Classify social image by integrating multi-modal content

Article *in* Multimedia Tools and Applications · April 2017

DOI: 10.1007/s11042-017-4657-2

CITATIONS
2

READS
214

5 authors, including:

Xiaoming Zhang
Beihang University (BUAA)
63 PUBLICATIONS **478** CITATIONS
SEE PROFILE

Senzhang Wang
Nanjing University of Aeronautics & Astronautics
107 PUBLICATIONS **770** CITATIONS
SEE PROFILE

Multimed Tools Appl
DOI 10.1007/s11042-017-4657-2

Classify social image by integrating multi-modal content

Xiaoming Zhang[1] · Xu Zhang[2] · Xiong Li[2] · Zhoujun Li[3] · Senzhang Wang[4]

Received: 17 July 2016 / Revised: 9 February 2017 / Accepted: 29 March 2017

Abstract There is a growing volume of social images with the development of social networks and digital cameras. Usually, these images are annotated with textual tags besides the visual content. It is quite urgent to automatically organize and manage this large number of social images. Image classification is the basic task of these applications and has attracted great research efforts. Though there are many researches on image classification, it is of considerable challenge to integrate the multi-modal content of social images simultaneously for classification, since the textual content and visual content are represented in two heterogeneous feature spaces. In this paper, we proposed a multi-modal learning method to integrate multi-modal features through their correlation seamlessly. Specifically, we learn two linear classification modules for the two types of features, and then they are integrated by the l_2 normalization method via a joint model. Each classier is normalized with $l_{2,1}$ to reduce the effect of the noisy features by selecting a subset of more important features. With the joint model, the classification based on visual features can be reinforced by the classification based on textual features, and vice verse. Then, the test image is classified based on both the textual features and visual features by combing the results of the two classifiers. Experiments conducted on real-world social image datasets demonstrate the superiority of our proposed method compared with the representative baselines.

✉ Xu Zhang
zhangxu@cert.org.cn

Xiaoming Zhang
yolixs@buaa.edu.cn

[1] Beijing Key Laboratory of Network Technology, Beihang University, Beijing, China

[2] National Computer Network Emergency Response Technical Team of China, Beijing, China

[3] State Key Laboratory of Software Development Environment, Beihang University, Beijing, China

[4] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China

Published online: 13 April 2017

Keywords Image classification · Multi-modal classification · Social image analysis

1 Introduction

With the rising popularity of camera devices and network terminals, the amount of user-contributed social images is increasing rapidly. Usually, these images contain rich information like tags and visual content as shown in Fig. 1. With such a large number of social images, it is of great challenge to analyze and organize the multimodal data effectively [1, 9, 10, 14], which also introduces new clues for research and applications in multimedia domain [24, 25, 39, 43]. As a fundamental task in image applications, classification has been widely used in social image annotation, retrieval, reranking etc. Multi-modality is a typical characteristic of social images. Usually, users upload images with corresponding keywords called tags. Therefore, a social image provides visual content and textual tags, and the two types of content are represented in two heterogeneous feature spaces. Both of these contents relate to the labels of these images and they are also correlated with each other. For example, the visual content of animal is different from those of architecture, and both of the two classes of images have their specific textual tags. However, how to fuse the multimodal features to classify social images is a challenging problem.

A number of methods [6, 12, 30, 40–42, 44, 45] have been proposed to conduct image classification. Many of these approaches classify the images based on the visual content directly [30, 40]. However, these approaches generally face a important problem, i.e., the so-called "semantic gap" between the lower level visual features and the higher level semantics of multimedia objects, which means that the visually similar multimedia objects are not always semantically similar. Therefore the performance is greatly affected by the ineffective representation of images. On the other side, there are also some works to classify social images by fusing multi-modal features, in which early fusion and late fusion are the two most common forms. The early fusion methods try to map multiple feature spaces to a unified space, on which traditional similarity evaluation can be conducted [12, 45]. However, since social images and their associated text information are of wide diversity, it is nearly infeasible to construct a suitable latent space to cover such a large diversity. The late fusion methods fuse the results learned from different features [3, 38, 41]. However, in social media systems, the multi-modal features are always correlated with each other and the late fusion method can not effectively capture the interaction among them. On the other side, images are represented by high dimensional feature vectors. Many of the features are noisy or redundant to image classification. Therefore, it is necessary to reduce the effect of these features to improve the performance of classification.

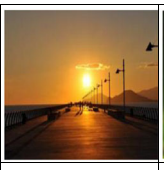
Image					
Tags	airplane, sunset, landing, aircraft, sky, sun, lights, aeroplane, land, wheels.	White, ocean, water, horse, sand, beach, sky, animal, island.	Ocean, beach, water, clouds, beautiful, travel, blue, sky, Clouds, week.	Sunset, sun, bridge, sand beach, water, golden glow, cloud	flower, plant, grass, fellow, park, tree, beautiful, leaves

Fig. 1 Examples of social images

For the above reasons, both of the late fusion and early fusion methods can't effectively explore the multi-modal features and their correlation for social image classification. In this paper, we propose to directly integrate the multi-modal features and the correlation for social image classifier learning. Specifically, we propose a multi-modal learning model (MML) to integrate different features with their correlation. For each type of features, a linear classification module with $l_{2,1}$-norm regularization is learned. By using the $l_{2,1}$-norm regularization, the features which are important to classification are more likely to be selected for classifier learning, which reduces the effect of noisy features. To capture the correlation between different types of features, a joint model with l_2-norm regularization is adopted. With the joint model, the classifiers based on visual features and textual features can be reinforced by each other. Then, the test image can be classified based on both the textual features and visual features by combing the results of the two classifiers. We conduct extensive experiments to evaluate the performance of our approach. The main contributions of this paper are summarized as follows:

- We propose to tackle the problem of social images classification with multi-modal content. We present a novel classification approach (MML)based on multi-modal learning, with a joint model to reinforce the learning of each classifiers. Meanwhile, the $l_{2,1}$-norm regularization is adopted to reduce the effect of noisy and redundant features.
- A efficient optimization algorithm is proposed to solve the object function of MML, and the final classification result is obtained by combining the classification results on two kinds of features.
- Extensive experiments are conducted on two real-world social image datasets to demonstrate the effectiveness of our approach.

The rest of the paper is organized as follows. Section 2 describes the related works. Our novel multi-modal learning approach is elaborated in Section 3. Social image classification based on multi-modal content is given in Section 4. Section 5 shows the experiments and Section 6 concludes the paper.

