed to convert the MIL problem into a standard single instance learning problem, then combining with transductive support vector machine (TSVM), we propose a novel semi-supervised multi-instance learning (SSMIL) algorithm. To the best of our knowledge, this is the first inductive PLSA method for MIL problem, in contrast to other feature representation scheme, PLSA is a proper generative model, and its probabilistic variant has a sound statistical foundation.
- When we use PLSA method to obtain the latent topic models of training bags, the term-document matrix is consisted of all the fuzzy word frequency vector of training bags. It is different form the traditional term-document matrix, in which each document is represented by the word frequency vector. Therefore, it can be viewed as a fuzzy term-document matrix, which will be more appropriate to describe the ambiguity relations between the "*visual-word*" and images.
- To demonstrate the promising performance of our method, we extensively compare our method with other state-of-the-art MIL methods in diverse applications, including drug activity prediction and image retrieval.

In this paper, the latent topic feature of each bag is regarded as its representation vector. After combining with TSVM, a new semi-supervised MIL method named PLSA-SSMIL algorithm is proposed for image retrieval problem. The framework of the proposed algorithm is shown in Fig. 1.

3. PLSA based feature representation scheme for MIL

As a generative model from the statistical text literature, the PLSA [19] was originally developed in the context of text modeling, where words are the elementary parts of documents. Documents are modeled as mixtures of intermediate hidden topics, and topics are characterized by a distribution over words. In the MIL problem, if each bag is regarded as a document, and the "*visual-word*" obtained by k-means clustering method as term, PLSA method can be used to discover training bag's latent topic model, which can obtain every bag's latent topic feature.

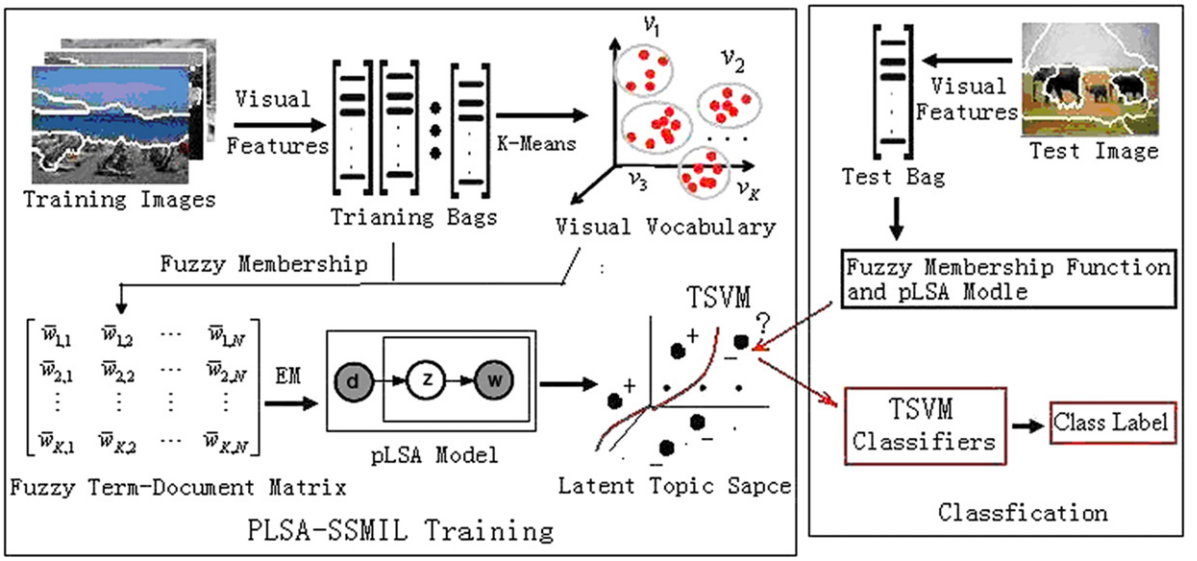

Fig. 1. The framework of the proposed PLSA-SSMIL algorithm.

3.1. Constructing visual vocabulary

To describe the MIL problem, we use the following notation throughout this paper. Let $L = \{(B_1, y_1), \ldots, (B_{|L|}, y_{|L|})\}$ be the labeled bags and let $U = \{B_1, \ldots, B_{|U|}\}$ be the unlabeled bags, and let $L = L^+ \cup L^-$, where L^+ is the positive labeled bags and L^- is the negative labeled bags. The ith positive bag denoted as B_i^+, and the jth instance in that bag denoted as X_{ij}^+, the bag B_i^+ consists of n_i^+ instances $X_{ij}^+ \in R^D$, $j = 1, 2, \ldots, n_i^+$, where R^D denote instance feature space. Similarly, B_i^-, X_{ij}^-, and n_i^- represent the ith negative bag, the jth instance in the bag, and the number of instances in the bag, respectively, the number of positive (negative) bags is denoted as $l^+(l^-)$. We are to learn a classification function that can accurately predict the label of any unseen bag. For the sake of convenience, we line up all the instances in every labeled training bag together, and re-index as:

$$InstSet = \{x_i | i = 1, \ldots, n.\} \tag{1}$$

where $n = \sum_{i=1}^{l^+} n_i^+ + \sum_{i=1}^{l^-} n_i^-$ is the total number of instances within labeled training bags.

When the image regions have the similar visual characteristic, the low-level visual feature vectors of them will group together into a cluster in the feature space. All instances are then grouped into a large number of clusters and those with similar visual characteristic are assigned into the same cluster. We regard the center of each cluster as a "*visual-word*", which represents a specific class of image region and having the same high-level concept, denoted as v_i. Then the set composed of all the "*visual-word*" is called "visual vocabulary", denoted as $\Omega = \{v_1, v_2, \ldots, v_K\}$, where K is the total number of the "*visual-word*".

Because k-means is well-known clustering method, and has been widely used to bags-of-words (BOW) [20] and probabilistic latent semantic analysis (PLSA) based image classification methods [21], so we chose it to group *InstSet* into K clusters, and regard the center of each cluster as a "*visual-word*", which represents a specific local pattern shared by all the features within this cluster.

