Pattern Recognition Letters 51 (2015) 63–69

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

The classification of multi-modal data with hidden conditional random field ☆

Xinyang Jiang, Fei Wu, Yin Zhang, Siliang Tang*, Weiming Lu, Yueting Zhuang

College of Computer Science and Technology, Zhejiang University, Zhejiang, China

ARTICLE INFO

Article history:
Received 12 November 2013
Available online 9 October 2014

Keywords:
Hidden conditional random field
Latent structure
Multi-modal classification

ABSTRACT

The classification of multi-modal data has been an active research topic in recent years. It has been used in many applications where the processing of multi-modal data is involved. Motivated by the assumption that different modalities in multi-modal data share latent structure (topics), this paper attempts to learn the shared structure by exploiting the symbiosis of multiple-modality and therefore boost the classification of multi-modal data, we call it Multi-modal Hidden Conditional Random Field (M-HCRF). M-HCRF represents the intrinsical structure shared by different modalities as hidden variables in a undirected general graphical model. When learning the latent shared structure of the multi-modal data, M-HCRF can discover the interactions among the hidden structure and the supervised category information. The experimental results show the effectiveness of our proposed M-HCRF when applied to the classification of multi-modal data.

1. Introduction

Nowadays, many real-world applications require the processing of multi-modal data. More and more information gathered from the real world inherently consists of data with different modalities, such as a web image with loosely related narrative text descriptions, and a news article with paired text and images. Therefore, it is desirable to support the classification of multi-modal data. The classification of multi-modal data is very important to many applications of practical interest, for instance, finding some similar data that best describe the rich literal and visual semantics about a topic.

The fundamental challenge dealing with multi-modal data is the appropriate modeling of correlations among multiple modalities. Some of the models like Canonical Correlation Analysis (CCA) [1,18,19,21] map the multi-modal data into one same subspace so that the data from different modalities can be processed together directly. Another kind of models like Gaussian-multinomial Mixture Latent Dirichlet Allocation (GM-LDA) [2] find latent structure among the multi-modal data and use the latent structure to discover the semantics the multi-modal data are sharing. Since the latent structure in multi-modal data bear a strong correlation between low-level features and high-level semantics, it is desirable to appropriately utilize latent structure to boost the semantic understanding of multi-modal data. For example, in the multi-modal document shown in Fig. 1, there is an image of a lion and a paragraph of corresponding description. Obviously, the textual units (e.g., words or sentences) and the visual units (e.g., patches or regions) are both describing several individual aspects of the lion respectively, such as appearance, habitat and biology. As a result, it is important to exploit the hidden structure such as the lion's appearance (i.e. mane and claws) and habitat (i.e. grassland and savanna).

In past years, some approaches have been proposed and achieved great advance in modeling correlations among modalities, such as CCA, GM-LDA, and Dual-wing Harmoniums (DWH) [3]. CCA finds the linear projections that maximally preserve the mutual correlations among multi-modal data. Latent Dirichlet Allocation [4,20] uses a hierarchical Bayesian probability model to discover the topics a document covers and provides a low dimensional embedding representation of each document in terms of topics. GM-LDA extends LDA from single modal data to multi-modal data. In order to obtain a desirable description of the correlations among multi-modal data, GM-LDA assumes that data with different modalities will share same latent topics. Although LDA offers clear semantics and manipulability by modeling the conditional dependence of variables with a directed graphical model, such Bayesian network can be quite expensive to inference due to the dependency among the different layers of hidden variables (note that the hidden topics are conditional independent though). DWH can be seen as an undirected variant of LDA. In order to make the inference easier, DWH assumes conditional independence among the hidden variables. The assumption in DWH makes it possible for the conditional probability of each hidden variable to be calculated

☆ This paper has been recommended for acceptance by A. Petrosino.

* Corresponding author. Tel.: +86 0571 87951853.
E-mail addresses: xinyangj@zju.edu.cn (X. Jiang), wufei@cs.zju.edu.cn (F. Wu), zhangyin98@zju.edu.cn (Y. Zhang), siliang@zju.edu.cn (S. Tang), luwm@zju.edu.cn (W. Lu), yzhuang@zju.edu.cn (Y. Zhuang).