2 Related works

Social image classification is an active research topic in recent years. Recently, many methods have been proposed for multi-modal classification [8, 28, 38, 48]. According to the strategy of integrating multi-modal content, the multi-modal classification methods can be grouped into two major categories: early fusion and late fusion.

A direct strategy of early fusion is to concatenate the different kinds of features into a long vector, which leads to the curse of dimensionality problem. Therefore, many sophisticated techniques are developed, which includes those similarity space fusion and multi-view subspace learning approaches. Most of the works on similarity space fusion conducted in the forms of kernel fusion, e.g., Multiple kernel learning (MKL) [22]. In [23], a combination of different kernels built on different features sets were utilized for protein prediction. In [26], an approach called MKL-DR is proposed to incorporates multiple kernel learning (MKL) into the training process of dimensionality reduction (DR) algorithms. It works with multiple base kernels, each of which is created based on a specific kind of visual feature, and combines these features in the domain of kernel matrices. Then, the formulation is illustrated with the framework of graph embedding [51], which presents a unified view for a large family of DR methods. Kloft et al. [19] extends the traditional l_1-norm MKL to arbitrary norms, and shows that the non-sparse MKL is superior to the state-of-the-art in

combining different feature sets for biometrics recognition. Another set of methods is on multi-modal subspace learning [49, 50], in which canonical correlation analysis (CCA) [16] is one of the most popular algorithm. SVM-2K [12] combines Kernel CCA (KCCA) and support vector machine (SVM) in a single optimization problem. M-LSA [13] maps original features to a latent space with lower dimensionality based on statistical methods. Another method learns a semantically refined visual bag-of-words (BOW) representation for image classification, which explores the correlation matrix between visual and textual words [29].

Schemes in the late fusion category either learn classifiers of different kinds of features independently or interactively. In [38], the SVM outputs are firstly converted to probabilistic scores, and then concatenated as the input of an SVM for final classification. A thorough study on the weighted voting methods for classifier fusion is presented in [46]. The other set of interactive fusion methods communicate information with other features when learning classifier of the current kind of feature. Many of these methods are semi-supervised and naturally co-training [2], and co-regularization [20]. Co-training is based on the compatibility and class conditional independence assumptions. For instance, Zhou and Li [54] developed a co-training style semi-supervised regression algorithm called CoREG. This algorithm employs two k-nearest neighbor (kNN) regressors, each of which labels the unlabeled data for the other during the learning process. The co-EM algorithm [34] bootstrapped samples in a similar way like co-training. It demonstrates that, when learning from labeled and unlabeled data, algorithms explicitly leveraging a natural independent split of the features outperform algorithms that do not. By formulating the linear classifier in a probabilistic framework, SVM is introduced as the base classifier in co-EM [5], which utilizes unlabeled data when the available attributes can be split into two independent subsets each of which has to be sufficient for learning. The co-regularization method enforced the classifiers of different features to be agreed on unlabeled data, in a graph-based semi-supervised framework [20]. Inspired by the co-regularization algorithm, Brefeld et al. [4] proposed a co-regression algorithm. This method learned multiple regressions for multiple features stimulatingly, with a cost function to impose the agreement among predictors on unlabeled examples.

Recently, with the development of deep learning, there are also many deep learning based image classification methods [7, 21, 31]. In these works, different types of deep models are adopted to classify images based on visual content. A semi-supervised classifier called discriminative deep belief networks (DDBN) is proposed to classify visual data in [27], in which the backpropagation strategy is adopted to optimize the classification results in training dataset by refining the parameter space. In [21], the convolutional deep networks are proposed to classify visual data. The constructed deep architecture is fine-tuned by gradient-descent based supervised learning with an exponential loss function. To reduce the effect of image resolution and objects sizes, a multi-scale convolutional neural network (CNN) is proposed to learn the scale-variant and scale-invariant features for deep image classification [36]. The attention model is also used in deep convolution neural network for fine-grained image classification [47]. Though the deep methods can improve the performance of image classification, these approaches mainly focus on processing the visual content of images, which can not be directly adopted to exploring the multi-modal features for social image classification.

3 Classifier learning for multi-modal content

In this section we first introduce the notations used in the paper, and then present the multi-modal learning method for social image classification.

3.1 Notations

In the remaining presentation, the following notations are used. Matrices are denoted by boldface uppercase letters, vectors by boldface lowercase letters, and scalars by lower case letters. For an matrix $\|\mathbf{A}\| \in \mathcal{R}^{n\times m}$, $\mathbf{A}^T$ denotes its transpose, $\mathbf{A}_i$ and $\mathbf{A}^j$ denote its *i-th* row and *j-th* column respectively, and $\|\mathbf{A}\|_{2,1}$ denotes the $l_{2,1}$-norm regularization [33], i.e., $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} (\mathbf{A}_{ij})^2}$.

Let $\mathbf{I}$ denotes the training dataset with n social images, and each image $I_i = \{\mathbf{x}_i, \mathbf{y}_i, \mathbf{l}_i\}_{I_i \in \mathbf{I}}$ consists of three atoms: $\mathbf{x}_i \in \mathcal{R}^d$ is the feature vector of the visual content; $\mathbf{y}_i \in \mathcal{R}^v$ is the feature vector of the textual content; $\mathbf{l}_i \in \{0, 1\}^{c\times 1}$ is the class indicator vector, where c is the number of image classes and l_{ij}=1 if the *i-th* image is labeled with the *j-th* class, and l_{ij}=0 otherwise. Specifically, let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$ denotes the visual feature matrix, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]$ denotes the textual feature matrix, and $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_n]^T$ denotes the class label matrix of the training image set.