3.2. Computing fuzzy term-document matrix

The starting point for building a PLSA model is to represent the entire corpus of documents by a term-document co-occurrence matrix of size $K \times N$, where N indicates the number of documents in the corpus and K the number of different words occurring across the corpus. Each matrix entry contains the number of times a specific word (row index) is observed in a given document (column index).

When PLSA is used to obtain the latent topic model of training bags in the MIL problem, we must use a term-frequency vector to represent each bag (document). Let the term-frequency vector of a bag B_j is

$$A_j = [n(v_1, B_j), n(v_2, B_j), \ldots, n(v_K, B_j)]^T \tag{2}$$

where $n(v_i, B_j)$ denotes how often the "*visual-word*" v_i occurred in a bag B_j. When calculate the term-frequency vector of B_j, the traditional method is [11]: if the Euclidean distance between instance x_{jt} and "*visual-word*" v_k is closest, then the k-th

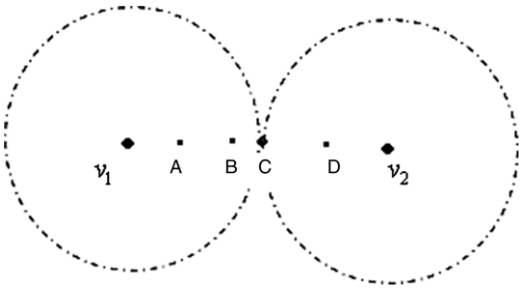

Fig. 2. The fuzzy relation between instances and "visual-word".

component of the term-frequency vector A_j will be added 1, that is

$$n(v_k, B_j) = n(v_k, B_j) + 1. \tag{3}$$

In this paper, the term-frequency vector calculated using Eq. (3) is called as "*traditional term-frequency vector*". When it is used for text analysis, there is not any problem, but when it is used for visual domain (i.e. image retrieval), there is following irrationality, as shown in Fig. 2, let v_1 and v_2 be two different "*visual-word*", A, B, C and D denote four different instances. That can be seen intuitively: the confidence degrees of A and B belong to v_1 are different, because the distances between A, B and v_1 are different, while C should belong to the v_1 or v_2 is ambiguity, as the distances between v_1, v_2 and C are same. When using Eq. (3) to statistics term-frequency vector, these difference and ambiguity are not taken into account. Aiming at these problems, according to the Euclidean distance between instance x and "*visual-word*" v, we define a fuzzy membership function:

$$f(x, v) = \exp\left(-\frac{\|x - v\|^2}{\delta^2}\right) \tag{4}$$

where δ^2 is a predefined scaling factor (we set $\delta^2 = 11$ in the following image retrieval experiments). Then, we regard that each instance within a bag can belong to all of the "*visual-word*", but only according to the distance, its membership degree is different. Therefore, the fuzzy term-frequency vector F_j of a bag B_j is defined as follows:

$$\begin{cases} F_j = [s(v_1, B_j), s(v_2, B_j), \ldots, s(v_K, B_j)]^T \\ s(v_k, B_j) = \sum_{t=1}^{n_j} f(x_{jt}, v_k) = \sum_{t=1}^{n_j} \exp\left(-\frac{\|x_{jt} - v_k\|^2}{\delta^2}\right). \end{cases} \tag{5}$$

Here, $s(v_k, B_j)$ can be interpreted as a measure of similarity between the "*visual-word*" v_k and a bag B_j that is determined by all the instances in this bag. In this paper, the term-frequency vector calculated using Eq. (5) is called as "*fuzzy term-frequency vector*". It has the following two main strongpoint, one is this method more appropriate to describe the ambiguity relations between "*visual-word*" and image when PLSA is used to visual domain, the other is it can overcome the sparse problem that exist in the traditional term-frequency vector, i.e., there are many observation pairs (v_k, B_j) with $n(v_k, B_j) = 0$. Especially in an image retrieval task, as considered in this paper, traditional term-frequency vector is sparse, since an image usually contains a few regions compared with the number of words in the visual vocabulary.

Because term weighting is a key technique in the PLSA method, so we apply popular *tf-idf* [11] term weighting scheme to the fuzzy term-frequency vector F_j, that is [21]

$$w_{k,j} = s(v_k, B_j) \times \log_2(1 + N/df_k) \tag{6}$$

where $s(v_k, B_j)$ is the fuzzy frequency of a "*visual-word*" v_k appears in a bag B_j, N is the total number of bags (images) in the training set, and df_k is the total number of bags having "*visual-word*" v_k. Then the weighted fuzzy term-frequency vector of a bag B_j is

$$W_j = [w_{1,j}, w_{2,j}, \ldots, w_{K,j}]^T. \tag{7}$$

Note that, before use the weighted fuzzy term-frequency vector W_j, we must carry out the following normalization

$$\begin{cases} \bar{W}_j = \left[\bar{w}_{1,j}, \bar{w}_{2,j}, \ldots, \bar{w}_{K,j}\right]^T \\ \bar{w}_{k,j} = (w_{k,j} - \min(W_j))/(\max(W_j) - \min(W_j)), \quad k = 1, 2, \ldots, K. \end{cases} \tag{8}$$

The purpose of normalization is to control the value into [0, 1] interval and let each component of feature vector get the same importance. The normalized weighted fuzzy term-frequency vectors of all training bags constitute the following fuzzy

term-document matrix

$$A_{K\times N} = [\bar{W}_1, \bar{W}_2, \ldots, \bar{W}_N] = \begin{bmatrix} \bar{w}_{1,1} & \bar{w}_{1,2} & \cdots & \bar{w}_{1,N} \\ \bar{w}_{2,1} & \bar{w}_{2,2} & \cdots & \bar{w}_{2,N} \\ \vdots & \vdots & \vdots & \vdots \\ \bar{w}_{K,1} & \bar{w}_{K,2} & \cdots & \bar{w}_{K,N} \end{bmatrix}. \tag{9}$$

Here, K is size of visual vocabulary, N is number of training bags (images), and each bag corresponds to a column vector, each "*visual-word*" corresponds to a row vector.