http://dx.doi.org/10.1016/j.patrec.2014.08.005
0167-8655/

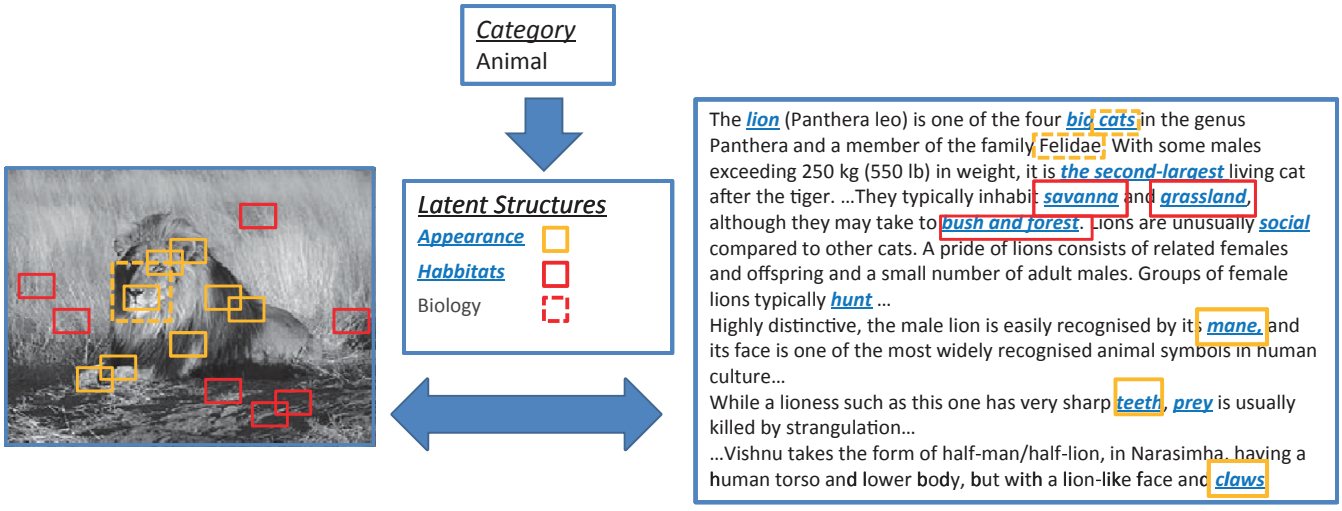

Fig. 1. An example of categorized multi-modal data. The textual units (e.g., words or sentences) and the visual units (e.g., patches or regions) are describing the individual aspects (i.e., latent structure) of the lion respectively, such as appearance, habitat and biology. The areas in the image and the words in the text highlighted by the same color share the same latent topics. For example, the red-colored regions in image and words in text (e.g., big) are used to describe the latent structure "appearance". (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

separately, although it may affect the performance of the model. By the introduction of Markov Random Field, DWH models the structure (topics) among multiple modalities as random vector rather than a random point in a simplex such as the traditional LDA. Although both aforementioned undirected models (e.g., DWH) and directed models (e.g., LDA) have made significant advance in past years, these models have not considered the interactions within the latent structure. As shown in our experiment, modeling such interactions have a positive effect on the performance of multi-modal classification.

Furthermore, both LDA and DWH model a joint probability of all the variables (i.e. category, hidden topics and multi-modal observed data in this paper), which makes it undoubtedly hard to be feasibly implemented. This is because directly modeling such dependence would make the inference intractable. That is why most of the generative models give a simple assumption of the distribution of the observed data, for instance, the conditional independence among all variables. Conditional random field [5] resolves this problem by modeling a conditional probability of the random variables (i.e. category and hidden topics) given the observed data (i.e. multi-modal data) and avoiding modeling the distribution of the observed data. However, it is well-known that models which include latent or hidden-state structure may be more expressive than fully observable models, and the traditional CRF cannot deal with latent variables that cannot be observed in training data, so the useful latent structure among multi-modal data are hard to be discovered by CRF.

We argue that supervised information (e.g. categories) plays a fundamental role to boost the discovery of structure in multi-modal data. A discriminative and supervised model like CRF can utilize category information to discover the hidden structure of multi-modal data in multi-modal classification. Although all these aforementioned unsupervised methods like CCA are proved to be quite effective for finding latent information shared by multi-modal data, they are seldom conducted to the multi-modal classification. A naive and commonly conducted way to use these unsupervised methods in prediction is to obtain the latent representations of the multi-modal data first and then use these representations as input features in a classifier like Support Vector Machine (SVM) [6]. However, in this way, the category information will not be used during the procedure of finding the latent representations. Therefore, we need a supervised model similar to CRF in order to take the category information into consideration. Some variant of LDA like [7] also encodes supervised information in the model, but in most of these models the supervised information is not used to help finding latent representations for multi-modal data. What is more, they still have deficiency we mentioned before, like disregarding of the interactions among hidden topics or only a naive assumption of observed data.

Here we argue that an appropriate utilization of interactions among *model factors* (e.g., categories, latent structure and observed multi-modal) is imperative to boost the performance of multi-modal classification in a supervised learning manner. For example, in Fig. 1, only the highlighted latent topics in latent structure is highly related to the category the image and text both belong to (i.e. appearance and habitats). By discovering the relationship among the categories, latent structure and the observed data, appropriate latent topics can be selected for further classification. To the best of our knowledge, there is no such discriminative probabilistic model that finds latent representations to address the classification of multi-modal data.

As a result, this paper proposes a model that not only discovers the hidden structure of multi-modal data like DWH and LDA, but also utilizes the category information to boost the classification performance like CRF. We call this model Multi-modal Hidden CRF (M-HCRF). M-HCRF is a natural extension of Hidden-state CRF (HCRF) [8,9], which uses hidden variables to discover the relationship between the observed data and the random data. To the best of our knowledge, HCRF has never been used in modeling multi-modal data before this paper. Our proposed M-HCRF extends HCRF to the processing of multi-modal data. Compared to HCRF, M-HCRF not only consider the relationship between the observed data and random data, it also tends to discover the relationship between different observed data modalities. By modeling the relationship between two modalities as the latent structure they share, the proposed M-HCRF can model the interactions among all the observed data modalities as well as the unobserved random variables (i.e. modeling the interactions among the category, the hidden structure and the observed multi-modal data). We hope that in this way, with the help of the category information, M-HCRF can find a more appropriate latent representations of multi-modal data specifically for the classification task.