3.2 Multi-modal classifier learning

With the label information, we can then learning a classifier for social image with multi-modal features. First, we adopt a linear model to predict the class label of image I_j based on its visual features as following:

$$f_i(I_j) = \mathbf{w}_i^T \mathbf{x}_j \tag{1}$$

where $w_j \in \mathcal{R}^d$ is the weight vector, and $\mathbf{W} = [\mathbf{w}_i, \mathbf{w}_2, ..., \mathbf{w}_c]$ denotes the weight matrix which also indicates the importance of each feature to different classes. Let the predicted label set $\mathbf{Q} \in \mathcal{R}^{n\times c}$. Hence, the classifier based on the visual features is formulated as following:

$$\min_{\mathbf{W},\mathbf{Q}} \|\mathbf{X}^T\mathbf{W} - \mathbf{Q}\|_F^2 + \beta\|\mathbf{W}\|_{2,1} \tag{2}$$

where the least square loss is used for the class label prediction, Usually, not all the features are equal to classification, and some of the features are even noisy or redundant. The project matrix $\mathbf{W}$ is also used for feature selection, and a sophisticated regularizer is needed to make $\mathbf{W}$ able to reflect the importance of different features. Thus, the $l_{2,1}$ normalization $\|\mathbf{W}\|_{2,1}$ is adopted to guarantee that $\mathbf{W}$ is sparse in rows [33, 52], which constrains the number of features to be selected since some features are unhelpful.

Since social images contain both visual features and textual features, the textual features also reflects the semantics and can be used to predict the class of each image. Therefore, textual features are used to complement the classification on visual features. Assume the predicted label set based on textual features is $\mathbf{Q}' \in \mathcal{R}^{n\times c}$, a similar classifier based on the textual feature matrix $\mathbf{Y}$ is defined as follows:

$$\min_{\mathbf{W}',\mathbf{Q}'} \|\mathbf{Y}^T\mathbf{W}' - \mathbf{Q}'\|_F^2 + \beta\|\mathbf{W}'\|_{2,1} \tag{3}$$

where $\mathbf{W}'$ is the weight matrix for the textual features. Meanwhile, the classification results based on both types of these features should be consistent with each other. We propose to use $\|\mathbf{Q} - \mathbf{Q}'\|_F^2$ to penalize the diversity of the labels predicted by different types of features.

Next, we build up the connection between the predicted labels and the ground truth labels $\mathbf{L} \in \{0, 1\}^{n\times c}$. Both $\mathbf{Q}$ and $\mathbf{Q}'$ is supposed to be consistent with $\mathbf{L}$ and we

propose to minimize $\|\mathbf{Q} - \mathbf{L}\|_F^2 + \|\mathbf{Q}' - \mathbf{L}\|_F^2$. To this end, we propose the joint object function for classification based on both visual features and textual features as follows:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{W}',\mathbf{Q},\mathbf{Q}'} \quad & \|\mathbf{X}^T\mathbf{W} - \mathbf{Q}\|_F^2 + \lambda\|\mathbf{Y}^T\mathbf{W}' - \mathbf{Q}'\|_F^2 \\ & +\beta(\|\mathbf{W}\|_{2,1} + \|\mathbf{W}'\|_{2,1}) + \delta\|\mathbf{Q} - \mathbf{Q}'\|_F^2 \\ & +\eta(\|\mathbf{Q} - \mathbf{L}\|_F^2 + \|\mathbf{Q}' - \mathbf{L}\|_F^2) \end{aligned} \tag{4}$$

where λ is a trade-off parameter to balance the importance of visual features and textual features to the classifier learning. A greater value of λ means that the textual features are more important to classify image. In this way, the module in (4) provides us a powerful and flexible tool for training the effective image classifier, which simultaneously conducts the classifier learned from both visual features and textual features, by exploring their intrinsic correlations. The fundamental design principle of the joint model lies in that the classifier on visual features and classifier on textual features should form a mutually-reinforcing learning loop. The learned weights of textual features should be well explored to be better embedded and help the learning of the weights of the visual features, and vice verse.

3.3 Optimization

It is difficult to solve (4) directly since it is nonconvex w.r.t. all the variables at the same time, and the non-smooth property of regularization makes it non-trivial to optimize the problem as a whole. To address these challenges, we devise an iterative optimization algorithm to optimize the model.

We first need to introduce a variational formulation for the $l_{2,1}$ norm. If we define $\phi(x) = \sqrt{x^2 + \varepsilon}$, the $l_{2,1}$ norm $\|\mathbf{W}\|_{2,1}$ and $\|\mathbf{W}'\|_{2,1}$ can be replaced with $\sum_i^d \phi(\|\mathbf{W}_i\|_2)$ and $\sum_i^v \phi(\|\mathbf{W}'_i\|_2)$ respectively, where d and v denote the numbers of rows of $\mathbf{W}$ and $\mathbf{W}'$ respectively, according to the analysis for $l_{2,1}$ in [17]. ε is a smoothing term which is usually set to be a small value. It can be proved that $\phi(x)$ satisfies all the conditions as following:

$$\begin{aligned} & x \longrightarrow \phi(x) \; is \; convex \; on \; \mathcal{R}, \\ & x \longrightarrow \phi(\sqrt{x}) \; is \; concave \; on \; \mathcal{R}_+, \\ & \phi(x) = \phi(-x), \; \forall x \in \mathcal{R}, \\ & \phi(x) \; is \; C^1 on \; \mathcal{R}, \\ & \phi''(x) > 0, \; \lim_{u \to \infty} \phi(x)/x^2 = 0. \end{aligned} \tag{5}$$

Then $\phi(\cdot)$ can be optimized in a half-quadratic way [35] according to the following Lemma 1 [17].

Lemma 1 *Let $\phi(\cdot)$ be a function satisfying all the conditions in (5), for a fixed $\|\mathbf{x}\|_2$, there exists a dual potential function $\varphi(\cdot)$, such that*

$$\phi(\|\mathbf{x}\|_2) = \inf_{p \in \mathcal{R}} \{p\|\mathbf{x}\|_2^2 + \varphi(p)\} \tag{6}$$

where p is determined by the minimizer function $\delta(\cdot)$ with respect to $\phi(\cdot)$.

