Information Sciences 199 (2012) 20–30

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Multi-view learning via probabilistic latent semantic analysis

Fuzhen Zhuang [a,b,*], George Karypis [c], Xia Ning [c], Qing He [a], Zhongzhi Shi [a]

[a] The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, China
[b] Graduate University of Chinese Academy of Sciences, China
[c] Department of Computer Science & Engineering, University of Minnesota, Twin Cities, United States

ARTICLE INFO

Article history:
Received 28 April 2011
Received in revised form 4 February 2012
Accepted 27 February 2012
Available online 6 March 2012

Keywords:
Multi-view learning
Generative model
Probabilistic Latent Semantic Analysis (PLSA)

ABSTRACT

Multi-view learning arouses vast amount of interest in the past decades with numerous real-world applications in web page analysis, bioinformatics, image processing and so on. Unlike the most previous works following the idea of co-training, in this paper we propose a new generative model for Multi-view Learning via Probabilistic Latent Semantic Analysis, called MVPLSA. In this model, we jointly model the co-occurrences of features and documents from different views. Specifically, in the model there are two *latent* variables y for the latent topic and z for the document cluster, and three *visible* variables d for the document, f for the feature, and v for the view label. The conditional probability $p(z|d)$, which is independent of v, is used as the bridge to share knowledge among multiple views. Also, we have $p(y|z, v)$ and $p(f|y, v)$, which are dependent of v, to capture the specifical structures inside each view. Experiments are conducted on four real-world data sets to demonstrate the effectiveness and superiority of our model.

1. Introduction

In general, the classification or clustering algorithms in machine learning focus on the scenarios that data are represented only by one single view [16,21,23,22,28,10,24]. However, there are numerous real-world applications in web page analysis, bioinformatics and image processing where data naturally have several different representations. A typical example is web page classification, in which the web page not only can be represented as a term vector by the words occurring in the web page, but also can be represented as a vector of link features exploited from hyperlink structure. We extract from an image different kinds of features, i.e., color, texture and shape. We can represent images by multiple views if one kind of features is regarded as one view. Moreover, we usually cannot merge all the features from different views together since different representations have their own intrinsic structure, i.e., people often want to exploit the information hidden in the hyperlink structure. Thus, the multi-view algorithms [3,20,2,29,27,17,13,11,12,25] are needed to handle data with multiple different representations.

The earliest works of multi-view learning are introduced by Blum and Mitchell [3] and Yarowsky [26]. Yarowsky [26] proposed an algorithm for word sense disambiguation, in which he exploited two classifiers, one from the nearby words (one view) and one from the co-occurrence words in the same document (another view), in an iterative bootstrapping manner to improve the accuracy. Blum and Mitchell [3] initiated the idea of co-training for semi-supervised classification. They first trained classifiers from each single view, and then iteratively let each classifier label unlabeled instances it predicts with the highest confidence. In that method, newly labeled example by one classifier may provide useful information for the other

* Corresponding author at: The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, China. Fax: +8610 82610254.
E-mail addresses: zhuangfz@ics.ict.ac.cn (F. Zhuang), karypis@cs.umn.edu (G. Karypis), heq@ics.ict.ac.cn (Q. He), shizz@ics.ict.ac.cn (Z. Shi).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2012.02.058

Table 1
The notation of variables.

d	The visible variable for document
f	The visible variable for feature
v	The visible variable for view label
y	The latent variable for latent topic
z	The latent variable for document cluster

classifier under the independent assumption between classifiers. There are several works following the initial idea of co-training [19,4,18,17]. Nigam and Ghani [19] developed a Co-EM algorithm, which can avoid tuning the parameter – the number of samples added for each iteration. Then Muslea and Ghani [17] combined the Co-Test algorithm [18] and Co-EM to address the application scenarios with incompatible, correlated views. However, the idea of co-training requires the conditional independent assumption to work well, co-training may also fail when the newly labeled samples are not reliable any more. Moreover, the idea co-training is designed to utilize only two views, thus being unable to exploit more views to improve the learning performance.

Long et al. [15] developed a general model for multi-view unsupervised learning, which is essentially an ensemble method. They find a best clustering pattern from the clustering results output by base clustering algorithms. In this paper, we study the multi-view learning problem via Probabilistic Latent Semantic Analysis (PLSA) [7,8], which is validated to model co-occurrence data very well on text data to find high-level latent topics.[1] Our algorithm is motivated by the following two observations. First, different features in the context may be grouped together to indicate a high-level concept, i.e., the words "price", "performance" "announcement" from an enterprise news may present the concept "produce announcement"; second, the methods working on only one single view may not perform well, since they cannot make full use of the knowledge from different views. Thus, we propose a generative model to jointly model the co-occurrences of features and documents from different views for multi-view learning. Specifically, in the model there are two *latent* variables y for the latent topic and z for the document cluster, and three *visible* variables d for the document, f for the feature, and v for the view label. The conditional probability $p(z|d)$, which is independent of v, is used as the bridge to share knowledge among multiple views. Also, we have $p(y|z, v)$ and $p(f|y, v)$, which are dependent of v, to capture the specifical structures inside each view. Experimental results in Section 4 show the effectiveness of our algorithm, and additional gains over compared methods working on only one single view. For the sake of clarity, we summarize the notation of variables in Table 1.

The formulation of this work is similar to the one in [30]. Zhuang et al. adapted PLSA model to cross-domain learning, in which the joint probability $p(y, z)$ is shared among multiple domains for knowledge transfer, while our model uses the same conditional probability $p(z|d)$ among multiple views to share knowledge. They also grouped the features from different data domains into the same latent topic space, which cannot deal with the multi-view learning problem directly, while our algorithm groups the features from different views into different latent topic spaces.

The rest of this paper is organized as follows: Section 2 describes the preliminary knowledge and problem formulation, followed by the EM solution in Section 3. In Section 4 we give the comprehensive experiments to evaluate our model. Finally, Section 5 concludes the paper.