2. The algorithm of M-HCRF

2.1. Conditional random field

First, we give a brief introduction to the basic conditional random field and hidden state CRF [5] is originally applied for

labeling sequence data. In CRF, a conditional probability distribution often uses an undirected graph to model the dependency among random variables. Let $\mathcal{G}$ be an undirected graph over random variables $\mathbf{y}$ and observed data $\mathbf{x}$. The conditional probability of the random variables $\mathbf{y}$ given $\mathbf{x}$ is:

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp \Phi(\mathbf{y}, \mathbf{x}; \theta)}{\exp \sum_{\mathbf{y}} \Phi(\mathbf{y}, \mathbf{x}; \theta)} \tag{1}$$

where Φ is a potential function parameterized by θ. We denote $C = \{(\mathbf{x}_c, \mathbf{y}_c)\}$ as the set of cliques in $\mathcal{G}$, then the potential function Φ can be defined according to a set of features f_k of each clique:

$$\Phi(\mathbf{y}, \mathbf{x}; \theta) = \sum_k \theta_k f_k(\mathbf{y}_c, \mathbf{x}_c) \tag{2}$$

where the parameters θ are the weights for the features f_k.

Due to the limitation that CRF cannot capture the latent structure of data,the hidden variables $\mathbf{h}$ are introduced into traditional CRF to model the relationship between $\mathbf{y}$ and $\mathbf{x}$. HCRF encodes a conditional probability distribution of both $\mathbf{y}$ and $\mathbf{h}$:

$$p(\mathbf{y}, \mathbf{h}|\mathbf{x}) = \frac{\exp \Phi(\mathbf{y}, \mathbf{h}, \mathbf{x}; \theta)}{\exp \sum_{\mathbf{y}} \Phi(\mathbf{y}_c, \mathbf{h}, \mathbf{x}_c; \theta)} \tag{3}$$

$$\Phi(\mathbf{y}, \mathbf{h}, \mathbf{x}; \theta) = \sum_k \theta_k f_k(\mathbf{y}_c, \mathbf{h}_c, \mathbf{x}_c) \tag{4}$$

2.2. The notations and formulation of M-HCRF

The hidden variables in our model not only describe the relationship between the observed data and the random variables, but also describe the relationship among the multi-modal observed data. In this section, the detailed formulation of M-HCRF will be introduced.

Assume that we have a set of labeled multi-modal training data $\{\mathbf{x}, \mathbf{z}, y\} \in D = \{\mathcal{X}, \mathcal{Z}, \mathcal{Y}\}$, where $\mathbf{x}$ and $\mathbf{z}$ are the data from modalities $\mathcal{X}$ and $\mathcal{Z}$ respectively, y is the corresponding category of $\mathbf{x}$ and $\mathbf{z}$.

As mentioned above, in order to boost the performance of classification, the appropriate modeling of interactions among the context of multi-modal data is fundamental. Here, motivated by the assumption that the data with multiple modalities share some latent structure (topics), we introduce m latent variables $h_i \in \mathbf{h} = \{h_1, h_2, \ldots, h_m\}$ to represent the shared structure and meanwhile serve as a bridge to activate the interactions among model factors. The value of each h_i can be a boolean value indicating if the $\mathbf{x}$ and $\mathbf{z}$ share the ith latent structure, or it can be a real number indicating the level of relevance between the topic and multi-modal data.

Given category information y and the latent structure $\mathbf{h}$, their joint probability (i.e. the conditional probability of y and $\mathbf{h}$ given $\mathbf{x}$ and $\mathbf{z}$) is as follows:

$$P(y, \mathbf{h}|\mathbf{x}, \mathbf{z}, \theta) = \frac{\exp \Phi(y, \mathbf{h}, \mathbf{x}, \mathbf{z}; \theta)}{\sum_{y', \mathbf{h}} \exp \Phi(y', \mathbf{h}, \mathbf{x}, \mathbf{z}; \theta)} \tag{5}$$

where θ are the parameters of the model and Φ is the potential function. The potential function can be defined according to an undirected graphical structure. Arbitrary interaction structure of hidden nodes $\mathbf{h}$ can be devised according to the nature of the problem. One example is shown in Fig. 2. In our experiments, we choose an ergodic structure for hidden nodes to make sure all the hidden nodes has interactions with each other. In M-HCRF, there are three kinds of interactions, namely interactions between category and each hidden variable representing latent structure, the interactions within latent structure itself, and the interactions between each variable node representing latent structure and observed data.

- **Category and hidden structure**: The interactions between category and the hidden variable representing latent structure can be interpreted as the probability that the multi-modal data with certain hidden structure belongs to a certain category.

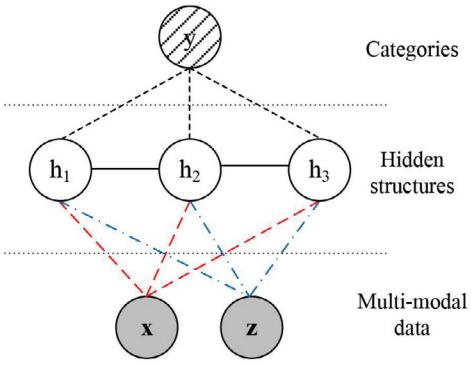

Fig. 2. An intuitive illustration of M-HCRF. There are three kinds of interactions, namely interactions among category and hidden structure, the interactions within latent structure itself, and the interactions among latent structure and observed data. Every interaction between pair of model factors is denoted as one undirected edge. A shaded color node indicates that the node is observed (e.g., multi-modal data $\mathbf{x}$ and $\mathbf{z}$ and the category y). For the simplicity, there are only three hidden nodes. In fact, the dependency among the hidden nodes in this example are in the form of a Markov Chain, in which the value of a node only depends on the value of its precursor and successor. It should be pointed out that the category y is observed in training and un-observed in testing. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- **Hidden structure**: The interactions within hidden structure indicates the relationships among all the hidden variables. Some of the hidden variables may tend to have similar values and some may not. This property of hidden variables is consistent with the property of the latent topics in real world. For example, an article about lions very likely contains both the topics 'habitat' and 'appearance' at the same time.
- **Observed data and hidden structure**: The interactions among observed data and hidden structure can be interpreted as how likely that the multi-modal data share certain hidden structure.