According to Lemma 1, using (6) on each $\phi(\|\mathbf{W}_i\|_2)$ for i, the object function (4) can be reformulated as follows:

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{W}',\mathbf{Q},\mathbf{Q}'} \|\mathbf{X}^T\mathbf{W}-\mathbf{Q}\|_F^2+\lambda\|\mathbf{Y}^T\mathbf{W}'-\mathbf{Q}'\|_F^2\\&+\beta(tr(\mathbf{W}^T\mathbf{D}\mathbf{W})+tr(\mathbf{W}'^T\mathbf{D}'\mathbf{W}))+\delta\|\mathbf{Q}-\mathbf{Q}'\|_F^2\\&+\eta(\|\mathbf{Q}-\mathbf{L}\|_F^2+\|\mathbf{Q}'-\mathbf{L}\|_F^2))\end{aligned} \tag{7}$$

where $\mathbf{D}=Diag(\mathbf{d})$ and $\mathbf{D}'=Diag(\mathbf{d}')$. $\mathbf{d}$ and $\mathbf{d}'$ are auxiliary vectors of the two $l_{2,1}$ norms respectively. The elements of $\mathbf{d}$ and $\mathbf{d}'$ are computed respectively as follows:

$$\begin{cases} d_i=\frac{1}{2\sqrt{\|\mathbf{W}_i\|_2^2+\varepsilon}}\\ d_i'=\frac{1}{2\sqrt{\|\mathbf{W}_i'\|_2^2+\varepsilon}}\end{cases} \tag{8}$$

where ε is a smoothing term, which is usually set to be a small constant value. This formulation is based on the half-quadratic optimization, i.e., $d_i=\delta(\|\mathbf{W}_i\|_2^2)$, where $\delta(.)$ is the minimizer function with respect to $\phi(.)$ shown in Lemma 1. Note that $\mathbf{D}$ and $\mathbf{D}'$ are actually depended on $\mathbf{W}$ and $\mathbf{W}'$ respectively. To handle this problem, we design an iterative algorithm, which updates $\mathbf{D}$ and $\mathbf{D}'$ in each iteration with $\mathbf{W}$ and $\mathbf{W}'$ of the previous iteration. That is, we update a matrix by fixing other matrices at each step.

Update Q and Q′ by fixing others When updating $\mathbf{Q}$ and $\mathbf{Q}'$ by keeping others fixed, we obtain the following sub-problem:

$$\begin{aligned}&\min_{\mathbf{Q},\mathbf{Q}'} \|\mathbf{X}^T\mathbf{W}-\mathbf{Q}\|_F^2+\lambda\|\mathbf{Y}^T\mathbf{W}'-\mathbf{Q}'\|_F^2\\&+\delta\|\mathbf{Q}-\mathbf{Q}'\|_F^2+\eta(\|\mathbf{Q}-\mathbf{L}\|_F^2+\|\mathbf{Q}'-\mathbf{L}\|_F^2)\end{aligned} \tag{9}$$

By setting the derivative of the above objective function w.r.t $\mathbf{Q}$ to zero, we have:

$$\mathbf{Q}=\frac{\mathbf{X}^T\mathbf{W}+\delta\mathbf{Q}'+\eta\mathbf{L}}{1+\delta+\eta} \tag{10}$$

By substituting (10) into (9) and setting the derivative of the objective function w.r.t $\mathbf{Q}'$ to zero, we arrive at

$$\frac{\delta(1+\eta)\mathbf{Q}'-\delta\mathbf{X}^T\mathbf{W}-\delta\eta\mathbf{L}}{1+\delta+\eta}+(\lambda+\eta)\mathbf{Q}'-\lambda\mathbf{Y}^T\mathbf{W}'-\eta\mathbf{L}=0 \tag{11}$$

$$\mathbf{Q}'=\frac{\lambda(1+\delta+\eta)\mathbf{Y}^T\mathbf{W}'+\eta(1+2\delta+\eta)\mathbf{L}+\delta\mathbf{X}^T\mathbf{W}}{\delta(1+\eta)+(\lambda+\eta)(1+\delta+\eta)} \tag{12}$$

Update W and W′ by fixing others When we fix all others except $\mathbf{W}$ and $\mathbf{W}'$, we have the following sub-problem:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{W}'}\quad &\|\mathbf{X}^T\mathbf{W}-\mathbf{Q}\|_F^2+\|\mathbf{Y}^T\mathbf{W}'-\mathbf{Q}'\|_F^2\\&+\beta(tr(\mathbf{W}^T\mathbf{D}\mathbf{W})+tr(\mathbf{W}'^T\mathbf{D}'\mathbf{W}))\end{aligned} \tag{13}$$

By setting the derivative of the above object function w.r.t $\mathbf{W}$ and $\mathbf{W}'$ to zero respectively, we have

$$\mathbf{W}=(\mathbf{X}\mathbf{X}^T+\beta\mathbf{D})^{-1}\mathbf{X}\mathbf{Q} \tag{14}$$

$$\mathbf{W}'=(\lambda\mathbf{Y}\mathbf{Y}^T+\beta\mathbf{D}')^{-1}\mathbf{Y}\mathbf{Q}' \tag{15}$$

Finally, we summarized the algorithm for solving the problem in Algorithm 1. It can be similarly proven that the propose iterative procedure in Algorithm can converge to the optimal solutions [33]. The complexity of calculating the inverse of a few matrices for $\mathbf{W}$ and $\mathbf{W}'$ are $O(d^3)(d \ll n)$ or $O(v^3)(v \ll n)$. The complexity of matrix multiplication is $O(cdn)(c \ll n)$ for $\mathbf{Q}$, $\mathbf{Q}'$,$\mathbf{W}$, and $\mathbf{W}'$. Therefore, the overall computational complexity is and $O(k(d^3 + v^3 + cdn))$, where k is the number of iterations.

Algorithm 1 Learning for MMLC

Input: Matrices $\mathbf{X}$ and $\mathbf{Y}$ of the training images; Parameters β, δ, η.
Output: $\mathbf{W}$, $\mathbf{W}'$.
Initialize $\mathbf{Q}$, $\mathbf{Q}'$, $\mathbf{W}$, $\mathbf{W}'$;
repeat
  Update $\mathbf{D}$ and $\mathbf{D}'$ according to Eq. (8);
  Update $\mathbf{Q}'$ according to Eq. (12);
  Update $\mathbf{Q}$ according to Eq. (10);
  Update $\mathbf{W}$ according to Eq. (14);
  Update $\mathbf{W}'$ according to Eq. (15);
until *Convergence*;

4 Image classification based on MML

Once the multi-modal classifier have been learned, we can then classify the test image I_q based on its visual features $\mathbf{x}_q$ and textual features $\mathbf{y}_q$. Since the output of MML contains two classifiers, both of the two kinds of features can be classified by the two classifiers. There are many methods to combine classification results based on different features. The first ones use the voting method to combine the classification result. They train multiple classifiers on different types of visual features, e.g., texture features, color features, and edge features, etc, and textual features. Then, the class which receives the most votes from the classifiers based on different features is considered as the final class of the test image. These methods handle different visual features and textual features equally, which neglects the heterogeneity and different importance between the visual features and textual features. The second type of methods use the linear combination strategy, which linearly combines the results derived from the two classifiers. With the linear strategy, the classification result of the test image $\mathbf{I}_q$ is obtained as follows:

$$\mathbf{l}_q = \alpha \mathbf{x}_q \mathbf{W} + (1 - \alpha)\mathbf{y}_q \mathbf{W}' \tag{16}$$

where α is a parameter to balance the classification results based on visual features and textual features respectively.