3.3. PLSA model and EM algorithm

After the fuzzy term-document matrix $A_{K\times N}$ for training bags is generated, we can then apply PLSA method to the MIL problem. Moreover, we also need introduce a latent topic variable $z_t \in Z = \{z_1, z_2, \ldots, z_T\}$ associated with each observation pair (v_k, B_j), here T is the number of topics, which is an import parameter in the PLSA model. The PLSA model is then given by the following generative latent class model [18,21]: (1) select a bag with probability $P(B_j)$, (2) select a topic with probability $P(z_t|B_j)$, and (3) generate a "*visual-word*" with probability, where

$$P(v_k|B_j) = \sum_{t=1}^{T} P(v_k|z_t)P(z_t|B_j). \tag{10}$$

The joint probability between a "*visual-word*" v_k and a bag B_j can be defined by the following mixture:

$$P(v_k, B_j) = P(B_j) \sum_{t=1}^{T} P(v_k|z_t)P(z_t|B_j). \tag{11}$$

Inverting $P(z_t|B_j)$ using the Bayes' rule results in:

$$P(v_k, B_j) = \sum_{t=1}^{T} P(z_t)P(v_k|z_t)P(B_j|z_t) \tag{12}$$

where $P(z_t)$ is the probability of topic variable z_t, and $P(z_t|B_j)$ is the probability of bag B_j occurring in a particular topic z_t. Given that the observation pairs (v_k, B_j) are assumed to be generated independently, we can obtain the following log-likelihood function:

$$LL = \sum_{k=1}^{K}\sum_{j=1}^{N} \bar{w}_{k,j} \log P(v_k, B_j). \tag{13}$$

The parameters $P(z_t)$, $P(v_k|z_t)$, and $P(B_j|z_t)$ can be estimated by maximizing the log-likelihood function using the EM algorithm. The E-step is given by

$$P(z_t|v_k, B_j) = \frac{P(z_t)P(v_k|z_t)P(B_j|z_t)}{\sum_{i=1}^{T} P(z_i)P(v_k|z_i)P(B_j|z_i)}. \tag{14}$$

In addition, the M-step is given by

$$P(z_t) = \frac{\sum_{k=1}^{K}\sum_{j=1}^{N} \bar{w}_{k,j}P(z_t|v_k, B_j)}{\sum_{k=1}^{K}\sum_{j=1}^{N} \bar{w}_{k,j}} \tag{15}$$

$$P(v_k|z_t) = \frac{\sum_{j=1}^{N} \bar{w}_{k,j}P(z_t|v_k, B_j)}{\sum_{k=1}^{K}\sum_{j=1}^{N} \bar{w}_{k,j}P(z_t|v_k, B_j)} \tag{16}$$

$$P(B_j|z_t) = \frac{\sum_{k=1}^{K} \bar{w}_{k,j}P(z_t|v_k, B_j)}{\sum_{k=1}^{K}\sum_{j=1}^{N} \bar{w}_{k,j}P(z_t|v_k, B_j)}. \tag{17}$$

After fitting the PLSA model to the set of training bags $D = \{(B_1, y_1), (B_2, y_2), \ldots, (B_N, y_N)\}$ using the EM algorithm, we can then derive the posterior probability of each training bag to every discovered topic by applying Bayes' rule as:

$$P(z_t|B_j) = \frac{P(B_j|z_t)P(z_t)}{\sum_{i=1}^{T} P(B_j|z_i)P(z_i)}. \tag{18}$$

This can be determined with the estimations in Eqs. (15) and (17). Hence, each training bag is represented using a latent topic vector $\phi(B_j)$:

$$\phi(B_j) = P(Z|B_j) = [P(z_1|B_j), P(z_2|B_j), \ldots, P(z_T|B_j)] \tag{19}$$

which can be considered as a representation of B_j (originally in the K-dimensional "*visual word*" space) in the T-dimensional topic space determined by the estimated $P(z_t|v_k, B_j)$ in Eq. (14).

For an unseen bag B, we first use Eqs. (5)–(8) to calculate its normalized weighted term-frequency vector $\bar{W}_B = \left[\bar{w}_{1,B}, \bar{w}_{2,B}, \ldots, \bar{w}_{K,B}\right]^T$, and then the specific mixing coefficients $P(z_t|B)$ are computed using the fold-in heuristic described in [11], which is achieved by running EM method in a similar manner to that used in the training process. The EM algorithm used to compute $P(z_t|B)$ can be summarized as follows:

$$E\text{-step} : P(z_t|v_k, B) = \frac{P(v_k|z_t)P(z_t|B)}{\sum_{i=1}^{T} P(v_k|z_i)P(z_i|B)} \tag{20}$$

$$M\text{-step} : P(z_t|B) = \frac{\sum_{k=1}^{K} \bar{w}_{k,B} P(z_t|v_k, B)}{\sum_{k=1}^{K}\sum_{j=1}^{N} \bar{w}_{k,B}}. \tag{21}$$

The method proposed in [11] consist in maximizing the likelihood of the bag B with a partial version of the EM algorithm described above, where $P(v_k|z_t)$ is kept fixed (not updated at each M-step). In doing so, $P(z_t|B_j)$ maximizes the likelihood of bag B with respect to the previously trained $P(v_k|z_t)$ parameters. The result is that the test bag B can also be represented by a latent topic vector:

$$\bar{\phi}(B) = P(Z|B) = [P(z_1|B), P(z_2|B), \ldots, P(z_T|B)]. \tag{22}$$

By Eqs. (19) and (22), we can obtain the latent topic features of the labeled and unseen bags respectively to represent the semantic of an image. If a bag is positive, the corresponding latent topic feature is labeled $+1$, if a bag is negative, labeled as -1, then the MIL problem is transformed into a standard SIL problem. In addition, because MIL allows for coarse labeling at the image level, instead of fine labeling at region level, it can significantly improve the efficiency of labeling training samples. As a result, the first two problems described in introduction, which are the representation of image semantic and training samples annotation, are resolved to some extent.

