Computer Vision and Image Understanding 136 (2015) 69–78

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Multiple-concept feature generative models for multi-label image classification

Minyoung Kim *

Department of Electronics and IT Media Engineering, Seoul National University of Science & Technology, Seoul 139-743, Republic of Korea

ARTICLE INFO

Article history:
Received 15 April 2014
Accepted 18 November 2014
Available online 1 December 2014

Keywords:
Probabilistic graphical models
Multi-label classification
Concept prediction
Image classification

ABSTRACT

We consider the problem of multi-label classification where a feature vector may belong to one of more different classes or concepts at the same time. Many existing approaches are devoted for solving the difficult estimation task of uncovering the relationship between features and active concepts, solely from data without taking into account any sensible functional structure. In this paper, we propose a novel probabilistic generative model that aims to describe the core generative process of how multiple active concepts can contribute to feature generation. Within our model, each concept is associated with multiple representative base feature vectors, which shares the central idea of sparse feature modeling with the popular dictionary learning. However, by dealing with the weight coefficients as exclusive latent random variables encoding contribution levels, we effectively frame the coefficient learning task as probabilistic inference. We introduce two parameter learning algorithms for the proposed model: one based on standard maximum likelihood learning via the expectation–maximization algorithm, the other focusing on maximally separating the margin of the true concept configuration away from the class boundary. In the latter we suggest an efficient approximate optimization method where each iteration admits closed-form update with no line search. For several benchmark datasets mostly from the multi-label image classification, we demonstrate that our generative model with proposed estimators can often yield superior prediction performance to existing methods.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

The multi-label classification is recognized as a very important problem in machine learning and visual pattern recognition. This is because it naturally arises in virtually any application area where data feature can be generated from, or related to, multiple different sources. The examples are, among others, the multi-label image classification where image consists of multiple concepts or objects of different categories therein. In the blind source separation task in signal processing, we aim to find a set of basic components that comprise the observed signal. In the text classification task, the documents often come from multiple different topics, and one wants to find the most relevant topics that are related to the given document.

Unlike the standard single-label classification, the multi-label classification deals with multiple class labels, and aims at finding all active classes that contribute to the observed data. A feature vector may belong to one of more classes at the same time, where the relevant classes can affect generation of the final feature vector in a rather complex way. The inherent difficulty lies in the exponential number of possible class configurations compared to the small number of training data available. Another notable aspect that makes the problem more challenging is that the class labels are often highly sparse, meaning that only a few of them are active for most samples.

In tackling the multi-label classification problem, recent approaches attempt to capture the central relationship between the feature space and the class label space [1,2] with consideration of the label co-occurrences [3,4] or exclusiveness [5] constraints. For detailed summary of the related work, please refer to Section 4. However, many approaches aim at solving the difficult estimation task of uncovering the relationship between features and active concepts, solely from data without taking into account any sensible functional structure. For instance, the latest method in [5] deals with a non-parametric model to find a linear representation that can be shared for both features and class labels. Such a blind estimation without any sensible structure can be difficult and often prone to overfitting.

In this paper, we propose a novel probabilistic generative model that aims to describe the core generative process of how multiple active concepts can contribute to feature generation. The main idea

* Fax: +82 2 970 7903.
E-mail address: mikim21@gmail.com

http://dx.doi.org/10.1016/j.cviu.2014.11.003
1077-3142/© 2014 Elsevier Inc. All rights reserved.

is to confine the model space by a sensible and plausible functional structure to compensate for relatively small number of sparse training data. To achieve this, we first consider to model each concept/class as a set of representative features which we call the base features. Once the active concepts are chosen, each active concept can select one of its base features, together with the significance level indicating its degree of contribution to the final feature vector. Then we form the observed feature as a significance-weighted sum of the selected base features, which can effectively mimic the complex process of multi-class fusion in the observed feature vector.

In this model each concept is associated with multiple representative base feature vectors, while the final feature vector is generated as a significance-weight linear combination of the base features. Similar in nature, this shares the central idea of sparse feature modeling with the popular sparse coding and dictionary learning [6–11]. However, by dealing with the weight coefficients as exclusive latent random variables, the sparsity, in terms of the number of chosen base features, can be enforced at each concept level (much like the exclusive group sparseness [12]), which can be more appealing in certain contexts. Furthermore, the coefficient learning can be effectively framed as a task of probabilistic inference.

In addition, we incorporate the label exclusive group constraints in parameter estimation, a sort of prior knowledge available from pre-inspection of data, whose usefulness was previously verified in several recent work in multi-label or multi-task learning [5,12]. From the observed data, it can be inferred that certain sets of concepts do not occur concurrently, e.g., book vs. truck. This can be learned from data using the recent Graph-Shift algorithm [13] which basically finds the modes of the graph defined with exclusive-labels edges.

There are latent Dirichlet allocation (LDA) like topic models that are also generative models for documents. As topic corresponds to a concept, these models can be used for multi-label classification naturally, where class prediction is done by inference on latent (observed in data in the multi-label setup) topic variables. However, as they model topic-wise multinomial distributions that generate words comprising a document, hence, the features are restricted to be a kind of histogram or any other form based on *frequencies* only. In our model, the features are not limited to histograms, but can be arbitrary as we can learn them in a way like dictionary learning. Also the contribution level of each active concept/topic is unable to be controlled in a principled way like ours.