2. Preliminary knowledge and problem formulation

2.1. Probabilistic latent semantic analysis

Probabilistic Latent Semantic Analysis (PLSA) [7,8] is a statistical model to analyze co-occurrence data by a mixture decomposition. Specifically, given the word-document co-occurrence matrix $\boldsymbol{O}$, element $O_{f,d}$ represents the frequency of word f appearing in document d. Here we use the more general notation f, instead of w, to represent feature in the sense that not all views use words as their features. PLSA models $\boldsymbol{O}$ by using a mixture model with latent topic (each topic is denoted by y) as follows,

$$p(f,d) = \sum_y p(f,d,y) = \sum_y p(f|y)p(y|d)p(d). \tag{1}$$

Fig. 1a shows the graphical model for PLSA. The parameters of $p(f|y)$, $p(y|d)$, $p(d)$ over all f, d, y can be obtained by the EM solution to the maximum likelihood problem.

In the PLSA model, the documents and words share the same latent variable y. However, documents and words usually exhibit different organizations and structures. Jiho and Choi [9] proposed a Dual-PLSA (DPLSA) model for clustering. Specifically, the model may have different kinds of latent topics, denoted by y for word topic and z for document class. Its graphical model is shown in Fig. 1b.

[1] The words "latent topic" and "concept" are used interchangeably in this paper.

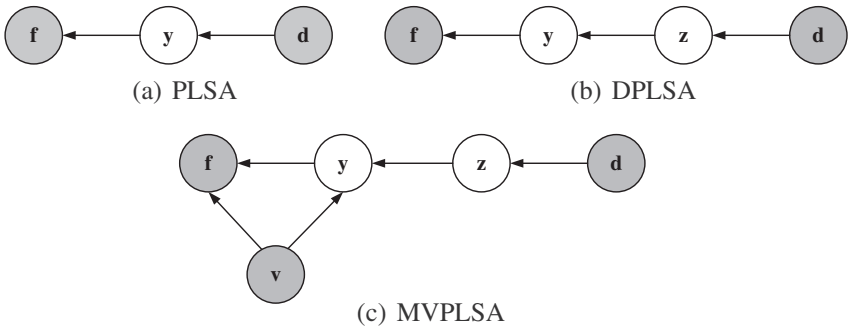


Fig. 1. The graphical models of PLSA, DPLSA and MVPLSA.

Given the word-document co-occurrence matrix $\boldsymbol{O}$, we can similarly arise a mixture model like Eq. (1),

$$p(f,d)=\sum_{y,z}p(f,d,y,z)=\sum_{y,z}p(f|y)p(y|z)p(z|d)p(d). \tag{2}$$

And the parameters of $p(f|y)$, $p(y|z)$, $p(z|d)$, $p(d)$ over all f, d, y, z can also be obtained by the EM solution.

2.2. Multi-view learning via PLSA

In this paper, we study the multi-view learning problem via probabilistic latent semantic analysis. As the graphical model shown in Fig. 1c, f and d are the variables for feature and document, and y, z are respectively the latent variables for feature topic and document class, while v is the variable for data view label. We have m views for the data, $1 \leqslant v \leqslant m$, and jointly model the co-occurrences to approximate all the parameters over all views of data. The variables f and y are dependent on the label of data view v, thus the features in each view can flexibly have different latent topic spaces. In this model, the data instances from all views share the same conditional probabilities $p(z|d)$, which can facilitate sharing knowledge among all views. Since our model is based on probabilistic latent semantic analysis for multi-view learning, thus we call it MVPLSA for short.

The details of generative process are as follows,

(1) choose a document $d \sim p(d)$, and a view $v \sim p(v)$;
(2) choose a document class $z \sim p(z|d)$;
(3) for each feature f in document d
  (3.1) choose a latent topic $y \sim p(y|z, v)$;
  (3.2) choose a feature $f \sim p(f|y, v)$;

(4) end.

Note that we repeat the steps (3.1) and (3.2) $|d|$ times, where $|d|$ is the number of features in document d. Obviously, the variables f and y are dominated by the data view label v. It is different from DPLSA (we can simply concatenate all the features together from all views and then perform DPLSA), which assumes all the features from all views are conditionally independent with each other when given the topic, e.g., $p(f_i, f_j|y) = p(f_i|y)p(f_j|y)$. In MVPLSA we relax this assumption by incorporating the variable v for data view, e.g., $p(f_i, f_j|y, v) = p(f_i|y, v)p(f_j|y, v)$ where f_i and f_j are from the same v-view. The experimental results on *20Newsgroups* show that MVPLSA model can benefit from this relaxation.

Supposed the multi-view data $\boldsymbol{X}$ have m representations from different feature spaces, $\boldsymbol{X} = \{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(m)}\}$, and the co-occurrences matrixes $\boldsymbol{O} = \{\boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \ldots, \boldsymbol{O}^{(m)}\}$, where $O_{f,\,d,\,v}$ $(1 \leqslant v \leqslant m)$ represents the co-occurrence of feature f and document d in the vth view. Then from the graphical model in Fig. 1c, we can write the mixture model with latent variables as follows,

$$p(f,d,v)=\sum_{y,z}p(f,d,y,z,v)=\sum_{y,z}p(f|y,v)p(y|z,v)p(z|d)p(d)p(v). \tag{3}$$

If we denote all the latent variables y, z as $\boldsymbol{Z}$, given the whole data $\boldsymbol{X}$ from different views we formulate the problem of maximum log likelihood as

$$\log\ p(\boldsymbol{X};\theta)=\log\ \sum_{\boldsymbol{Z}}p(\boldsymbol{Z},\boldsymbol{X};\theta), \tag{4}$$

where θ includes all the parameters of $p(f|y, v)$, $p(y|z, v)$ $(1 \leqslant v \leqslant m)$, $p(z|d)$ and $p(d)$. In the next section we derive the EM algorithm to solve this maximum log likelihood problem.