4. PLSA-SSMIL algorithm

In the machine learning, to get a lot of labeled training samples often is difficult, while to gain a large number of unlabeled samples is very easy. For the problem of small sample learning, we use semi-supervised TSVM to train classifiers, which can take advantage of large number of unlabeled images to improve the classification accuracy.

We now assume we have $|L|$ labeled examples $\{\phi(B_i), y_i\}$, where $i = 1, 2, \ldots, |L|$, $y_i \in \{+1, -1\}$, and $|U|$ unlabeled examples $\{\bar{\phi}(B_j) \mid j = 1, 2, \ldots, |U|\}$. Our goal is to construct a TSVM classifier $\text{sign}(w^T \cdot \bar{\phi}(B))$ that utilizes unlabeled data, typically in situations where $|L| \ll |U|$. The following optimization problem is setup for standard TSVM [22]:

$$\begin{cases} \min(w, y'_1, \ldots, y'_{|U|}) : \dfrac{\lambda}{2}\|w\|^2 + \dfrac{1}{2|L|}\displaystyle\sum_{i=1}^{|L|} l(y_i w^T \phi(B_i)) + \dfrac{\lambda_u}{2|U|}\sum_{j=1}^{|U|} l(y'_j w^T \bar{\phi}(B_j)) \\ \text{subject to: } \dfrac{1}{|U|}\displaystyle\sum_{j=1}^{|U|} \max[0, \text{sign}(w^T \bar{\phi}(B_j))] = r \end{cases} \tag{23}$$

where, $|L|$ is the number of labeled samples, $|U|$ is the number of unlabeled samples, λ is regularization parameter, λ_u is a user-provided parameter that provides control over the influence of unlabeled data, if there is enough labeled data, λ, λ_u can be tuned by cross-validation. $l(z)$ is a hinge loss function, usually, $l(z) = l_1(z) = \max(0, 1 - z)$. The labels on the unlabeled data, $y'_1, y'_2, \ldots, y'_{|U|}$, are $\{+1, -1\}$-valued variables in the optimization problem. In other words, TSVM seeks a hyper-plane

w and a labeling of the unlabeled examples, so that the TSVM objective function is minimized, subject to the constraint that a fraction r of the unlabeled data be classified positive. Let w^* is the optimal solution, and bag-level classifier is defined as

$$\text{label}(B) = \text{sign}(w^{*T}\bar{\phi}(B)). \tag{24}$$

Finally, the detailed steps of PLSA-SSMIL algorithm can be summarized as follows

Algorithm: PLSA-SSMIL algorithm

1. *PLSA-SSMIL training.*

Input: A set of labeled bags L, unlabeled bags U, parameter K and T.

Output: PLSA parameters $P(z_t)$, $P(v_k|z_t)$, $P(z_t|B_j)$, and TSVM classifier w^*.

Step 1: Re-index the instances of all the labeled bags together, denoted as *InstSet*, then clustered it into K group using k-means clustering method, and regard every center of the clusters as "*visual-word*"to construct a visual vocabulary $\Omega = \{v_1, v_2, \ldots, v_K\}$.

Step 2: For $\forall B_i \in L$.
Calculate the normalized weighted term-frequency vector $\bar{W}(B_i)$ via Eq. (8), and obtain a fuzzy term-document matrix $A_{K\times N}$ via Eq. (9).

Step 3: Learning PLSA model.
According to Eqs. (14)–(17), we use EM method to fitting the PLSA parameters $P(z_t)$, $P(v_k|z_t)$ and $P(z_t|B_j)$.

Step 4: In latent topic space, we train a TSVM classifier w^* using the latent topic features of all the labeled and unlabeled bag, which are obtained via Eqs. (19) and (22) respectively.

(2) *PLSA-SSMIL predicting.*

Let B be an unseen bag. We first use Eqs. (8) and (22) to calculate its normalized project feature $\bar{W}(B)$ and latent semantic feature $\phi(B)$ respectively, and then use TSVM classifier w^* to predict its label via Eq. (25).

5. Experiments and analysis

5.1. Image data set

To evaluate our method in image retrieval problem, we applied it on the COREL data set, a widely used standard benchmark data set for image retrieval. The data set consists of 2000 images in JPEG format with size 256×384 or 384×256. There are all-together 20 different categories, each containing 100 images. The twenty categories (labeled from 0 to 19) are: Africa people and villages, beaches, buildings, buses, dinosaurs, elephants, flowers, horses, mountains and glaciers, food, dogs, lizards, fashion models, sunset scenes, cars, waterfalls, antique furniture, battle ships, skiing, and desserts. Fig. 3 shows one sample image from each of the 20 categories. The categories are ordered in a row-wise manner from the upper-leftmost image (category 0) to the lower-rightmost image (category 19).

Since we will compare PLSA-SSMIL with MILES and DD-SVM approaches, we adopt the same image segmentation and feature extraction algorithms as described in [14,15]. A brief summary about the imagery features is given as follows. To segment an image, the method first partitions the image into non-overlapping blocks of size 4×4 pixels, and a 6-dimensional feature vector is extracted for each block. Three of them are the average LUV color components in a block. The other three represent square root of energy in the high frequency bands of the wavelet transforms, i.e., the square root of the second order moment of wavelet coefficients in high frequency bands. Because the coefficients of the wavelet transforms in different frequency bands show variations in different directions, hence they can capture the texture properties of a block. Then a modified k-means algorithm is applied to group the feature vectors into clusters, each of which corresponds to a region in the segmented image. After segmentation, three extra features are computed for each region to describe its shape properties, they are normalized inertia of order 1, 2, and 3. As a result, each region in any image is characterized by a 9-dimensional feature vector, which characterizing the color, texture, and shape properties of the region. Descriptions that are more detailed can be found in [14,15]. The original data set is used as two data sets, the first one (i.e. Corel 1k) only used the first ten categorizes images while the second (i.e. Corel 2k) used all the categories images. According to MIL definition, we treat each image as a bag and the 9-dimensional feature vector of each patch in this image as an instance.