5 Experiments

In this section, the proposed model MML is evaluated for social image classification on two benchmark datasets. We first describe the experimental setup, including the introduction of the two benchmark datasets and the implementation details. Moreover, our classification approach is compared with other closely related methods on the two benchmark datasets.

5.1 Experimental setup

We select two benchmark datasets for performance evaluation. The first dataset is PASCAL VOC'07 [11] that contains around 10,000 images. Each image is annotated by users with a set of tags, and the total number of tags used here is reduced to 804 by the same preprocessing step as [15]. This dataset is organized into 20 classes. Moreover, the second dataset is MIR FLICKR [18] that contains 25,000 images annotated with 457 tags. This dataset is organized into 38 classes. For the PASCAL VOC'07 dataset, we use the standard training/test split, while for the MIR FLICKR dataset we split it into 20,000/5,000 training/test images.

To represent the visual content of social image, we concatenated 64-D color histogram, 144-D color correlogram, 73-D edge direction histogram, 128-D wavelet texture, 225-D block-wise color moments, and 1000-D bag of words to set a 1634-D representation for the visual content of image. Each dimension is mean-centered. As the raw tag list is very sparse, the textual feature vector is high-dimensional and many of the element entries are zero, which may affect the effectiveness of classification on text features as shown in (5). We include a new representation for text content by combining the geographical information. First, Word2Vec [32] is used to represent each tag with a 500-D vector, which is trained on Wikipedia articles and Web news of about 100 million words. Then, all the image tag lists are clustered into 2000 groups, and a 2000-D dictionary is obtained. Finally, each tag is represented by a 2000-D sparse codes learned from the geographical dictionary, and the tag list of each image is represented by a 2000-D feature vector by max pooling all the tags in the list.

In order to assess the performance, we compare our approach with the early fusion approaches, i.e., SMKL [50], LMKL [16], and SBOW [29], and the late fusion approaches, i.e., InfR [13]. The parameters of all these methods are set according to their papers. Besides, we also compare our approach with the deep methods of classification, i.e., DDBN [27] and CDN [21]. Since these two models require the size of the input image to be 20×20, we resize the images in the datasets. Assume the resolution of the image is $m \times n$, we fetch the middle square of $d \times d$ pixels and resample it to 20×20, where $d = min(m, n)$. Then, the deep learning models are constructed according to the two works [21, 27].

In the two datasets, the positive and negative samples are quite unbalanced. Thus, the accuracy criterion is not proper any more, and we use the criterion of average precision (AP) for evaluation. In this paper, AP is the ranking performance computed under each label. Usually, the mean value over all labels, i.e., mAP is reported. For each label, we rank the test images (assume m images) based on the classification scores. For each of the ranked images, we compute the recall value r_i. For each recall value r_i, the greatest precision value p_i is computed, and then the average precision value is computed as:

$$AP = \sum_i \frac{p_i}{m} \tag{17}$$

Finally, mAP is computed by average over the AP values of all the classes.

5.2 Performance analysis

In the first experiment, we compare the performance between our approach and other fusion methods, and the result is shown in Fig. 2. From this figure, several conclusions can be derived. First, the multi-modal classifier learned by the model MML significantly

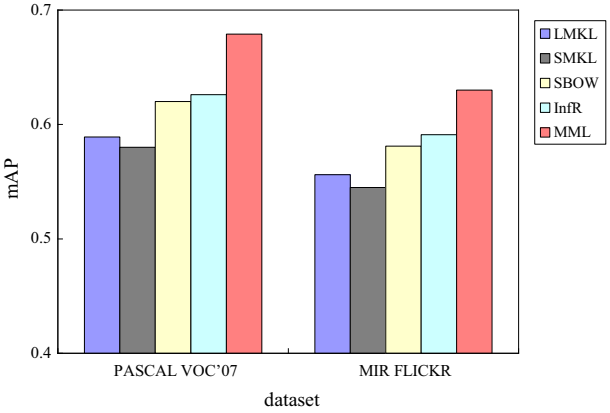


Fig. 2 Performance comparison for different approaches

outperforms other early fusion and late fusion approaches. This is because that we integrate multi-modal features with the latent correlation for social image classification. The classification based on textual features and visual features can be reinforced by each other with the joint model. Second, it shows that the late fusion method outperform the early fusion methods. This is because that textual features and visual features have different effectiveness of reflecting the semantic content of social image, and usually, the textual features are more effective to capture the semantic content. Moreover, social images and the associated text information are of wide diversity. It is difficulty to conduct a suitable latent space to cover such a large diversity and heterogeneity. However, the late fusion method learns the classifier for each kind of features independently, which ignore the latent correlation between different kinds of features. Thus, its performance is also affected.

Second, we assess whether the correlation between different kinds of features is effective for social image classification. To evaluate the performance, we compare our approach with the ensemble classification model [37] which is conducted as follows: it first learns two classifiers based on visual features and textual features using (2) and (3) independently, and then they are combined using a linear strategy with (16). We call the implementation En_Classifier for presentation convenience. Then we compare the performance for MML and En_Classifier with that different percentages of training data are used in Fig. 3. From this figure, several conclusions can be drawn. First, both of the two approaches are affected by the size of training dataset. Second, it can be observed that MML outperforms En_Classifier consistently. Therefore, exploiting the correlation between different kinds of features is helpful to the classifier learning. It can improves the performance of classifier which direct combines the classification results on different types of features.

Then, we compare our approach with the deep learning methods on the two datasets. In this experiment, different percentages of the training dataset are used to evaluate the performance. The performance comparison is shown in Fig. 4. From this figure, we can see that the performance of our approach is better to the deep based methods in some degree. There are mainly two reasons. First, these deep models mainly use the visual content for image classification. They neglect the the plentiful information of text content which are usually more effective to model the semantics of images. Second, the deep model usually need a large number of images to train. Therefore, the performance may be affected by the training dataset.