3. EM solution to MVPLSA

We rewrite Eq. (4) as follows,

$$\mathcal{L}_0 = \log\ p(\boldsymbol{X};\theta) = \log\ \sum_{\boldsymbol{Z}} p(\boldsymbol{Z},\boldsymbol{X};\theta) = \log\ \sum_{\boldsymbol{Z}} p(\boldsymbol{Z},\boldsymbol{X};\theta)\frac{p(\boldsymbol{Z}|\boldsymbol{X};\theta_{old})}{p(\boldsymbol{Z}|\boldsymbol{X};\theta_{old})} = \log\ \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X};\theta_{old})\frac{p(\boldsymbol{Z},\boldsymbol{X};\theta)}{p(\boldsymbol{Z}|\boldsymbol{X};\theta_{old})}. \tag{5}$$

According to Jensen's inequality:

$$\mathcal{L}_0 \geqslant \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X};\theta_{old}) \log\frac{p(\boldsymbol{Z},\boldsymbol{X};\theta)}{p(\boldsymbol{Z}|\boldsymbol{X};\theta_{old})} = \underbrace{\sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X};\theta_{old}) \log p(\boldsymbol{Z},\boldsymbol{X};\theta)}_{\mathcal{L}} - \underbrace{\sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X};\theta_{old}) \log p(\boldsymbol{Z}|\boldsymbol{X};\theta_{old})}_{\text{const}} = \mathcal{L} + \text{const}. \tag{6}$$

Omitting the constant in the second item, an EM algorithm is to maximize the lower bound $\mathcal{L}$. Now we write the observed data $\boldsymbol{X}$ and latent variable $\boldsymbol{Z}$ in details,

$$\mathcal{L} = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X};\theta_{old})\ \log\ p(\boldsymbol{Z},\boldsymbol{X};\theta) = \sum_{f,d,v} O_{f,d,v} \sum_{y,z} p(y,z|f,d,v;\theta_{old})\ \log\ p(f,d,y,z,v;\theta), \tag{7}$$

where $O_{f,d,v}$ is the co-occurrent number of f and d in the v-view, then we find θ_{new} according to

$$\theta_{new} = \arg\max_{\theta}\ \mathcal{L}. \tag{8}$$

We also omit the update formulas for $p(d)$, $p(v)$ here since they are independent of θ, thus the developed EM algorithm to obtain all the parameters is,

E-step

$$p(y,z|f,d,v;\theta_{old}) = \frac{p(f,d,y,z,v;\theta_{old})}{\sum_{y,z} p(f,d,y,z,v;\theta_{old})}. \tag{9}$$

The joint probability $p(f, d, y, z, v; \theta_{old})$ $(1 \leqslant v \leqslant m)$ can be computed according to Eq. (3).

M-step

For the parameter $p(f|y, v)$, we maximize $\mathcal{L}$ with its parameters by Lagrangian Multiplier method and extract the terms containing $p(f|y, v)$. Then, we have

$$\mathcal{L}_{[p(f|y,v)]} = \sum_{y,z,f,d,v} O_{f,d,v} p(y,z|f,d,v;\theta_{old}) \cdot \log\ p(f|y,v). \tag{10}$$

Applying the constraint $\sum_f p(f|y, v) = 1$ into the following equation:

$$\frac{\partial[\mathcal{L}_{[p(f|y,v)]} + \lambda(1 - \sum_f p(f|y,v))]}{\partial p(f|y,v)} = 0, \tag{11}$$

then

$$p(f|y,v) = \frac{\sum_{z,d} O_{f,d,v}\ p(y,z|f,d,v;\theta_{old})}{\lambda}. \tag{12}$$

Considering the constraint $\sum_f p(f|y, v) = 1$,

$$1 = \sum_f p(f|y,v) = \frac{\sum_f \sum_{z,d} O_{f,d,v}\ p(y,z|f,d,v;\theta_{old})}{\lambda}, \tag{13}$$

the value of λ can be computed:

$$\lambda = \sum_f \sum_{z,d} O_{f,d,v}\ p(y,z|f,d,v;\theta_{old}). \tag{14}$$

Finally, the update formula of $p(f|y, v)$ can be obtained,

$$p(f|y,v) = \frac{\sum_{d,z} O_{f,d,v}\ p(y,z|f,d,v;\theta_{old})}{\sum_{f,d,z} O_{f,d,v}\ p(y,z|f,d,v;\theta_{old})}, \tag{15}$$

Similarly,

$$p(y|z,v) = \frac{\sum_{f,d} O_{f,d,v}\ p(y,z|f,d,v;\theta_{old})}{\sum_{f,d,y} O_{f,d,v}\ p(y,z|f,d,v;\theta_{old})}, \tag{16}$$

$$p(z|d) = \frac{\sum_v \sum_{f,y} O_{f,d,v}\ p(y,z|f,d,v;\theta_{old})}{\sum_v \sum_{f,y,z} O_{f,d,v}\ p(y,z|f,d,v;\theta_{old})}. \tag{17}$$

Algorithm 1. Multi-view Learning via Probabilistic Latent Semantic Analysis (MVPLSA)

Input: Given the data with views $X = \{X^{(1)}, X^{(2)}, \ldots, X^{(m)}\}$; the number of latent topics for each view $Y^{(1)}, \ldots, Y^{(m)}$; T, the number of iterations

Output: the classification results on unlabeled data, or the clustering results if there is not any labeled data

1: Initialization. $p^{(0)}(f|y, v)$ and $p^{(0)}(y|z, v)$ $(1 \leqslant v \leqslant m)$ are set randomly, the initialization of $p^{(0)}(z|d)$ associated with unlabeled data is detailed in Section 4.1
2: $t := 1$
3: **for** $v := 1$ to m
  Update $p^{(t)}(y, z|f, d, v)$ according to Eq. (9) in **E-step**
4: **end**
5: **for** $v := 1$ to m
  Update $p^{(t)}(f|y, v)$ according to Eq. (15) in **M-step**
  Update $p^{(t)}(y|z, v)$ according to Eq. (16) in **M-step**
6: **end**
7: Update $p^{(t)}(z|d)$ according to Eq. (17) in **M-step**
8: $t := t + 1$
9: **If** $t < T$, turn to Step 3
10: Obtain the final probabilities $p^{(t)}(f|y, v)$, $p^{(t)}(y|z, v)$ $(1 \leqslant v \leqslant m)$ and $p^{(t)}(z|d)$
11: Conduct the classification or clustering according to the output conditional probability $p^{(t)}(z|d)$ and Eq. (18)

3.1. Semi-supervised classification

In this subsection, we show how to adapt our model to multi-view semi-supervised classification, i.e., how to incorporate some labeled information to supervise the EM algorithm.

