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/

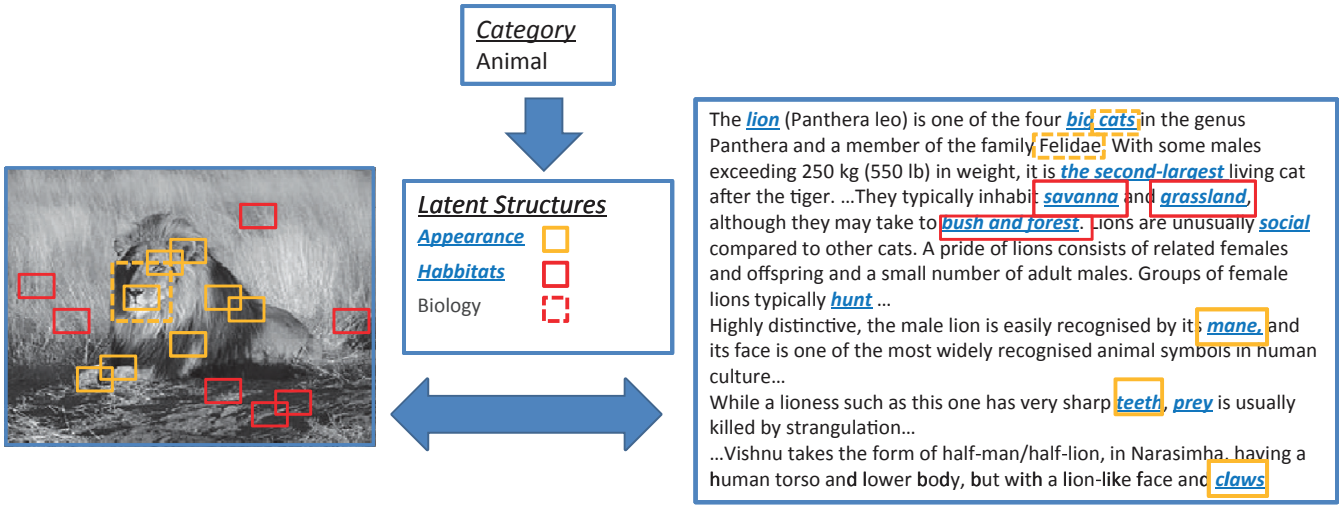

Fig. 1. An example of categorized multi-modal data. The textual units (e.g., words or sentences) and the visual units (e.g., patches or regions) are describing the individual aspects (i.e., latent structure) of the lion respectively, such as appearance, habitat and biology. The areas in the image and 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