The set of the edges between two hidden nodes is denoted as E_1, and $(h_i, h_j) \in E_1$ signifies the edge between two hidden nodes h_i and h_j. Similarly we denote the set of edges between a hidden node h_j and an output node y as E_2. As a result, the interactions among model factors can be explicitly formulated *via* Φ as follows:

$$\begin{aligned}\Phi(\mathbf{y}, \mathbf{h}, \mathbf{x}, \mathbf{z}) = &\sum_{j=1}^{m} \sum_k f_{1,k}(j, h_j, \mathbf{x}) \theta_{1,k} + \sum_{j=1}^{m} \sum_k f_{2,k}(j, h_j, \mathbf{z}) \theta_{2,k} \\ &+ \sum_{(h_i, h_j) \in E_1} \sum_k f_{3,k}(i, j, h_i, h_j) \theta_{3,k} \\ &+ \sum_{(y, h_j) \in E_2} \sum_k f_{4,k}(j, y_i, h_j) \theta_{4,k}\end{aligned} \tag{6}$$

where the three terms are feature functions respectively: $f_{1,k}$ and $f_{2,k}$ describes the interactions among input $\mathbf{x}$, $\mathbf{z}$ and the hidden nodes $\mathbf{h}$; $f_{3,k}$ describes the interactions between pairs of connected hidden nodes $(h_i, h_j) \in \mathbf{h}$; $f_{4,k}$ describes the interactions between one hidden node $h_i \in \mathbf{h}$ and the category node y. The specific form of $f_{i,k}$ can be defined based on the nature of the tasks the model applied to. Usually the feature function of two variables can be interpreted as a measure of the compatibility of these two variables. For example, for variables with discrete values, $f_{i,k}$ can be a indicator function and only returns 1 when the variables equal to certain values. Another example form of feature function can be the inner-product of certain feature vectors of the two variables, such as $f_{1,k}(j, h_j, \mathbf{x}) = \phi(h_j)\psi(\mathbf{x})$.

2.3. The optimization of M-HCRF

Like most of the probabilistic models, we use maximum likelihood estimation to learn the parameters. Given a training set D with

n labeled multi-modal data $(\mathbf{x}_i, \mathbf{z}_i, y_i) \in D = \{\mathcal{X}, \mathcal{Z}, \mathcal{Y}\}$, the objective function of our model can be written as follows:

$$\begin{aligned} L(\theta) &= \sum_i \log P(y_i|\mathbf{x}, \mathbf{z}) - \frac{1}{2\delta^2}\|\theta\|^2 \\ &= \sum_i \log \sum_{\mathbf{h}} P(y_i, \mathbf{h}|\mathbf{x}, \mathbf{z}) - \frac{1}{2\delta^2}\|\theta\|^2 \end{aligned} \tag{7}$$

where the first term is the likelihood of the training data and the second term is a regularization term of the parameters. The coefficient in regularization term δ is a pre-chosen value, which can be tuned to adjust the tradeoff between the likelihood and the model complexity. The goal is to search for the optimal parameters that maximize the objective function. We can use a quasi-Newton method (e.g., using L-BFGS [10]) to obtain optimal parameters here, since it is commonly used for unconstrained optimization. Because of the introduction of the hidden variables, the optimization of the loss function becomes *non-convex*, so a good initialization of parameters is crucial. Here, we choose several start points and search for the best local optimum.

Next we introduce how to get the derivatives for the optimization algorithm. The derivatives of the likelihood with respect to the parameters $\theta_{1,k}$ of the ith training sample is (note that we omit the regularization term):

$$\begin{aligned} \frac{\partial L_i}{\partial \theta_{1,k}} &= \sum_{\mathbf{h}} P(\mathbf{h}|y_i, \mathbf{x}_i, \mathbf{z}_i, \theta)\frac{\partial \Phi(y_i, \mathbf{h}, \mathbf{x}_i, \mathbf{z}_i; \theta)}{\partial \theta_{1,k}} \\ &\quad - \sum_{y', \mathbf{h}} P(y', \mathbf{h}|\mathbf{x}_i, \mathbf{z}_i, \theta)\frac{\partial \Phi(y', \mathbf{h}, \mathbf{x}_i, \mathbf{z}_i; \theta)}{\partial \theta_{1,k}} \\ &= \sum_{j,a} P(h_j = a|y_i, \mathbf{x}_i, \mathbf{z}_i, \theta) f_{1,k}(j, a, \mathbf{x}_i) \\ &\quad - \sum_{y', j, a} P(h_j = a, y'|\mathbf{x}_i, \mathbf{z}_i, \theta) f_{1,k}(j, a, \mathbf{x}_i) \end{aligned} \tag{8}$$

From the deduction above, we find that the derivatives of $\theta_{1,k}$ can be expressed in terms of $P(h_j = a|\mathbf{x}_i, \mathbf{z}_i)$ and $P(y|\mathbf{x}_i, \mathbf{z}_i)$, both of which can be calculated using existing inference methods. Due to the structure of the model in our experiment, an exact inference method is intractable, so we use loopy belief propagation [11] to obtain one approximative result.