Actually, we can easily inject the labeled information by initializing the conditional probability $p(z|d)$ with true labels. Specifically, if the document d belongs to class l, then $p(z_l|d) = 1$, else $p(z_k|d) = 0 (k \neq l)$; for the unlabeled data, the probability $p(z|d)$ is assigned based on the output of baseline methods in the experiments and normalized as $\sum_k p(z_k|d) = 1$. We update only the conditional probability $p(z|d)$ of unlabeled data during an EM iteration, while keep the ones associated with labeled data unchanged to supervise the optimization. After the EM iteration, we can predict the unlabeled data according to the output conditional probability $p(z|d)$,

$$l = \arg\max_k p(z_k|d). \tag{18}$$

The MVPLSA algorithm is detailed in Algorithm 1.

4. Experimental evaluation

4.1. Experimental setup

4.1.1. Data preparation

We use the data sets WebKB[2] and *20Newsgroups*[3] to evaluate our algorithm. WebKB data set has naturally two views of different representations, including the content features of the web pages and the link features exploited from the link structures. This data set consists of 877 web pages from computer science departments in four universities, i.e., Cornell, Washington, Wisconsin and Texas, and each university has five document classes, i.e., course, faculty, student, project and staff. The web pages from Cornell and Washington are selected as our experimental data, and we construct two-class problem similarly to the work [3]. Specifically, the class course is considered as the target class (positive class), and the rest of home pages belonging to other classes is negative. Therefore, for Cornell data we have 46 and 153 web pages respectively in positive and negative class, while for Washington data, the corresponding number of web pages is 66 and 164. We use the *tf* weighting scheme to represent the content features of documents. The description of these two multi-view problems is detailed in Table 2.

20Newsgroups is one of the widely used data sets for classification and clustering. This corpus has approximately 20,000 newsgroup documents, which are evenly divided into 20 subcategories. We follow the method in [15] to construct from this data set two multi-view learning problems (denoted as NewsGv2 and NewsGv4, respectively) as follows,

$$\text{NewsGv2}: \begin{bmatrix} & \textit{class 1} & \textit{class 2} \\ \textit{view 1} & \textit{windows.misc} & \textit{pc.hardware} \\ \textit{view 2} & \textit{politics.mideast} & \textit{politics.misc} \end{bmatrix},$$

[2] http://www.cs.purdue.edu/commugrate/data_access/all_data_sets.php?page=6.
[3] http://www.people.csail.mit.edu/jrennie/20Newsgroups/.

Table 2
The detailed description of the data sets. For the data sets from WebKB, V1 is the dimension of content features, while V2 stands for the dimension of link features.

Data sets	Size	Class	V1		V2	
Cornell	195	2	1703		195	
Washington	230	2	1703		230	
	Size	Class	V1	V2	V3	V4
NewsGv2	1000	2	3837	7452	–	–
NewsGv4	1500	3	6783	6307	7717	9336

$$\text{NewsGv4}: \begin{bmatrix} & \textit{class } 1 & \textit{class } 2 & \textit{class } 3 \\ \textit{view } 1 & \textit{alt.atheism} & \textit{comp.graphics} & \textit{windows.misc} \\ \textit{view } 2 & \textit{comp.windows.x} & \textit{misc.forsale} & \textit{rec.autos} \\ \textit{view } 3 & \textit{sport.hockey} & \textit{sci.crypt} & \textit{sci.electronics} \\ \textit{view } 4 & \textit{politics.guns} & \textit{politics.mideast} & \textit{politics.misc} \end{bmatrix}.$$

Specifically, considering the construction of data set NewsGv2, the data view 1 of class 1 from *windows.misc* is $\boldsymbol{X}^{(1)} \in \mathbb{R}_+^{N\times M^{(1)}}$ (where $\mathbb{R}_+$ denotes the non-negative real numbers, and N, $M^{(1)}$ are respectively the number of documents and features in view 1), and the data view 2 of class 1 from *politics.mideast* is $\boldsymbol{X}^{(2)} \in \mathbb{R}_+^{N\times M^{(2)}}$ ($M^{(2)}$ is the number of features in view 2). When concatenating the features from view 1 and view 2, we have $N \times N$ combinations, i.e., generating sample $x \in \boldsymbol{X}$ by concatenating vector $x^{(1)} \in \boldsymbol{X}^{(1)}$ from view 1 and vector $x^{(2)} \in \boldsymbol{X}^{(2)}$ from view 2 randomly drawn. Therefore, the constructed data from class 1 is $\boldsymbol{X} \in \mathbb{R}_+^{N\times(M^{(1)}+M^{(2)})}$. In this experiments, we consider only one randomly generated combination. The NewsGv4 data set is similarly constructed. We use the $tf \cdot idf$ weighting scheme to represent the document and the document frequency with the value of 5 adopted to cut down the number of word features. There are 500 documents in each class and the characteristics of the constructed multi-view learning problems are summarized in Table 2. Note that this construction may lead to a small portion of features co-occur in all views, which violates the assumption that all features from all views are conditionally independent.

4.1.2. Parameter setting

In our MVPLSA model, there are several parameters to be set, e.g., the number of latent topics for all views and number of iterations. To simplify our model, we set the same number of latent topics for all views, i.e., $Y^{(1)} = \cdots = Y^{(m)}$. We set the number of topics to 12 for the problems from WebKB and 128 for the tasks from *20Newsgroups*. The number of iterations is set to 150. For all the tasks, the number of document clusters is set to the true number of document classes.

We carefully tune the parameters of co-training [3][4] in the following section, and the parameters of the compared MVC algorithm [15] are set the same as in the original paper. Specifically, for MVC the number of clusters is set as the number of true clusters and the k-means algorithm is also used as the base clustering algorithm. It is also worth to mention that the EM algorithm usually finds only local optimal solutions and converges slowly, so it is very important to initialize properly the probability $p(z|d)$. In the experiments, we assign the initial value of $p(z|d)$ as the ensemble results of Naïve Bayesian model [14,16] for classification and as the results of multi-view clustering algorithm MVC [15] for clustering.