5.2. Experimental setup

For each category, we use the "one-versus-the-rest" strategy to evaluate the performance, during each trial, 40 positive images are randomly selected from one category and 40 negative images are randomly selected from the other categories to form the labeled training set, and all the remaining images to form the test set, also regarded as unlabeled training set. In PLSA-SSMIL method, the SVMlin software (available at http://vikas.sindhwani.org/svmlin.html) is used to train TSVM classifiers. Aiming at the problem of training parameter need to be predefined in TSVM, in all subsequent experiments, we set the positive class fraction r of unlabeled image to 0.05, and fix $\lambda = 1$, then chose λ_u from $\{0.001, 0.01, 0.1, 1\}$, a 2-fold cross-validation is conducted on the training set to find the best λ_u.

In addition, in order to verify the "*fuzzy term-frequency vector*" (i.e. Eq. (5)) is effective for PLSA in visual domain, we also use the topic vectors that are trained form the "*traditional term-frequency vector*" (i.e. Eq. (3)) to represent all bags, and combine with TSVM, as a new MIL method, named TPLSA-SSSVM method.

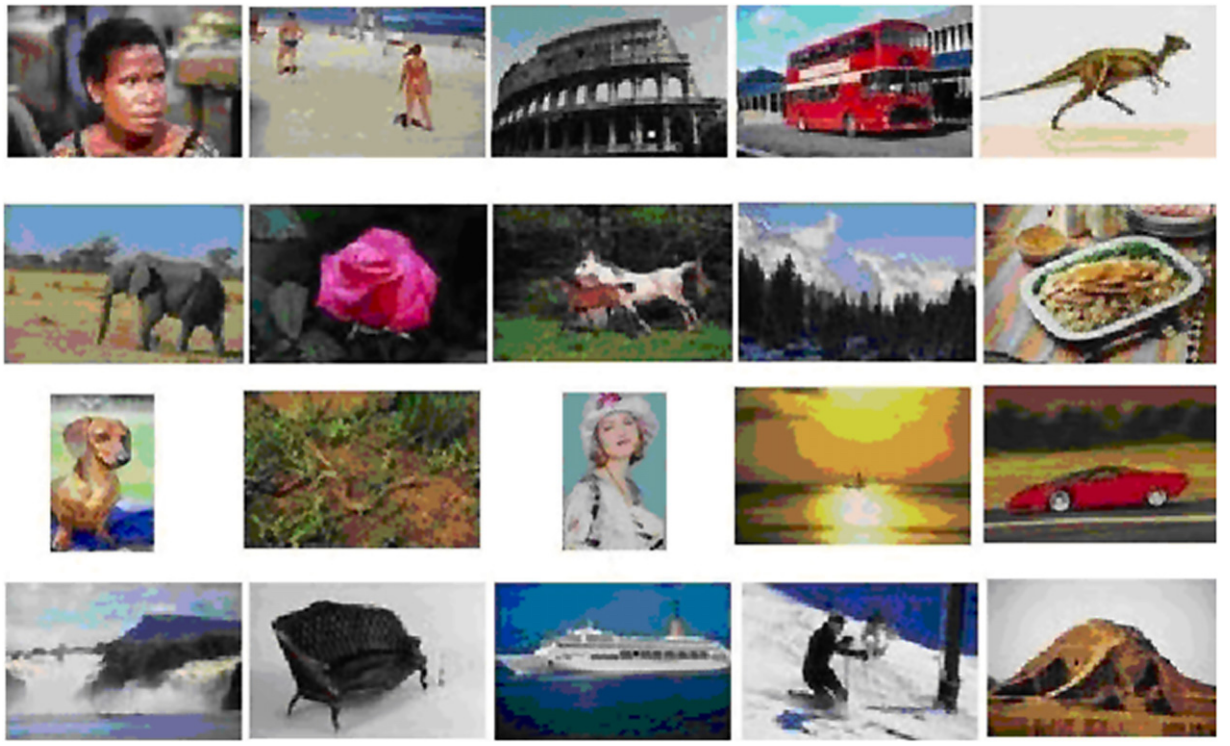

Fig. 3. Sample images from the 20 categories of the COREL image set.

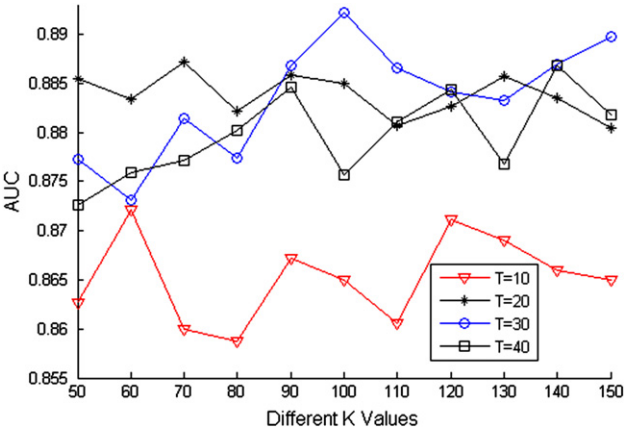

Fig. 4. The retrieval results of different K and T values for PLSA-SSMIL on the 1000 images in Category 0 to Category 9.

5.3. Sensitivity to the number of K and T

While we using k-means clustering and PLSA methods to obtain "*visual-word*" and the latent topic space, two important parameters K and T must to be predefined. In order to confirm the influence of K and T value to retrieval accuracy, we chose K from 50 to 150 with step size 10, and set $T = 10, 20, 30$ and 40 respectively, then test PLSA-SSMIL method using 1000 images in Category 0 to Category 9. Over 10 rounds repeated training and testing, the average AUC (area under the ROC curve) values with 95% confidence interval are shown in Fig. 4.