We introduce two parameter learning algorithms for the proposed model: one based on standard maximum likelihood learning via the expectation–maximization algorithm, the other focusing on maximally separating the margin of the true concept configuration away from the class boundary. In the latter we suggest an efficient approximate optimization method where each iteration reduces to a convex quadratic problem that admits a closed-form update with no line search. In both learning algorithms, we resort to posterior distributions on the latent variables. Our model incorporates fully-connected pairwise random fields over the concept variables, and the probabilistic inference is performed approximately with the loopy belief propagation (LBP) algorithm [14,15].

The paper is organized as follows: In the subsequent section, we formally describe the problem and introduce the notations used in the paper. The proposed model is introduced in Section 2 where the parameter learning algorithms are described in Section 3. After reviewing some related prior work in Section 4, we provide the empirical evaluations on some benchmark image classification datasets in Section 5.

1.1. Problem setup and notations

We denote by $\mathbf{x}$ a d-dim feature vector, which for image data is typically formed by combining one or more popular image features like SIFT [16,17] and color information. Each feature vector $\mathbf{x}$ is associated with multiple class labels (there are K classes in total), indicated by a K-dim 0/1 vector $\mathbf{c}$ whose j-th concept $c_j = 1(0)$ implies the presence (absence) of the concept/class j in $\mathbf{x}$ for $j = 1, \ldots, K$.

Throughout the paper we consider a fully supervised setup, where one is given the n dyadic training samples $\mathcal{D} = \{(\mathbf{c}^i, \mathbf{x}^i)\}_{i=1}^n$. The ultimate goal of the multi-label classification is to estimate the prediction function $\mathbf{c} = \mathbf{f}(\mathbf{x})$ such that the predicted label vector $\mathbf{c}$ for a new test feature $\mathbf{x}$ is as close as possible to the true labels.

2. Multi-concept feature generative model

We propose a probabilistic generative model $P(\mathbf{c}, \mathbf{x})$ that represents the core generative process of how multiple active concepts impact on the generation of a feature vector. To make a sensible model, we will introduce a set of latent random variables that are useful for modeling the generative process. In the below the major concept-feature generation process is described while the detailed formal description follows in Section 2.1.

Our main motivation is that each concept j ($j = 1, \ldots, K$) is characterized by a set of representative feature vectors which we call the *base features*. Instead of having a single representative feature vector, we consider *multiple* base features per concept since a concept usually exhibits several different observations. In image classification, for instance, each object/concept can bear multiple different sub-categories, also with diverse appearances depending on view points and/or pose variations.

The base features of the concept j are denoted by[1] $\mathbf{B}^j = [\mathbf{b}_1^j, \mathbf{b}_2^j, \ldots, \mathbf{b}_M^j]$ where $\mathbf{b}_i^j \in \mathbb{R}^d$. Our modeling intention is that when the j-th concept is active, one of its base features is selected and contributed to form the feature $\mathbf{x}$. To indicate the index of the selected base feature, we hence introduce the hidden random variable, z_j. The prior distribution $P(z_j)$ is modeled as the multinomial over $z_j \in \{1, 2, \ldots, M\}$, where the parameters might be different across j, and can be learned from data. We often use the vector notation $\mathbf{z} = [z_1, z_2, \ldots, z_K]^\top$ to refer to the selection across all K concepts.

The selected base features across the active concepts may have equal effects on the final feature vector. However, it is more reasonable to allow them to have different contributions to the feature generation. In image classification, for instance, certain objects may appear large in an image while others occupy only smaller portions. In such cases one can naturally observe larger impacts of the former features in the overall image descriptor than the latter. To account for it, we consider a $[0, 1]$-scaled significance value to indicate the level of contribution that each concept makes.

More formally, for each active concept j ($c_j = 1$), we decide its significance level s_j, a positive valued (bounded above by 1) latent random variable that encodes the amount of contribution the j-th concept makes in the final feature generation. To make the probabilistic inference computationally feasible, we confine s_j to be discrete, taking one of the S predefined values, $s_j \in \{a_1, a_2, \ldots, a_S\}$, typically we choose $a_i = i/S$. For instance, for $S = 10$ different contribution levels, we get $s_j \in \{0.1, 0.2, \ldots, 1.0\}$. The conditional distribution $P(s_j | c_j = 1)$ can be modeled as the multinomial, whose parameters can be learned from data. We also need to deal with the case of inactive concepts, where a natural choice is to fix it as a delta function at $s_j = 0$, i.e., $P(s_j = 0 | c_j = 0) = 1$, since inactive concepts give zero contribution to the feature generation. Similarly, we use the vector notation $\mathbf{s} = [s_1, s_2, \ldots, s_K]^\top$.

[1] Although one can define different numbers of base features across concepts, we let all concepts have the same number of bases M, for simplicity.

Download English Version:

https://daneshyari.com/en/article/525694

Download Persian Version:

https://daneshyari.com/article/525694

Daneshyari.com