4.1.3. Evaluation metrics

As our evaluation metric for classification the accuracy is used, defined as follows:

$$auc = \frac{|\{d|d \in \mathcal{D} \wedge f(d) = y\}|}{|\mathcal{D}|}, \tag{19}$$

where y is the true label of document d, $\mathcal{D}|$ is number of documents and the function $f(d)$ gives d a prediction label.

We also adopt two popular metrics *normalized mutual information* (*NMI*) and *Pairwise F-measure* (*PF* for short) for clustering evaluation. *NMI* [6] measures how closely the clustering algorithm could reconstruct the underlying label distribution in the data. *PF*, defined in [1], is the harmonic mean of pairwise precision and recall. If L is the random variable denoting the underlying class labels on the data, and P is the random variable denoting the cluster assignments, then *NMI* measure is defined as

$$NMI = \frac{I(P,L)}{(H(P)+H(L))/2}, \tag{20}$$

where $I(X;Y) = H(X) - H(X|Y)$ is the mutual information between the random variables X and Y, $H(X)$ is the Shannon entropy of X, and $H(X|Y)$ is the conditional entropy of X given Y.

[4] For co-training, we tune the parameters u, p and n (where u is the number of samples in the pool, and p, n are respectively the number of positive and negative samples selected by the classifiers in each iteration), and report the best results.

PF is defined as follows,

$$Precision = \frac{\#PairsCorrectlyPredictedInSameCluster}{\#TotalPairsPredictedInSameCluster}, \tag{21}$$

$$Recall = \frac{\#PairsCorrectlyPredictedInSameCluster}{\#TotalPairsActuallyInSameCluster}, \tag{22}$$

$$PF = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{23}$$

4.2. Classification results

To validate the classification performance of our MVPLSA algorithm, we first compare it with several baseline methods:

- The Naïve Bayesian model (NBC) [16], which learns the models from each single view.
- The ensemble results of NBC over all views, denoted as NBC_E.
- The state-of-the-art multi-view co-training [3][5] approach.

We conduct the experiments over all four data sets depicted in Table 2, and randomly sample documents with the rate r from the data sets as labeled instances. The value of the rate r ranges from 0.01 to 0.05 with interval 0.01 to create a wide range of application scenarios. We perform 20 independent runs for each scenario, and the average performances and standard deviations are both recorded in Table 3. To make a clear comparison, the best results are marked with bold font. From these detailed results, we have following promising observations:

(1) MVPLSA significantly outperforms all the the compared approaches. Especially, our algorithm MVPLSA can improve the accuracy from 65.79% (the best result of NBC on four views) to 99.43% on the data set NewsGv4 when there are only 1% of data sampled as labeled information.
(2) The most important founding is that MVPLSA can obtain very good results even if there is only a small portion of labeled data, i.e., 1%, which indicates that MVPLSA is very effective and robust for semi-supervised learning. This also can dramatically alleviate the human labor to collect labeled samples or manually label samples.
(3) The performances of all the algorithms improve with the increasing sampling rate r. The more labeled data we have, the higher accuracy we obtain for all the classification algorithms. We find also that all the algorithms perform much more stable when more labeled data are incorporated since the standard deviation decreases with the increasing sampling rate r.
(4) The link features make almost no contribution to the performance improvement, especially in Cornell data set the accuracy of CoTr_V2 is very similar to NBC_V2.

We further compare MVPLSA with DPLSA on all four data sets, whose results are shown in Table 4. As discussed in Section 2.2, the DPLSA model conducted on all combined features assumes that all features are conditionally independent with other, while MVPLSA relaxes this condition. Indeed, some features from different views may correlate with each other, i.e., in the constructed data set NewGv2 a small portion of features co-occurs in both views. The experimental results show that MVPLSA is better than DPLSA. Overall, our model outperforms all compared algorithms according to the t-test with confidence 95%, and is more robust than DPLSA.

4.3. Clustering results

Similarly, we perform the document clustering experiments over all four data sets, and the average results of 20 independent runs are reported in Fig. 2. We compare MVPLSA with the k-means algorithm from the software package CLUTO[6] applied to each single view, and the best (Kmeans_Best) and average (Kmeans_Avg) results are shown in Fig. 2. The recently proposed multi-view clustering method MVC [15] is compared, and the k-means algorithm is also used as the base clustering algorithm for MVC.

NMI and *PF* measures are shown in Fig. 2a and b, respectively. The results show that MVPLSA is better than all the compared algorithms on three data sets Washington, NewsGv2 and NewsGv4, except the very sparse Cornell data set. We conjecture that without supervision information, MVPLSA may not model well the sparse co-occurrence to approximate the parameters. All the results shown above confirm the superiority and effectiveness of our MVPLSA model.

[5] Here we do not compare our method with the improved versions of co-training, since most of them focus on some specific scenarios, e.g., whether the views are sufficiently or insufficiently compatible for multi-view learning [5]. The consideration of these specific scenarios is out of the scope of this paper.
[6] The code from http://www.glaros.dtc.umn.edu/gkhome/cluto/cluto/ download, and default parameters are adopted.

Table 3
Classification accuracy (average performance and standard deviation) of all the algorithms on four data sets (two from WebKB, two from *20Newsgroups*).[a]