As seen in Fig. 4, the K values have little effect on the retrieval accuracy, but the T values have bigger influences on the retrieval accuracy. The reason is, the "*fuzzy term-frequency vector*" can help to overcome the sparse problem that exists in the "*traditional term-frequency vector*", and so the sizes of visual vocabulary K have little effect to the PLSA-SVM method. However, because the optimal number of latent topic is almost equal for a given data set, when T is too small, it will lose useful classification information and affect classification accuracy, when T is too big, there is more redundant noises topic in the latent topic feature and cannot help to improve retrieval accuracy.

Table 1
Average AUC values (in percent) with 95% confidence interval over 20 rounds of test on the COREL image set.

Image Category	PLSA-SSMIL	TPLSA-SSMIL	MILES [15]	DD-SVM [14]	EC-SVM [3]	GMIL [23]	MISSL [24]
Dinosaurs	99.1 ± 0.4	94.9 ± 1.0	97.6 ± 0.7	96.8 ± 0.9	97.8 ± 0.7	94.6 ± 0.6	97.7 ± 0.3
Flowers	97.2 ± 0.6	95.0 ± 1.0	86.5 ± 1.9	94.1 ± 1.0	95.1 ± 1.0	94.0 ± 0.6	88.9 ± 0.7
Horses	96.3 ± 0.8	87.4 ± 1.4	94.1 ± 1.2	86.9 ± 3.0	87.3 ± 2.1	93.1 ± 0.8	90.5 ± 1.1
Buses	94.2 ± 1.0	91.7 ± 1.1	94.1 ± 1.1	95.1 ± 1.2	93.2 ± 1.0	94.1 ± 0.6	93.3 ± 0.9
Antique furniture	93.6 ± 1.1	85.1 ± 1.1	85.3 ± 1.3	85.6 ± 2.1	85.5 ± 1.1	85.3 ± 0,8	78.2 ± 1.6
Food	92.1 ± 1.1	89.3 ± 1.5	89.5 ± 1.5	91.8 ± 1.3	91.3 ± 1.5	89.5 ± 0.8	93.9 ± 0.9
Sunset scenes	91.2 ± 1.1	85.8 ± 1.1	88.2 ± 1.5	89.4 ± 3.1	87.6 ± 2.1	88.2 ± 1.2	80.4 ± 3.5
Elephants	90.1 ± 1.2	85.2 ± 1.2	86.0 ± 1.3	83.4 ± 2.3	88.1 ± 2.0	85.0 ± 1.3	84.5 ± 0.8
Waterfalls	88.5 ± 1.3	82.5 ± 1.5	89.0 ± 1.2	88.4 ± 2.8	86.5 ± 1.8	81.2 ± 1.5	82.1 ± 2.8
Fashion models	88.2 ± 1.5	73.9 ± 2.3	83.4 ± 1.7	77.7 ± 2.8	76.6 ± 2.5	90.0 ± 2.1	88.4 ± 2.8
Cars	86.1 ± 1.4	80.3 ± 2.0	85.8 ± 1.6	78.8 ± 3.5	83.2 ± 3.2	80.4 ± 1.7	80.7 ± 2.0
Skiing	85.3 ± 1.6	73.1 ± 2.3	87.7 ± 1.7	76.1 ± 3.9	76.5 ± 3.1	84.8 ± 1.6	68.0 ± 5.2
Beach	84.6 ± 1.5	81.3 ± 2.5	84.0 ± 1.5	78.4 ± 3.0	84.2 ± 2.1	83.7 ± 1.7	83.4 ± 2.7
Battle ships	84.4 ± 1.6	76.1 ± 2.2	79.4 ± 1.3	74.0 ± 2.3	79.6 ± 3.3	81.0 ± 1.5	69.6 ± 2.5
Lizards	84.0 ± 1.7	86.3 ± 2.1	81.1 ± 1.4	80.4 ± 2.2	84.3 ± 2.0	79.4 ± 1.3	77.3 ± 4.3
Dogs	83.5 ± 1.6	81.1 ± 1.7	85.2 ± 1.3	85.2 ± 1.7	85.6 ± 1.4	81.1 ± 1.4	73.8 ± 3.4
Africa	80.8 ± 1.8	77.5 ± 2.2	79.6 ± 1.3	73.2 ± 2.8	81.8 ± 2.2	79.6 ± 1.3	76.8 ± 5.2
Building	80.1 ± 2.0	83.6 ± 2.2	77.0 ± 1.7	73.2 ± 2.5	83.1 ± 2.1	77.0 ± 1.7	70.2 ± 2.9
Mountains	76.3 ± 2.1	72.6 ± 3.2	78.1 ± 1.7	74.0 ± 3.1	75.4 ± 2.1	76.4 ± 1.2	62.4 ± 4.3
Desserts	75.3 ± 2.3	73.4 ± 2.9	76.4 ± 1.2	66.4 ± 3.4	75.8 ± 2.6	69.5 ± 1.6	51.6 ± 2.6
Average	**87.55**	**82.80**	**85.4**	**82.45**	**84.93**	**84.39**	**79.58**

5.4. Overall retrieval results

Considering the fact that MILES [15] is the state-of-the-art method on the COREL image data set, we choose MILES as the baseline for comparison. As in [3,15], we choose λ from 0.1 to 0.6 with step size 0.05 and δ^2 from 5 to 15 with step size 1. We find that $\lambda = 0.2$ and $\delta^2 = 11$ give the best test performance for MILES on the COREL data set. Hence, we fix $\lambda = 0.2$ and $\delta^2 = 11$ for MILES [3,15] in all the following experiments. For the parameters in PLSA-SSMIL method, experiments found that when $K = 100, T = 30$, it can give the best retrieval results. Therefore, we fix these parameters in all the comparative experiments. Based on the same training and testing sets, a comparison of PLSA-SSMIL with other key existing MIL algorithms was listed in Table 1, which includes MILES [15], DD-SVM [14], EC-SVM [3], GMIL [23] and MISSL [24], etc. The numbers listed are the average AUC values for all 20 categories with 95% confidence interval over 20 rounds of training and test.