The derivatives with respect to $\theta_{3,k}$ is in the same form of $\frac{\partial L_i(\theta_{1,k})}{\partial \theta_{1,k}}$, just replace the feature function in Eq. (8) with $f_{2,k}$. The derivatives with respect to $\theta_{3,k}$ and $\theta_{4,k}$ are slightly different:

$$\begin{aligned} \frac{\partial L_i}{\partial \theta_{2,k}} &= \sum_{(j,k)\in E, a, b} P(h_j = a, h_k = b|y_i, \mathbf{x}_i, \mathbf{z}_i, \theta) f_{2,k}(j, k, a, b) \\ &\quad - \sum_{y', (j,k)\in E, a, b} P(h_j = a, h_k = b, y'|\mathbf{x}_i, \mathbf{z}_i, \theta) f_{2,k}(j, k, a, b) \end{aligned} \tag{9}$$

$$\begin{aligned} \frac{\partial L_i}{\partial \theta_{1,k}} &= \sum_{j,a} P(h_j = a|y_i, \mathbf{x}_i, \mathbf{z}_i, \theta) f_{3,k}(j, y_i, a) \\ &\quad - \sum_{y', j, a} P(h_j = a, y'|\mathbf{x}_i, \mathbf{z}_i, \theta) f_{3,k}(j, y', a) \end{aligned} \tag{10}$$

both of which can also be calculated with alike approximative inference methods.

The pseudocode of the parameter estimation process is as Algorithm 1.

Algorithm 1 Parameter Estimation for M-HCRF

Input: $\mathcal{X}$ and $\mathcal{Z}$ (multi-modal data); $\mathcal{Y}$ (category information); $\Theta^{(0)}$ (initial model parameters); δ; T (number of cycles)
Output: Θ
for $t = 1, ..., T$ **do**
  for $i = 1, ..., N$ **do**
    Compute the all marginal probability needed for the derivatives with approximative inference methods
    Compute derivatives of i_{th} training samples $\frac{\partial L_i(\theta_{1,k})}{\partial \Theta}$
  end for
  Compute the derivatives of all training samples

$$\frac{\partial L(\theta_{1,k})}{\partial \Theta} = \sum_i \frac{\partial L_i(\theta_{1,k})}{\partial \Theta} - \frac{1}{2\delta^2}\|\theta\|^2 \tag{11}$$

  Update Θ with derivatives
end for

3. Experiments

In this section, we evaluate the performance of our proposed M-HCRF approach. As a running example, we specify the multi-modal data as paired image-text data. We compare our M-HCRF approach with other methods for the classification of multi-modal data. Since the hidden variables (structure) are conducted for the activation of interactions, how the number of hidden variables of M-HCRF representing latent structure affects the performance of classification is also reported.

3.1. Experiment setup

Four datasets are used in this experiment, i.e. PASCAL VOC07 [12], MIR Flickr [13], NUS-WIDE [14] and Wikipedia featured articles [15].

The Pascal VOC'07 dataset (Pascal07) is a benchmark in category recognition and has been commonly used for evaluating multi-modal classification (e.g., [12]). It contains 10,000 images from 20 different categories. In the dataset, SIFT points are extracted from each image and are represented as a feature vector with bag of visual words (BoVW) model. In every experiment. 804 corresponding tags are downloaded from Flickr for each image in the dataset, and are represented as a 804-dimensional feature vector, each of whose dimension indicates if a tag appears. There are 9587 images in the dataset whose user tags are available, which are the image-tag pairs we use in the experiment. Five thousand eleven images and the corresponding tags are sampled for the training set and the rest of the dataset are used as testing set.

MIR Flickr (MIR) is another widely used benchmark for multi-modal retrieval and classification. It contains 25,000 images from 24 categories collected from Flickr. Each image in MIR is represented as a 1000-dimensional feature vectors with BoVW model. The user tags appear at least 50 times are selected, resulting a vocabulary of 457 tags. The corresponding tags of each image are represented as a 458-dimensional feature vector. Seven thousand five hundred images and their corresponding tags are sampled for the training set and the rest of the dataset are used as testing set.

Wikipedia featured articles contains 2886 articles with the corresponding images in each of the articles. All of Wikipedia articles are categorized as one of the 10 concepts. A feature vector is extracted from each article with the bag of words model (BoW). SIFT points are extracted from each image and are represented as a feature vector with bag of visual words (BoVW) model. Due to the fact that the number of positive samples is much smaller than the number of negative samples, in order to make sure both training and test sets have enough number of positive samples, we randomly pick 150 positive samples and 250 negative samples for each category in the training set and the rest samples are for the testing set.

NUS-WIDE contains 26,948 images with 1000 associated tags from Flickr. Each image with corresponding tags has several of the 81 concepts as groundtruth. In order to use this dataset in classification, one concept from the ground-truth is selected for binary classification (does or does not has the certain concept) in every experiment.