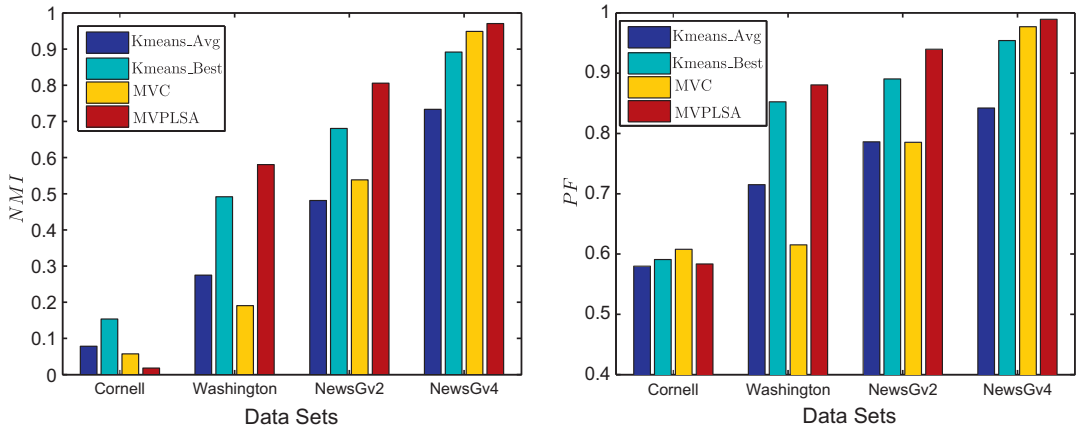
	$r = 0.01$	$r = 0.02$	$r = 0.03$	$r = 0.04$	$r = 0.05$
	Cornell				
NBC_V1	82.06 ± 3.74	79.13 ± 1.00	82.47 ± 3.16	81.64 ± 2.30	85.79 ± 3.36
NBC_V2	77.24 ± 8.91	79.05 ± 0.61	77.55 ± 7.04	79.44 ± 0.74	80.08 ± 1.08
NBC_E	81.82 ± 3.48	78.92 ± 0.73	81.89 ± 2.82	80.99 ± 2.03	84.89 ± 3.25
CoTr_V1	82.86 ± 5.35	80.58 ± 4.63	83.91 ± 5.06	88.01 ± 5.72	91.09 ± 4.36
CoTr_V2	79.43 ± 2.24	79.55 ± 1.31	79.97 ± 1.20	80.86 ± 2.15	82.55 ± 2.34
CoTr_E	82.97 ± 4.88	80.95 ± 4.48	84.07 ± 4.73	88.23 ± 5.63	91.33 ± 4.56
MVPLSA	93.49 ± 4.54	94.32 ± 2.91	94.65 ± 2.17	95.54 ± 0.96	95.90 ± 1.14
	Washington				
NBC_V1	82.05 ± 7.10	86.83 ± 4.92	79.84 ± 5.19	85.23 ± 7.45	89.95 ± 3.95
NBC_V2	71.96 ± 0.48	72.32 ± 0.84	71.97 ± 0.72	72.32 ± 1.06	72.33 ± 1.23
NBC_E	80.88 ± 6.79	86.18 ± 5.33	79.10 ± 4.89	84.50 ± 7.24	88.85 ± 4.13
CoTr_V1	83.85 ± 10.22	89.91 ± 5.35	87.98 ± 8.83	90.43 ± 6.15	92.12 ± 2.06
CoTr_V2	74.96 ± 3.70	77.19 ± 2.87	77.06 ± 4.03	78.23 ± 3.17	78.85 ± 2.52
CoTr_E	83.55 ± 9.69	89.78 ± 5.15	87.89 ± 8.57	89.95 ± 5.91	91.94 ± 2.13
MVPLSA	93.55 ± 1.45	94.06 ± 1.30	94.22 ± 1.42	94.50 ± 1.21	93.96 ± 1.62
	NewsGv2				
NBC_V1	61.01 ± 4.50	67.92 ± 3.86	68.14 ± 5.48	70.29 ± 6.38	73.68 ± 2.78
NBC_V2	68.58 ± 4.60	77.91 ± 5.02	82.15 ± 4.44	86.11 ± 2.81	86.85 ± 2.40
NBC_E	70.60 ± 5.93	81.77 ± 5.37	85.29 ± 4.61	87.99 ± 3.14	90.48 ± 1.93
CoTr_V1	82.54 ± 10.83	86.87 ± 3.95	87.96 ± 3.45	88.43 ± 2.95	89.18 ± 2.65
CoTr_V2	89.83 ± 13.11	94.55 ± 1.77	95.19 ± 1.13	95.15 ± 0.99	95.65 ± 1.11
CoTr_E	90.10 ± 13.10	94.95 ± 2.54	95.65 ± 1.75	95.92 ± 1.31	96.34 ± 1.28
MVPLSA	94.44 ± 4.74	97.10 ± 0.66	96.71 ± 0.93	97.16 ± 0.76	97.36 ± 0.55
	NewsGv4				
NBC_V1	58.76 ± 8.44	70.01 ± 5.89	74.41 ± 5.51	78.28 ± 2.75	79.46 ± 2.42
NBC_V2	63.89 ± 5.63	75.11 ± 6.09	79.13 ± 5.04	81.58 ± 5.23	85.38 ± 1.90
NBC_V3	65.79 ± 8.20	76.04 ± 6.64	78.51 ± 8.04	87.59 ± 4.87	89.86 ± 2.94
NBC_V4	51.85 ± 5.23	65.35 ± 5.51	71.85 ± 4.00	76.17 ± 3.74	79.93 ± 2.22
NBC_E	78.73 ± 6.06	91.26 ± 2.48	94.10 ± 2.10	96.43 ± 1.56	97.80 ± 0.67
MVPLSA	99.43 ± 0.11	99.50 ± 0.06	99.48 ± 0.09	99.50 ± 0.06	99.49 ± 0.06

[a] NBC_V*i* denotes the the accuracy of NBC when training and testing on view *i*. Similarly, CoTr_V*i* denotes the the accuracy of co-training when training and testing on view *i*. For the data set NewsGV4 with more than two views, co-training is not applicable. CoTr_E is the combination results of CoTr_V1 and CoTr_V2 with same weighted ensemble.