It can be seen from Table 1, among all the images in the 20 categories, PLSA-SSMIL achieves the best overall retrieval performance. Similar to the problems of synonymy and polysemy in the textual retrieval, the issue of "semantic gap" is the biggest hurdle in image retrieval. The existence of semantic gap means that although visual features are similar, their high-level semantic may be different, as well as that although the high-level semantic is the same, the corresponding visual features may be completely different. For example, blue sky and blue sea, their visual features are similar, but their semantics are different, while blue sky and cloudy sky, their visual features are dissimilar, but their semantics are the same. Because PLSA based feature representation scheme not only can reduces the dimensionality of the original projection feature, but also can captures the important underlying semantic structure in the association of terms (*visual-words*) and documents (images), so PLSA is able to overcome the problem of synonymy and polysemy. As a result, the latent topic feature extracted by Eq. (19) can better represent the bag in MIL problem.

Another also can be seen from Table 1, the average AUC values of PLSA-SSMIL and TPLSA-SSMIL algorithms are 88.13 and 82.80 respectively, increased by 5.33%. It shows that the fuzzy term-frequency vector is effective, and it is more suitable to describe the ambiguity relationship between "*visual-word*" and image in the visual domain.

In order to verify the validity of semi-supervised learning in the MIL problem, we also compared the PLSA-SSMIL with MILES [14] that is a supervised MIL method on the "Corel 1k" data set. With the number $|L|$ of labeled images increase, the average precision curves of 10 randomized repeated experiments are shown in Fig. 5(A); And then $|L|$ is fixed at 80, with the number $|U|$ of unlabeled images increase, the average precision curves of 10 randomized repeated experiments are shown in Fig. 5(B). From Fig. 5, we can see the performance of PLSA-SSMIL always outperforms MILES. This experiment indicates that using the unlabeled image to help training classifiers can improve the retrieval accuracy surly.

5.5. Speed

Under the assumption that both image segmentation and feature extraction work have been done previously, Table 2 lists only the training and test time (in Matlab 7.01 on a AMD 4200+PC running the windows XP operating system with 1G memory) required by PLSA-SSMIL DD-SVM, MILES, and DD-SVM methods. "Training" refers to the average time for training a binary classifier for all the 20 categories when 40 positive and 40 negative images are used as the training set. "Test" refers to the average time to finish a testing when all the remaining 1920 images are used as test set.

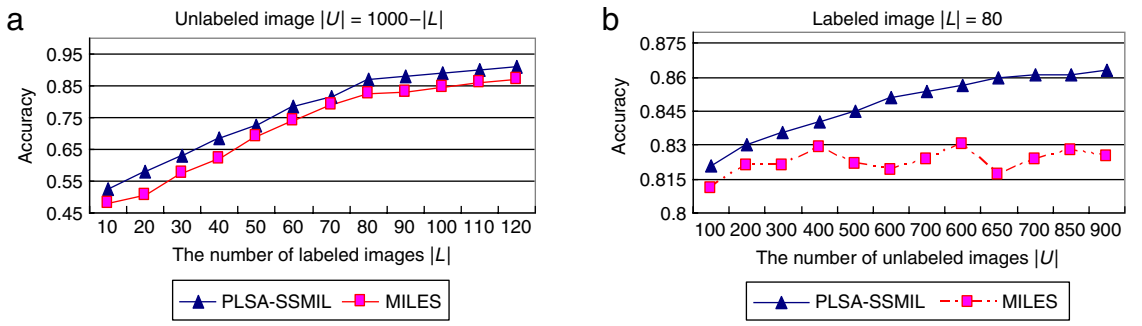

Fig. 5. Comparison of PLSA-SSMIL and MILES on the Corel 1k data set for image retrieval. (A) Comparison of the accuracy of two algorithms with different numbers of labeled images $|L|$. (B) Comparison of the accuracy of two algorithms with different numbers of unlabeled images $|U|$.

Table 2
Training and test time comparison (in minutes).

MIL methods	PLSA-SSMIL	DD-SVM	MILES	MI-SVM
Training	0.15	2.6	0.31	1.03
Test	0.08	1.01	0.12	0.31

For PLSA-SSMIL, the training time is mainly spent in three aspects: (1) generate a collection of "*visual-word*" to construct a visual vocabulary by K-means clustering method; (2) using EM to fitting bag's latent topic models; (3) TSVM training. The training set consists of 80 images. There are around 340 different instances in all bags. When we set $K = 100$ and $T = 30$, k-means clustering spends around 1.06 s to extract "*visual-word*". Then it uses the method in Section 3.3 to obtain training bags' latent topic model in around 5.82 s (Matlab code), and train a binary TSVM classifier in around 2.11 s (C code). Because DD-SVM method must learn a collection of instance prototypes based on DD function to construct a new feature space by quasi-Newton search algorithm with every instance in every positive bag as starting points, it incur significantly higher computation cost than PLSA-SSMIL. MILES method uses all the instances from the training bags instead of the prototypes used with DD-SVM to construct a new feature space, so the projection space for representing bags is of very high dimensionality. This not only increases the computation cost of the projection feature of every bag, but also increases the training cost of SVM.

The test efficiency mainly depends on the dimension of the feature space. Because the PLSA based feature only has 30 dimensions and MILES uses 1-norm SVM to select a few important instances to construct the feature space during its training, it decreases the original feature space dimension greatly. However, DD-SVM is less efficient since the DD function is a continuous and highly nonlinear function with multiple peaks and valleys (or local maximums and minimums), so the quantity of "prototypes" searched by quasi-Newton algorithm is large. Therefore, the dimensionality of its feature space is also the highest among the three methods we compared, which results in the least efficiency. Table 2 illustrates that PLSA-SSMIL method is more efficient than other two MIL methods.