Table 1
The performance comparison in terms of F_1 on Pascal VOC'07, MIR Flickr, Wiki featured article and NUS-WIDE. The average performance of five random training/testing sets groups are reported. The results shown in boldface are the best results.

	M-HCRF	SVM	CRF	CCA + SVM	CCA + CRF	Uni-modal model
Pascal07	**0.4943 ± 0.0162**	0.4706 ± 0.0194	0.4398 ± 0.0167	0.4720 ± 0.0192	0.4584 ± 0.0163	0.3134 ± 0.0147
MIR	**0.5110 ± 0.0131**	0.4873 ± 0.0157	0.4697 ± 0.0159	0.4944 ± 0.0143	0.4744 ± 0.0154	0.3635 ± 0.0200
Wiki	**0.9118 ± 0.0162**	0.8408 ± 0.0608	0.7951 ± 0.0217	0.8856 ± 0.0270	0.8473 ± 0.0747	0.7810 ± 0.0263
NUS	**0.2386 ± 0.0092**	0.2270 ± 0.0180	0.2146 ± 0.0298	0.2282 ± 0.0163	0.2073 ± 0.0317	0.1965 ± 0.0205

Similar to the Wikipedia featured articles, a 500-dimensional feature vector based on SIFT BoVW is used to represent each image. The corresponding annotated tags of each image are represented as a 1000-dimensional vector, each dimension of 1000-dimensional vector is a binary indicator to indicate whether a tag appears or not. One thousand positive samples and 1000 negatives samples are sampled to get a balance training set for each category in the experiment. Two thousand images and their corresponding tags are sampled in the rest of the dataset for testing in every experiment. We ran five rounds of experiments on both datasets and the average performance is reported.

We use F_1 score as the evaluation criterion for the classification of multi-modal data. F_1 score considers both precision and recall of a test. Precision is the number of correct results divided by the number of all returned results and recall is the number of correct results divided by the number of results that should have been returned. F_1 is the harmonic mean of precision and recall. The larger the F_1 is, the better the performance is.

3.2. The performance comparison

The compared algorithms with our proposed M-HCRF are listed as follows:

- **SVM**: Given the paired image-text, the visual and textual features are catenated in SVM to learn the classifier without distinguishing their different modality.
- **CRF**: Similar to SVM, both visual and textual features are catenated in CRF without distinguishing modality.
- **CCA + SVM** and **CCA + CRF**: CCA is first performed to map images and texts into a subspace , and we use the mapped features of images and texts in SVM or CRF for classification respectively based on the framework in Ref. [16].
- **Uni-modal model**: A CRF model is trained independently for each of the two modalities and the two models are combined together by multiplying the probabilities computed by each CRF.

Table 1 reports the performance of different algorithms in Pascal07, MIR, NUS-WIDE and Wiki featured articles in terms of F_1. From the results in Table 1, we make the several observations.

Firstly, the proposed M-HCRF achieves the best performance of classification of multi-modal data in term of F_1 score over all the for data sets, thanks to its introduction of latent structure and the interactions between model factors (observed multi-modal data, latent structure and category) .

Secondly, on Pascal07, MIR and NUS-WIDE, preprocessing multi-modal data with CCA before training the classifier boosts the performance of the models (i.e., CCA + SVM and CCA + CRF achieve better performance than SVM and CRF respectfully). This proves that the latent space discovered by CCA on these three datasets is suitable for classification. On the other hand, on Wiki featured articles, the performance achieved by CCA + SVM is almost the same as the performance achieved by SVM, and the performance achieved by CCA + CRF is even worse than the performance of CRF. The reason is that the latent space extracted by CCA is not suitable for classifying the multi-modal data in Wiki featured articles, which makes it hard for classifier to classify the data in the latent space. It shows the importance of finding latent structure suitable for classification (i.e. a discriminative latent space), so it is imperative to utilize category information to find the suitable latent structure.

Thirdly, compared to Pascal and MIR, it is noted that classification performance over Wikipedia featured articles are noticeably high, and the classification performance over NUS is relatively low. This is probably because the images as well as their corresponding tags in NUS dataset are very richer and diverse in semantics, and the semantics in the Wiki dataset is relatively more clear and distinct. We will further verify this conclusion in the next section.

At last, the uni-modal model achieves relatively lower performance compared to all the other models, which are all multi-modal classification algorithms. This is multi-modal classification algorithms simultaneously optimize the model parameters of both modalities in the training process, while the uni-modal model independently optimizes the parameters from two modality, so the uni-modal model does not consider the inter-relationship between two modalities, which is crucial for multi-modal classification.

We also report the performance of algorithms using feature vectors with different dimensions. As shown in Table 2, all the algorithms achieve better performance when using high dimensional features. However, the performance of SVM and CRF is affected significantly by the number of dimensions of the feature vectors. On the other hand, M-HCRF can still achieve relatively good performance when using lower dimension feature vectors. It shows M-HCRF's ability to capture appropriate latent structure from the multi-modal data with much less information.

3.3. The influence of latent structure

In this section, we report how the number of hidden variables representing latent structure affects the classification performance in M-HCRF, and the discriminative ability of the latent space.