Table 4
The comparison results of DPLSA and MVPLSA on four data sets.[a]

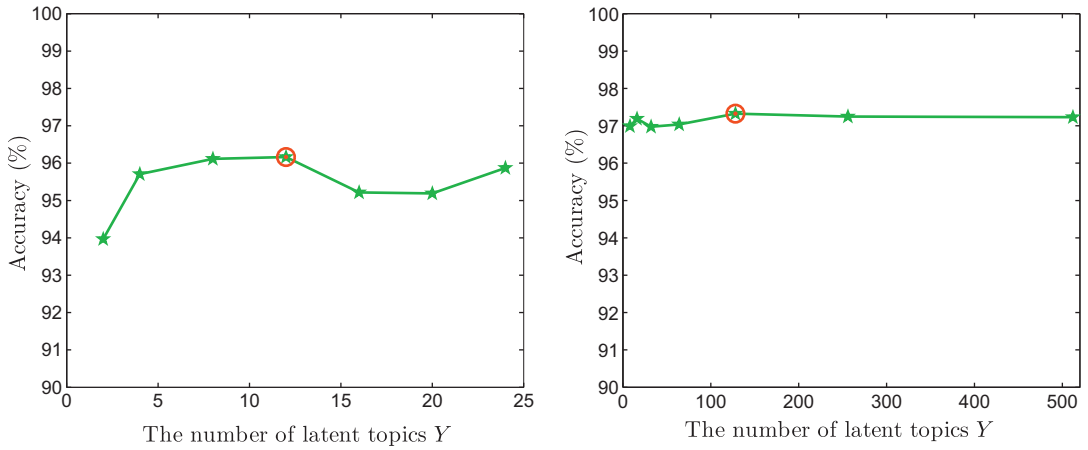
	$r = 0.01$	$r = 0.02$	$r = 0.03$	$r = 0.04$	$r = 0.05$
	Cornell				
DPLSA_V1	91.64 ± 4.54	92.61 ± 6.61	93.46 ± 2.78	94.78 ± 1.26	94.40 ± 1.50
DPLSA_V2	74.66 ± 8.63	74.76 ± 3.01	73.96 ± 9.28	77.77 ± 3.72	81.20 ± 2.78
DPLSA	92.97 ± 3.27	94.16 ± 2.49	94.18 ± 2.91	95.35 ± 1.01	95.52 ± 1.34
MVPLSA	93.49 ± 4.54	94.32 ± 2.91	94.65 ± 2.17	95.54 ± 0.96	95.90 ± 1.14
	Washington				
DPLSA_V1	93.10 ± 1.19	93.17 ± 0.98	93.04 ± 1.22	93.41 ± 0.95	93.19 ± 1.12
DPLSA_V2	78.46 ± 4.88	81.29 ± 3.87	77.29 ± 4.16	80.05 ± 5.07	81.87 ± 2.24
DPLSA	93.07 ± 1.37	93.93 ± 1.38	93.26 ± 1.54	93.61 ± 1.29	93.06 ± 1.72
MVPLSA	93.55 ± 1.45	94.06 ± 1.30	94.22 ± 1.42	94.50 ± 1.21	93.96 ± 1.62
	NewsGv2				
DPLSA_V1	78.28 ± 3.58	83.15 ± 1.87	83.64 ± 2.45	84.67 ± 1.42	85.35 ± 0.94
DPLSA_V2	89.56 ± 7.63	93.07 ± 1.10	92.63 ± 2.43	93.28 ± 0.88	93.67 ± 1.08
DPLSA	85.38 ± 12.70	90.97 ± 3.60	89.80 ± 3.21	90.40 ± 2.70	92.15 ± 2.45
MVPLSA	94.44 ± 4.74	97.10 ± 0.66	96.71 ± 0.93	97.16 ± 0.76	97.36 ± 0.55
	NewsGv4				
DPLSA_V1	90.23 ± 1.43	91.36 ± 0.75	91.84 ± 0.47	91.76 ± 0.51	91.98 ± 0.74
DPLSA_V2	90.79 ± 0.34	90.95 ± 0.27	91.15 ± 0.15	91.25 ± 0.28	91.54 ± 0.31
DPLSA_V3	97.23 ± 0.26	97.40 ± 0.19	97.34 ± 0.27	97.43 ± 0.20	97.55 ± 0.18
DPLSA_V4	91.89 ± 0.66	92.60 ± 0.87	92.86 ± 0.57	93.26 ± 0.49	93.20 ± 0.54
DPLSA	98.66 ± 0.41	98.99 ± 0.19	99.00 ± 0.26	99.07 ± 0.12	99.13 ± 0.11
MVPLSA	99.43 ± 0.11	99.50 ± 0.06	99.48 ± 0.09	99.50 ± 0.06	99.49 ± 0.06

[a] DPLSA_V*i* denotes the accuracy of DPLSA model on view *i*, while DPLSA is the one on all combined feature space.



(a) Clustering evaluation in terms of *NMI* (b) Clustering evaluation in terms of *PF*

Fig. 2. Clustering evaluation of all the algorithms on four data sets (two from WebKB, two from *20Newsgroups*).



(a) Cornell data set (b) NewsGv2 data set

Fig. 3. The effect of number of latent topics on accuracy (sampling rate $r = 0.05$).

4.4. Parameter tuning

In our experiments, we also empirically investigate the performance of MVPLSA affected by different setting of the number of topics Y on Cornell and NewsGv2 data sets. We examine seven values of Y, $Y \in \{2, 4, 8, 12, 16, 20, 24\}$ for the Cornell data set and $Y \in \{8, 16, 32, 64, 128, 256, 512\}$ for the NewsGv2 data set. The results are shown in Fig. 3 under sampling rate $r = 0.05$. As a matter of fact, our MVPLSA model is not sensitive to the number of latent topics, especially on the data sets from *20Newsgroups*. For the Cornell data set, the performance of MVPLSA is also very stable under different number of feature topics, even if there are only two latent topics. MVPLSA can obtain the high accuracy up to 94% (similar with the best result 96%).

We get the best results when the number of latent topics equals 12 for the Cornell data set and 128 for the NewsGv2 data set (Fig. 3). Thus, the number of latent topics $Y = 12$ is adopted for the data sets from WebKB, and $Y = 128$ for the data sets from *20Newsgroups* in all our experiments.

4.5. Discussion

MVPLSA model is a general model that not only can deal with data with naturally multiple representations, but also can improve the performance of data with only one view. We use the data set NewsGv4 for this study. For each representation of data set NewsGv4, we randomly split the features into four parts to construct multi-view scenario, and MVPLSA runs on these new four views data. The comparison results of NBC and MVPLSA are shown in Table 5. MVPLSA model can even

Table 5
Accuracy (%) of NBC and MVPLSA models under multiple views for each representation of the NewsGv4 data set (1% labeled data).