6. Conclusions and future work

Focusing on the three major problems of semantic image retrieval under the framework of supervised learning and based on PLSA and TSVM, a novel SSMIL algorithm, named PLSA-SSMIL, has been presented in this paper. First, we use the nature of training bag in the MIL, the labels are assigning to image rather then region, so it can greatly improve the efficiency of hand-labeled the training samples. Second, because PLSA based feature representation scheme cannot only reduces the dimensionality of the original term-document vector, but also captures the important underlying semantic in the images, so the latent topic feature extracted by PLSA method can better represent the semantic of bag in image retrieval problem. Finally, the semi-supervised TSVM is used to train the classifier, which can take advantage of a large number of unlabeled images to improve the classifier performance. Hence, the small sample learning problems can be resolved also. Therefore, for the data sets used in our experiments, the proposed approach achieved retrieval accuracies comparable to the state-of-the-art MIL algorithms while being much more efficient.

By obtaining the latent topic features of bags (images), PLSA could remove most noise presented in original term-document matrix. When the number of topic is too small, important information may be lost. On the other hand, if the number of topic is too high, it may fail to remove sufficient amount of noises. What is the optimal number of topic is the future work we must to do. Furthermore, in CBIR, it is easy to get a large number of unlabeled images from the image repository. Hence, we will explore other semi-supervised learning methods in our future research.

Acknowledgments

This research is fully supported by the Youth (1090428) & Doctoral Research (1091216) Foundation of the Xi'an University of Posts and Telecommunications.

References

[1] Ying Liu, Dengsheng Zhang, Guojun Lu, Wei-Ying Ma, A survey of content-based image retrieval with high-level semantics, Pattern Recognition 40 (1) (2007) 262–282.
[2] Qingyong Li, Siwei Luo, Zhongzhi Shi, Fuzzy aesthetic semantics description and extraction for art image retrieval, Computers & Mathematics with Applications 57 (6) (2009) 1000–1009.
[3] Wu-Jun Li, Dit-Yan Yeung, Localized content-based image retrieval through evidence region identification, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, June, 2009, pp. 1666–1673.
[4] T.G. Dietterich, R.H. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artificial Intelligence 89 (12) (1997) 31–71.
[5] Jun Wang, Jean-Daniel Zucker, Solving the multiple-instance problem: a lazy learning approach, in: Proceedings of the 17th International Conference on Machine Learning, San Francisco, CA, 2000, pp. 1119–1125.
[6] O. Maron, A.L. Ratan, Multiple-instance learning for natural scene classification, in: Proceedings of the 15th International Conference on Machine Learning, Madison, WI, 1998, pp. 341–349.
[7] Q. Zhang, S. Goldman, Em-dd: an improved multipleinstance learning technique, in: Advances in Neural Information Processing Systems, 2002, pp. 1073–1080.
[8] Min-Ling Zhang, Zhi-Hua Zhou, A multi-instance regression algorithm based on neural network, Journal of Software 14 (7) (2003) 1238–1242.
[9] S. Andrews, T. Hofmann, I. Tsochantaridis, Multiple instance learning with generalized support vector machines, in: Proceedings of the 18th National Conference on Artificial Intelligence, Edmonton, Canada, 2002, pp. 943–944.
[10] P.V. Gehler, O. Chapelle, Deterministic annealing for multiple-instance learning, in: Proceedings of the 11th International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March, 2007, pp. 123–130.
[11] T. Gartner, P.A. Flach, A. Kowalczyk, A.J. Smola, Multi-instance kernels, in: Proceedings of the 19th International Conference on Machine Learning, Sydney, Australia, 2002, pp. 179–186.
[12] J.T. Kwok, P.-M. Cheung, Marginalized multi-instance kernels, in: Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hydrabad, India, 2007, pp. 901–906.
[13] Z.-H. Zhou, Y.-Y. Sun, Y.-F. Li, Multi-instance learning by treating instances as non-i.i.d. samples, in: Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Quebec, Canada, 2009, pp. 1249–1256.
[14] Y.X. Chen, J.Z. Wang, Image categorization by learning and reasoning with regions, Journal of Machine Learning Research 5 (8) (2004) 913–939.
[15] Yi-xin Chen, Jin-bo Bi, James Z. Wang, MILES: multiple-instance learning via embedded instance selection, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (12) (2006) 1931–1947.
[16] Z. Zhou, M. Zhang, Solving multi-instance problems with classifier ensemble based on constructive clustering, Knowledge and Information Systems 11 (2) (2007) 155–170.
[17] Yasser EL-Manzalawy, Vasant Honavar, MICCLLR: Multiple-Instance Learning using Class Conditional Log Likelihood Ratio, in: Discovery Science, vol. 5808, 2009, pp. 80–91.
[18] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Machine Learning 41 (2001) 177–196.
[19] G. Salton, C. Buckley, Automatic structuring and retrieval of large text files, Communications of the ACM 32 (2) (1994) 97–107.
[20] L. Fei-Fei, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR, Pasadena, CA, USA, June, 2005, pp. 524–531.
[21] Zhi-wu Lu, Yu-xin Peng, Horace H.S. Ip, Image categorization via robust pLSA, Pattern Recognition Letters 31 (1) (2010) 36–43.
[22] Vikas Sindhwani, S. Sathiya Keerthi, Large scale semi-supervised linear SVMs, in: Proceedings of the 29th annual international ACM SIGIR conference on Research and Development in Information Retrieval, Seattle, Washington, USA, 2006, pp. 477–484.
[23] Changhu Wang, Lei Zhang, Hong-Jiang Zhang, Graph-based multiple-instance learning for object-based image retrieval, in: Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval, Vancouver, British Columbia, Canada, October, 2008, pp. 156–163.
[24] Rouhollah Rahmani, Sally A. Goldman, MISSL: multiple-instance semi-supervised learning, in: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, June, 2006, pp. 705–712.