As mentioned in the last section, the classification performance on NUS-WIDE and Wikipedia featured articles shows significant difference. We analyze the reason of the performance difference by comparing the influence of the model's latent structure on these two datasets. Table 3 shows the performance of our model with different number of hidden variables over Wiki and NUS. As shown in Table 3, the classification performance by M-HCRF over NUS-WIDE is increasing when the number of hidden variables is increasing, and the best performance is achieved when the number of hidden variables is 3. After that, the performance of M-HCRF over NUS-WIDE is decreasing. On the other hand, the best performance of M-HCRF is achieved when the number of hidden variables is 1 over Wiki featured articles, and then the performance is decreasing. This observation further verifies that images as well as their corresponding tags in NUS-WIDE dataset have much richer and more diverse semantics than the ones in the Wiki dataset, hence more hidden variables are preferred to represent the latent structure shared by images and texts and to capture the correlations between them in NUS-WIDE dataset.

Now we explain why the classification performance peaks when a certain number of hidden variables are introduced in the model and then decreases. From the experiments, we draw following conclusion: if the latent structure is too complex, it may have negative

Table 2
The performance comparison in terms of F_1 scores on Wiki Feature Articles. The performance of algorithms using feature vectors with different number of dimensions is reported. The first row of the table shows the performance of the algorithms using 500-dimensional image feature and 1000-dimensional text feature. The second row of the table shows the performance dimensional text feature. The average performance of five random training/testing sets groups is reported. The results shown in boldface are the best results.

	M-HCRF	SVM	CRF	CCA + SVM	CCA + CRF	Uni-modal model
500-dimensional/1000-dimensional	**0.8545 ± 0.0307**	0.7840 ± 0.0501	0.7201 ± 0.1295	0.7942 ± 0.0298	0.7169 ± 0.0667	0.6823 ± 0.0311
1000-dimensional/5000-dimensional	**0.9118 ± 0.0342**	0.8408 ± 0.0608	0.7951 ± 0.0217	0.8856 ± 0.0270	0.8473 ± 0.0747	0.7810 ± 0.0263

Table 3
The performance comparison in terms of F_1 scores on Wiki featured article and NUS-WIDE when the number of hidden variables representing latent structure is set to different values. The average performance of five random training/testing sets are reported. The results shown in boldface are the best results.

	No. of topics				
	1	2	3	5	10
NUS-WIDE	0.2222 ± 0.03826	0.2239 ± 0.0140	**0.2386 ± 0.0092**	0.2131 ± 0.0951	0.1979 ± 0.0094
Wiki	**0.9118 ± 0.0162**	0.8646 ± 0.0349	0.8448 ± 0.0228	0.8226 ± 0.0148	0.7723 ± 0.0297

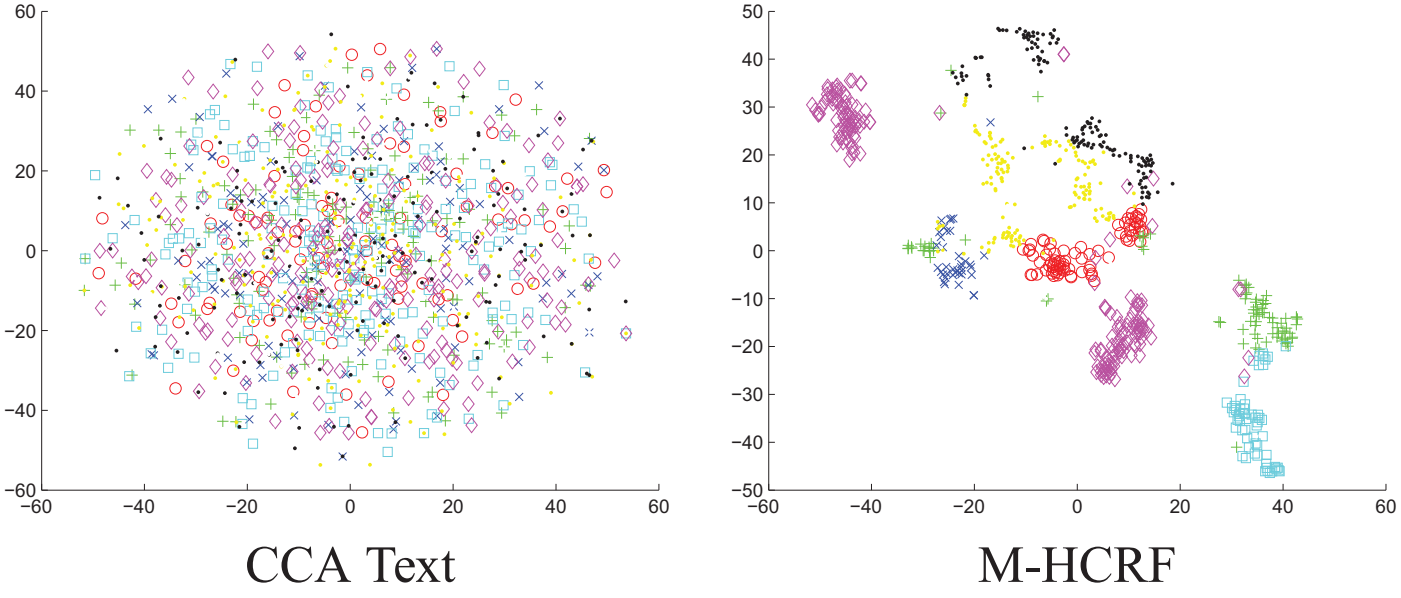

Fig. 3. The joint space discovered by CCA (left), M-HCRF(right). The figure on the left shows text data distribution in the CCA learned latent space. Due to lack of space, we omit the image data distribution because it is very similar to the text data distribution. The figure on the right shows the distribution of the latent structure learned by M-HCRF in the 2-dimensional space. The markers with different colors and shapes represent the data from different categories. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