	View 1	View 2	View 3	View 4
NBC	58.76	63.89	65.79	51.85
MVPLSA	77.52	87.80	94.99	77.43

improve very low accuracy of NBC (51.85%) to 77.43%, what is a significant improvement and validates again the superiority of our model.

5. Conclusions

In this work we study the multi-view problem of data with several different representations, and propose a new multi-view learning algorithm MVPLSA based on the generative model for probabilistic latent semantic analysis. Compared with the baseline model, i.e., DPLSA, our MVPLSA model can jointly model the co-occurrences from multiple views and obtain additional gains. Furthermore, MVPLSA can handle the data with only one view by splitting the feature space, and improve the performance. To derive the solution to the MVPLSA model, an EM algorithm is developed. Finally, the experimental results on four real-world data sets show the effectiveness and advantage of the proposed model.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (Nos. 60933004, 60975039, 61175052, 61035003, and 61072085).

References

[1] S. Basu, A. Banerjee, R.J. Mooney, Active semi-supervision for pairwise constrained clustering, in: Proc. of the SIAM International Conference on Data Mining (SDM), 2004, pp. 333–344.
[2] S. Bickel, T. Scheffer, Multi-view clustering, in: Proc. of the IEEE International Conference on Data Minig (ICDM), 2004, pp. 19–26.
[3] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proc. of the 11th Annual Conference on Computational Learning Theory, 1998, pp. 92–100.
[4] U. Brefeld, T. Scheffer, Co-em support vector learning, in: Proc. of International Conference of Maching Learning (ICML), 2004, pp. 16–23.
[5] C. Christoudias, R. Urtasun, T. Darrell, Multi-view learning in the presence of view disagreement, in: Proc. of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI), 2008, pp. 88–96.
[6] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley Interscience, 1991.
[7] T. Hofmann, Probabilistic latent semantic analysis, in: Proc. of 15th Conference on Uncertainty in Artificial Intelligence (UAI), 1999, pp. 289–296.
[8] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Machine Learning (2001) 177–196.
[9] Y. Jiho, S.J. Choi, Probabilistic matrix tri-factorization, in: Proc. of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2009, pp. 1553–1556.
[10] J.J. Castillo, A WordNet-based semantic approach to textual entailment and cross-lingual textual entailment, Journal of Machine Learning and Cybernetics 2 (3) (2011) 177–189.
[11] S.M. Kakade, D.P. Foster, Multi-view regression via canonical correlation analysis, in: Proceedings of the 20th Annual Conference on Learning Theory, 2007, pp. 82–96.
[12] Y.M. Kim, M. Amini, C. Goutte, P. Gallinari, Multi-view clustering of multilingual documents, in: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010, pp. 821–822.
[13] S. Koço, C. Capponi, A boosting approach to multiview classification with cooperation, in: Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases – Volume Part II, 2011, pp. 209–228.
[14] D. Lewis, M. Riguette, A comparison of two learning algorithms for text categorization, in: Proc. 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994, pp. 81–93.
[15] B. Long, P.S. Yu, Z.F. Zhang, A general model for multiple view unsupervised learning, in: Proc. of the SIAM International Conference on Data Mining (SDM), 2008, pp. 822–833.
[16] A. McCallum, K. Nigam, A comparison of event models for naive bayes text classification, in: Proc. of the AAAI Workshop on Learning for Text Categorization, 1998, pp. 41–48.
[17] I. Muslea, R. Ghani, Active + semi-supervised learning = robust multi-view learning, in: Proc. of International Conference of Maching Learning (ICML), 2002, pp. 435–442.
[18] I. Muslea, S. Minton, C.A. Knoblock, Selective sampling with redundant views, in: Proc. of the Seventeenth National Conference on Artificial Intelligence (AAAI), 2000, pp. 621–626.
[19] K. Nigam, R. Ghani, Analyzing the effectiveness and applicability of co-training, in: Proc. of the 9th ACM Conference on Information and Knowledge Management (CIKM), 2000, pp. 86–93.
[20] S. Rüping, T. Scheffer, Learning with multiple views, in: Proc. of International Conference of Maching Learning Workshop on Learning with Multiple Views, 2005, pp. 1–86.
[21] X.Z. Wang, C.R. Dong, Improving generalization of fuzzy if–then rules by maximizing fuzzy entropy, Fuzzy Systems, IEEE Transactions on 17 (3) (2009) 556–567.
[22] X.Z. Wang, L.C. Dong, J.H. Yan, Maximum ambiguity based sample selection in fuzzy decision tree induction, IEEE Transactions on Knowledge and Data Engineering (2011).
[23] X.Z. Wang, J.H. Zhai, S.X. Lu, Induction of multiple fuzzy decision trees based on rough set technique, Information Sciences 178 (16) (2008) 3188–3202.
[24] L. Wenyin, X. Quan, M. Feng, B. Qiu, A short text modeling method combining semantic and statistical information, Information Sciences 180 (20) (2010) 4031–4041.
[25] R. Xia, C. Zong, S. Li, Ensemble of feature sets and classification algorithms for sentiment classification, Information Sciences (2011).
[26] D. Yarowsky, Unsupervised word sense disambiguation rivaling supervised methods, in: Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL), 1995, pp. 189–196.

[27] D. Zhang, F. Wang, C.S. Zhang, T. Li, Multi-view local learning, in: Proc. of the 23nd Conference on Artificial Intelligence (AAAI), 2008, pp. 752–757.
[28] Y. Zhang, R. Jin, Z.H. Zhou, Understanding bag-of-words model: a statistical framework, Journal of Machine Learning and Cybernetics 1 (1–4) (2010) 43–52.
[29] D.Y. Zhou, J.C. Christopher, Burges, Spectral clustering and transductive learning with multiple views, in: Proc. of International Conference of Maching Learning (ICML), 2007, pp. 1159–1166.
[30] F.Z. Zhuang, P. Luo, Z.Y. Shen, Q. He, Y.H. Xiong, Z.Z. Shi, H. Xiong, Collaborative dual-PLSA: mining distinction and commonality across multiple domains for text classification, in: Proc. of the 19th ACM Conference on Information and Knowledge Management (CIKM), 2010, pp. 359–368.