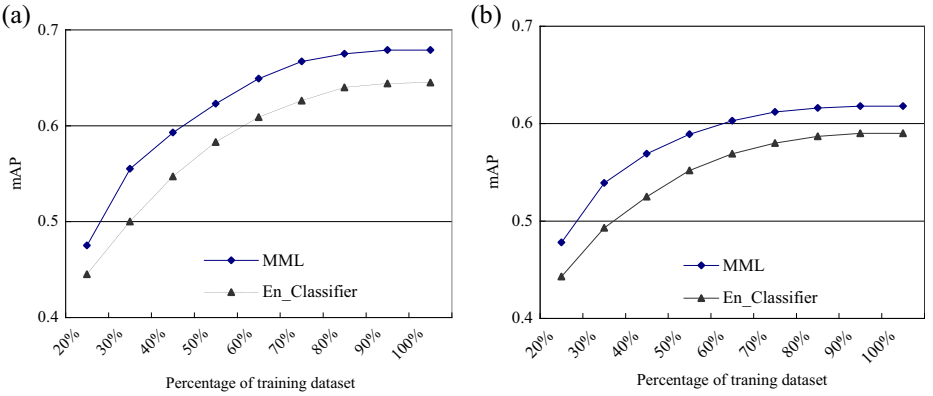


Fig. 3 Effect of feature correlation: (**a**) PASCAL VOC'07, (**b**) MIR FLICKR

5.3 Parameters sensitivity

For our learning model MML, there are four trade-off parameters in the learning of MML, i.e., λ, β, δ, η, and one parameter α in the classification process as shown in (16). In the experiments, all the parameters are selected by cross-validation on the training set, and the parameters β, δ, η are turned in the range {0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000}.

Taking the dataset of PASCAL VOC'07 as a example, the performance of our approach with the parameters set with different values is shown in Fig. 5. According to these results, we can conclude that the approach obtains the best performance when β=0.01, δ=0.01, η=0.1, λ=4, and α=0.4 for the dataset PASCAL VOC'07. From the figure, it can be concluded that the textual features are more important to determine the class labels of social images. This is because that textual tags are usually more effective to model the semantics of social images than visual features. We apply the optimal parameters in the comparison experiments.

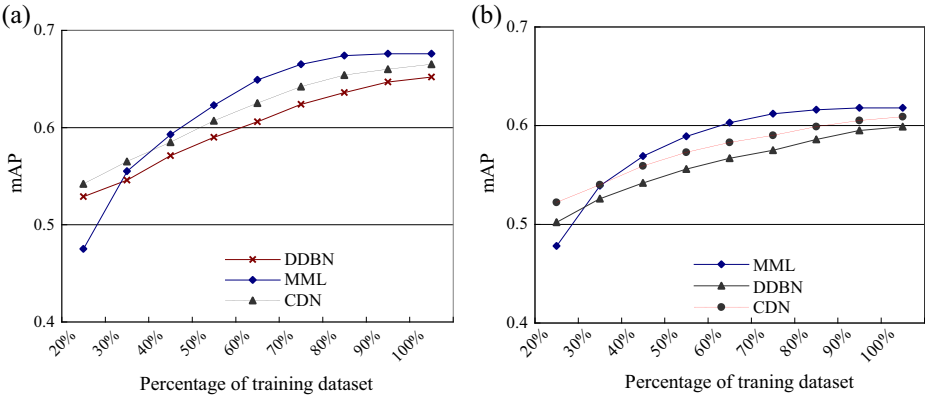


Fig. 4 Performance comparison with the deep methods with different percentages of training dataset used

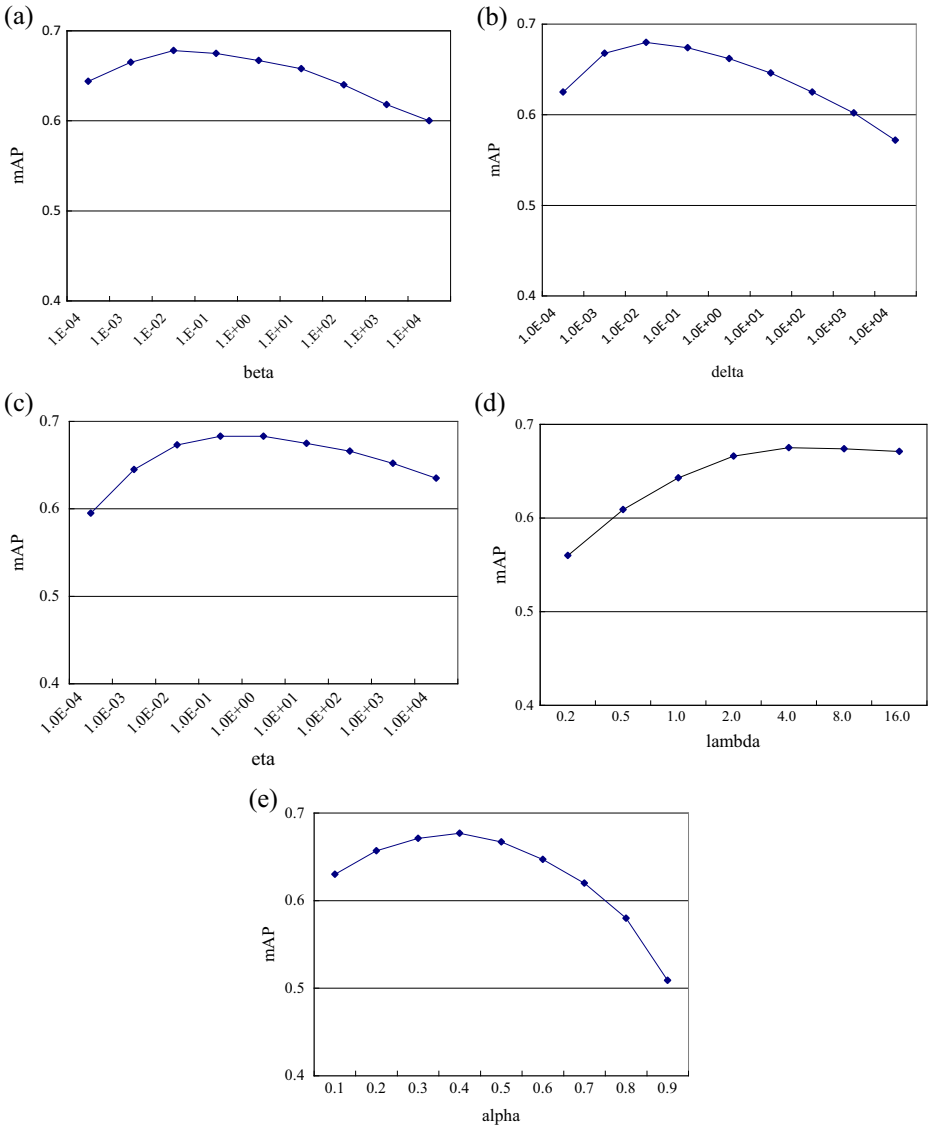


Fig. 5 Effect of parameters

6 Conclusion and future works

With the large amount of social images, it is interesting and challenging to classify these images with multi-modal content, i.e., visual content and textual description. Though there are many works on social image classification. There is still a problem of exploiting the latent correlation between different kinds of features for classifier learning. To address this problem, we have proposed a novel model to capture the correlation among the multi-modal