effects on the performance of the classification. For multi-modal data, it's impossible for every modality to have exactly the same semantics. Some of the semantics the latent structure (topics) represents may not strongly relate to the category of the multi-modal data. A larger number of latent variables leads to more complex latent structure, which makes it more possible that part of the semantics the latent structure contains are irrelevant to the category of the multi-modal data. Obviously, latent structure with too much irrelevant information will affect the classification performance. As a result, as we can see in Table 3, the performance of the model peaks when a certain number of latent variables is introduced in the model, then it starts decreasing. As mentioned in the previous paragraph, multi-modal data in datasets with richer semantics like NUS-WIDE may contain more mutual information, so more latent variables are needed to achieve good classification performance.

Fig. 3 shows the discriminative ability of M-HCRF. We learn latent space of the multi-modal data with both CCA and M-HCRF (the number of hidden variables is set to 10 for M-HCRF) on Wikipedia Featured Articles. Then we use t-SNE [17] to map the learned latent space into a 2-dimensional space. As mentioned in the last section, on Wiki dataset, the latent space learned by CCA is not suitable for the classification. This is further verified by Fig. 3, as the latent space learned by CCA does not present a very good pattern. On the other hand, the joint space discovered by M-HCRF has a very strong grouping pattern, and there are clear margins between data from different categories.

4. Conclusions

This paper proposes an approach called M-HCRF for the classification of multi-modal data in a supervised learning manner. The latent variables are introduced in M-HCRF to capture the interactions among category, latent structure and observed multi-modal data. The model utilizes the category information to find latent representations of the multi-modal data more suitable for classification than unsupervised models. The proposed approach outperforms current state-of-art algorithms for the classification of multi-modal data.

Acknowledgments

This work is supported in part by 973 Program (2010CB327900), NSFC (61103099, 61128007), Zhejiang Provincial Natural Science Foundation of China (LQ13F020001), Chinese Knowledge Center of Engineering Science and Technology (CKCEST).

References

[1] H. Hotelling, Relations between two sets of variates, in: S. Kotz, N. Johnson (Eds.), Breakthroughs in Statistics, Springer Series in Statistics, Springer, 1992, pp. 162–190.
[2] D.M. Blei, M.I. Jordan, Modeling annotated data, in: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, ACM, 2003, pp. 127–134.
[3] E.P. Xing, R. Yan, A.G. Hauptmann, Mining associated text and images with dual-wing harmoniums, in: UAI, AUAI Press, 2005, pp. 633–641.
[4] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[5] J.D. Lafferty, A. McCallum, F.C.N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 282–289.
[6] C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining Knowl. Discov. 2 (1998) 121–167.
[7] C. Wang, D. Blei, F.F. Li, Simultaneous image classification and annotation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009, CVPR 2009, 2009, pp. 1903–1910.
[8] A. Quattoni, M. Collins, T. Darrell, Conditional random fields for object recognition, in: In NIPS, MIT Press, 2004, pp. 1097–1104.
[9] S.B. Wang, A. Quattoni, L. Morency, D. Demirdjian, T. Darrell, Hidden conditional random fields for gesture recognition, in: CVPR, 2006, vol. 2, 2006, pp. 1521–1527.
[10] D. Liu, J. Nocedal, On the limited memory bfgs method for large scale optimization, Math. Programming 45 (1989) 503–528.
[11] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.
[12] M. Everingham, L. Gool, C. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2010) 303–338.
[13] M.J. Huiskes, M.S. Lew, The mir flickr retrieval evaluation, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, MIR '08, ACM, New York, NY, USA, 2008, pp. 39–43.
[14] T.S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '09, ACM, New York, NY, USA, 2009, pp. 48:1–48:9.
[15] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: Proceedings of the International Conference on Multimedia, MM '10, ACM, New York, NY, USA, 2010, pp. 251–260.
[16] S. Ji, L. Tang, S. Yu, J. Ye, Extracting shared subspace for multi-label classification, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, ACM, New York, NY, USA, 2008, pp. 381–389.
[17] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[18] X. Lu, F. Wu, S. Tang, Z. Zhang, X. He, Y. Zhuang, A low rank structural large-margin method for cross-modal ranking, SIGIR 2013 (Full Paper), 2013, 433–442.
[19] F. Wu, X. Lu, Y. Zhang, Z. Zhang, S. Yan, Y. Zhuang, Cross-Media Semantic Representation via Bi-directional Learning to Rank, Proceedings of the 2013 ACM International Conference on Multimedia (ACM Multimedia, Full Paper), 2013, 877–886.
[20] Y. Wang, F. Wu, J. Song, X. Li, Y. Zhuang, Multi-modal Mutual Topic Reinforce Modeling for Cross-media Retrieval, Proceedings of the 2014 ACM International Conference on Multimedia (ACM Multimedia, FULL paper), 2014.
[21] F. Wu, X. Tan, Y. Yang, S. Tang, D. Tao, Y. Zhuang, Supervised Nonnegative Tensor Factorization with Maximum-Margin Constraint, Proceeding of the Twenty-Seventh Conference on Artificial Intelligence (AAAI, Oral Paper), 2013, 962–968.