features, and the classification result is reinforced by different kinds of features. Moreover, the $l_{2,1}$ normalization method is applied to reduce the effect of noisy and redundant features. Then, a efficient optimization algorithm is posed to solve the object function of the model. Experiments on two real-world datasets demonstrate the superiority of our approach.

There are also many potential future extensions of this work. It would be interesting to investigate other social information, like uploader's interests and social relationship, for social image classification. Our approach can also be extended to integrate more than two types of heterogeneous features for social image classification.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (No. 61202239, No. U1636210, U163610106, and No. 61403090), in part by the Fundamental Research Funds for the Central Universities (No. YWF-14-JSJXY-16), and in part by the Fund of the State Key Laboratory of Software Development Environment (No. SKLSDE-2015ZX-11)

References

1. Ayoub et al (2016) Personalized social image organization, visualization, and querying tool using low- and high-level features. IEEE CSE
2. Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In: Proc. Annu. conf. comput. learn. theory, pp 92–100
3. Bo Y, Mei T, Hua X-S, Yang L, Yang S-Q, Li M (2007) Online video recommendation based on multimodal fusion and relevance feedback. In: Proceedings of ACM CIVR conference, pp 73–80
4. Brefeld U, Gartner T, Scheffer T, Wrobel S (2006) Efficient co-regularised least squares regression. In: Proceedings of the 23rd international conference on machine learning. ACM, pp 137–144
5. Brefeld U, Scheffer T (2004) Co-EM support vector learning. In: Proc. int. conf. mach. learn., pp 121–128
6. Cabral R, De la Torre F, Costeira JP, Bernardino A (2015) Matrix completion for weakly-supervised multi-label Image classification. IEEE Trans Pattern Anal Mach Intell 37(1):121–135
7. Chan TH, Jia K, Gao S, Lu J, Zeng Z, Ma Y (2015) PCANet: a simple deep learning baseline for image classification. IEEE Trans Image Process 24(12):5017–5032
8. Clinchant S, Ah-Pine J, Csurka G (2011) Semantic combination of textual and visual information in multimedia retrieval. In: Proceedings of ACM international conference on multimedia retrieval
9. Crampes M et al (2009) Visualizing social photos on a Hasse diagram for eliciting relations and indexing new photos. IEEE TVCG
10. Eklund P et al (2006) An intelligent user interface for browsing and search MPEG-7 images using concept lattices. Inter. LNAI'06 Conf. Springer
11. Everingham M, Van Gool L, Williams C, Winn J, Zisserman A (2007) The PASCAL visual object classes challenge 2007 results. http://www.pascalnetwork.org/challenges/VOC/voc2007
12. Farquhar JDR, Hardoon DR, Meng H, Shawe-Taylor J, Szedmak S (2005) Two view learning: SVM-2K, theory and practice. In: Proc. Adv. neural inf. process. syst., pp 355–362
13. Fazel M (2002) Matrix rank minimization with applications. Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford
14. Ferr S (2007) CAMELIS: organizing and browsing a personal photo collection with a logical information system. In: Proc. of inter. CLA conf.
15. Guillaumin M, Verbeek J, Schmid C (2010) Multimodal semi-supervised learning for image classification. In: Proc. IEEE Conference on computer vision and pattern recognition (CVPR), pp 902–909
16. Hardoon D, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: an overview with application to learning methods. Neural Comput 16(12):2639–2664
17. He R, Tan T, Wang L, Zheng W-S (2012) $l_{2,1}$ regularized correntropy for robust feature selection. In: IEEE Conference on computer vision & pattern recognition, pp 2504–2511
18. Huiskes M, Lew M (2008) The MIR Flickr retrieval evaluation. In: Proc ACM international conference on multimedia information retrieval (MIR), pp 39–43
19. Kloft M, Brefeld U, Sonnenburg S, Zien A (2011) l_p,-norm multiple kernel learning. J Mach Learn Res 12:953–997
20. Krishnapuram B, Williams D, Xue Y, Carin L, Figueiredo MAT, Hartemink A (2004) On semi-supervised classification. In: Proc. adv. neural inf. process. syst., pp 721–728

21. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25(2):2012
22. Lanckriet GRG, Cristianini N, Bartlett P, Ghaoui LE, Jordan MI (2002) Learning the kernel matrix with semidefinite programming. In: Proc. int. conf. mach. learn., pp 323–330
23. Lanckriet GRG, Cristianini N, Bartlett P, Ghaoui LE, Jordan MI (2004) Learning the kernel matrix with semidefinite programming. J Mach Learn Res 5:27–72
24. Li Z, Liu J, Tang J, Lu H (2015) Robust structured subspace learning for data representation. IEEE Trans Pattern Anal Mach Intell 37(10):2085–2098
25. Li Z, Tang J (2015) Weakly supervised deep metric learning for community-contributed image retrieval. IEEE Trans Multimed 17(11):1989–1999
26. Lin Y-Y, Liu T-L, Fuh C-S (2008) Dimensionality reduction for data in multiple feature representations. In: Proc. adv. neural inf process. syst., pp 961–968
27. Liu Y, Zhou S, Chen Q (2011) Discriminative deep belief networks for visual data classification. Pattern Recogn 44(10):2287–2296
28. Liu N, Dellandrea E, Chen L, Zhu C, Zhang Y, Bichot C-E, Bres S, Tellez B (2013) Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme. Comput Vis Image Underst 117(5):493–512
29. Lu Z, Wang L, Wen J-R (2014) Direct semantic analysis for social image classification. In: Proceedings of the twenty-eighth AAAI conference on artificial intelligence
30. Luo Y, Liu T, Tao D, Xu C (2014) Decomposition-based transfer distance metric learning for image classification. IEEE Trans Image Process 23(9):3789–3801
31. Maggiori E, Tarabalka Y, Charpiat G, Alliez P (2016) Convolutional neural networks for large-scale remote-sensing image classification. IEEE Trans Geosci Remote Sens
32. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: Proceedings of the first international conference on learning representations, pp 4089–4114
33. Nie F, Huang H, Cai X, Ding C (2010) Efficient and robust feature selection via joint $l_{2,1}$ -norms minimization. In: Proceedings of the 24th conference on neural information processing systems, pp 1813–1821
34. Nigam K, Ghani R (2000) Analyzing the effectiveness and applicability of co-training. In: Proc. int. conf. inf. knowl. manage., pp 86–93
35. Nikolova M, Ng MK (2005) Analysis of half-quadratic minimization methods for signal and image recovery. SIAM J Sci Comput 27(3):937–966
36. Noord NV, Postma E (2016) Learning scale-variant and scale-invariant features for deep image classification. Pattern Recogn 61:583–592
37. Polikar R (2006) Ensemble based systems in decision making. IEEE Circ Syst Mag Third Quart:21–45
38. Snoek CGM, Worring M, Smeulders AWM (2005) Early versus late fusion in semantic video analysis. In: Proc. 13th Annu. ACM int. conf. multimedia, pp 399–402
39. Suh B, Bederson B (2007) Semi-automatic photo annotation strategies using event based clustering and clothing based person recognition interacting with computers
40. Tao D, Cheng J, Lin X, Yu J (2015) Local structure preserving discriminative projections for RGB-D sensor-based scene classification. Inf Sci. doi:10.1016/j.ins.2015.03.031
41. Tollari S, Glotin H (2007) Web image retrieval on ImagEVAL: evidences on visualness and textualness concept dependency in fusion model. In: Proceedings of ACM CIVR conference, pp 65–72
42. Wang H, Wang J (2014) An effective image representation method using kernel classification. In: 2014 IEEE 26th international conference on tools with artificial intelligence. pp 853–858
43. Wang J, Shi L, Wang H, Meng J, Wang JJY, Sun Q, Gu Y (2016) Optimizing top precision performance measure of content-based image retrieval by learning similarity function. arXiv:1604.06620
44. Wang L, Zhao Z, Su F (2015) Efficient multi-modal hypergraph learning for social image classification with complex label correlations. Neurocomputing 171(C):242–251
45. Wang X, Sun J-T, Chen Z, Zhai CX (2006) Latent semantic analysis for multiple-type interrelated data objects. In: Proceedings of ACM SIGIR conference, pp 236–243
46. Wozniak M, Jackowski K (2009) Some remarks on chosen methods of classifier fusion based on weighted voting. In: Proc. 4th int. conf. hybrid artif intell. syst., pp 541–548
47. Xiao T, Xu Y, Yang K, Zhang J (2015) The application of two-level attention models in deep convolutional neural network for fine-grained image classification. Comput Vis Pattern Recogn:130–160

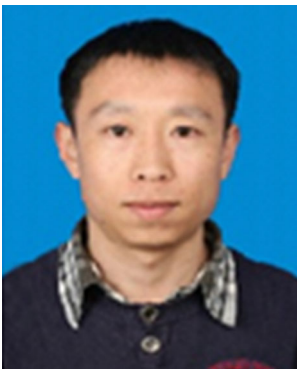
48. Xu C, Tao D, Xu C (2013) A survey on multi-view. Learn Comput Sci
49. Xu C, Tao D, Xu C (2014) Large-margin multi-view information bottleneck. IEEE Trans Pattern Anal Mach Intell 36(8):1559–1572
50. Xu C, Tao D, Xu C (2015) Multi-view intact space learning. IEEE Trans Pattern Anal Mach Intell. doi:10.1109/TPAMI.2015.2417578
51. Yan S, Xu D, Zhang B, Zhang H, Yang Q, Lin S (2007) Graph embedding and extensions: a general framework for dimensionality reduction. PAMI
52. Yang Y, Shen HT, Nie F, Ji R, Zhou X (2011) Nonnegative spectral clustering with discriminative regularization. In: Proceedings of the twenty-fifth AAAI conference on artificial intelligence, pp 555–560
53. Zhou S, Chen Q, Wang X (2013) Convolutional deep networks for visual data classification. Neural Process Lett 38(1):17–27
54. Zhou ZH, Li M (2005) Semi-supervised regression with co-training. In: International joint conference on artificial intelligence (IJCAI)

Xiaoming Zhang was born in Hunan, China, on December 7, 1980. He received the B.Sc. degree, and the M.Sc. degrees in computer science and technology from the National University of Defence Technology, China, in 2003, 2007 respectively. He received his Ph.D degrees in computer science from Beihang University, in 2012. He is currently working at the school of computer, Beihang University, and he has been the lecturer since 2012. His major interests are text mining, image tagging, and TDT.

Xu Zhang received her Master's degree in Computer Science and Technology from Institute of Computing Technology, Chinese Academy of Sciences in 2010. She is currently an engineer in National Computer Network Emergency Response Technical Team of China. Her research interests include machine learning, multimedia analysis and retrieval.

Xiong Li received the PhD degree in pattern recognition and intelligence system from Shanghai Jiao Tong University, China, in 2013. He is currently a senior engineer in National Computer Network Emergency Response Technical Team, China. His research interests include hybrid generative discriminative learning and probabilistic graphical model.

Zhoujun Li received his M.Sc and Ph.D degrees in computer science from the National University of Defence Technology, China, in 1984 and 1999, respectively. He is currently working at the school of computer, Beihang University, and he has been the professor since 2001. His research interests include data mining, information retrieval, and database.



Senzhang Wang was born in Yantai, China, in 1986. He received the M.Sc. degree in Southeast University, Nanjing, China, in 2009. He received his Ph.D degrees in computer science from Beihang University, in 2015. He is currently the associate professor of the School of Nanjing University of Aeronautics and Astronautics, Nanjing, China. His current research interests include data mining and social network analysis. He has publishes more than 10 papers on the famous international journals and conferences, such as KAIS, SIGKDD, AAAI, SDM, DASFAA, etc.