Information Processing and Management 52 (2016) 682–697

Contents lists available at ScienceDirect

Information Processing and Management

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

Probabilistic latent variable models for unsupervised many-to-many object matching

Tomoharu Iwata*, Tsutomu Hirao, Naonori Ueda

NTT Communication Science Laboratories, Japan

ARTICLE INFO

Article history:
Received 3 March 2015
Revised 22 December 2015
Accepted 25 December 2015
Available online 5 February 2016

Keywords:
Latent variable model
Mixture model
Object matching

ABSTRACT

Object matching is an important task for finding the correspondence between objects in different domains, such as documents in different languages and users in different databases. In this paper, we propose probabilistic latent variable models that offer many-to-many matching without correspondence information or similarity measures between different domains. The proposed model assumes that there is an infinite number of latent vectors that are shared by all domains, and that each object is generated from one of the latent vectors and a domain-specific projection. By inferring the latent vector used for generating each object, objects in different domains are clustered according to the vectors that they share. Thus, we can realize matching between groups of objects in different domains in an unsupervised manner. We give learning procedures of the proposed model based on a stochastic EM algorithm. We also derive learning procedures in a semi-supervised setting, where correspondence information for some objects are given. The effectiveness of the proposed models is demonstrated by experiments on synthetic and real data sets.

1. Introduction

Object matching is an important task for finding the correspondence between objects in different domains. Examples of object matching include matching an image with an annotation (Socher & Fei-Fei, 2010), an English word with a French word (Tripathi, Klami, & Virpioja, 2010), and user identification in different databases for recommendation (Li, Yang, & Xue, 2009). Most object matching methods require similarity measures between objects in the different domains, or paired data that contain correspondence information. When a similarity measure is given, we can match objects by finding pairs of objects that maximize the sum of the similarities. When correspondence information is given, we can obtain a mapping function from one domain to another by using supervised learning methods, and then we can calculate the similarities between objects in different domains.

However, similarity measures and correspondences might not be available. Defining similarities and generating correspondences incur considerable cost and time, and they are sometimes unobtainable because of the need to preserve privacy. For example, dictionaries between some languages might not exist, and different online stores cannot share user identification. For these situations, unsupervised object matching methods have been proposed; they include kernelized sorting (Quadrianto, Smola, Song, & Tuytelaars, 2010), least squares object matching (Yamada & Sugiyama, 2011), matching canonical correlation analysis (Haghighi, Liang, Berg-Kirkpatrick, & Klein, 2008), and its Bayesian extension (Klami, 2012; 2013).

* Corresponding author. Tel.: +81 774 93 5161.
E-mail addresses: iwata.tomoharu@lab.ntt.co.jp (T. Iwata), hirao.tsutomu@lab.ntt.co.jp (T. Hirao), ueda.naonori@lab.ntt.co.jp (N. Ueda).

http://dx.doi.org/10.1016/j.ipm.2015.12.013
0306-4573/© 2016 Elsevier Ltd. All rights reserved.

These methods find one-to-one matches. However, matching is not necessarily one-to-one in some applications. For example, when matching English and German documents, multiple English documents with the similar topic could correspond to multiple German documents. In image annotation, related annotations ‘tree’, ‘wood’ and ‘forest’ can be attached to multiple images that look similar to each other. Other limitations of these methods are that the number of domains is limited to two, and the numbers of objects in the different domains must be the same. There can be more than two domains in some applications, for example matching multilingual documents such as English, French and German, and the number of documents for each language can be different.

In this paper, we propose a probabilistic latent variable model for finding correspondence between object clusters in multiple domains without correspondence information. We assume that objects in different domains share a hidden structure, which is represented by an infinite number of latent vectors that are shared by all domains. Each object is generated from one of the latent vectors and a domain-specific linear projection. The latent vectors used for generating objects are unknown. By assigning a latent vector to each object, we can allocate objects in different domains to common clusters, and find many-to-many matches. The number of clusters is automatically inferred from the given data by using a Dirichlet process prior. The proposed model can handle more than two domains with different numbers of objects. We infer the proposed model using a stochastic EM algorithm. The proposed model can ignore arbitrary linear transformations for different domains by inferring the domain-specific linear projection, and can find cluster matching in different domains, where similarity cannot be calculated directly.

The proposed model assumes a Gaussian distribution for each observed variable, and its mean is determined by a latent vector and a linear projection matrix. It is an extension of probabilistic principle component analysis (PCA) (Tipping & Bishop, 1999b) and factor analysis (FA) (Everitt, 1984), which are representative probabilistic latent variable models. With probabilistic PCA and FA, each object is associated with a latent vector. On the other hand, with the proposed model, the latent vector that is assigned to each object is hidden. When the number of domains is one, and every object is assigned to a cluster that is different from those of other objects, the proposed model corresponds to probabilistic principle component analysis.

The proposed model can be also used in a semi-supervised setting, where correspondence information for some objects is given (Jagarlamudi, Juarez, & Daumé III, 2010; Quadrianto et al., 2010). The information assists matching by incorporating a condition stating that the cluster assignments of the corresponding objects must be the same. We derive learning procedures for the semi-supervised setting by modifying the learning procedures for unsupervised setting.

This paper is an extended version of Iwata, Hirao, and Ueda (2013). We newly proposed the inference procedure for a semi-supervised setting, and added derivations and experiments. The remainder of this paper is organized as follows. In Section 2, we review related work. In Section 3, we formulate the proposed model and describe its efficient learning procedures. We also present the learning procedures for a semi-supervised setting and for missing data. In Section 4, we demonstrate the effectiveness of the proposed models with experiments on synthetic and real data sets. Finally, we present concluding remarks and a discussion of future work in Section 5.

2. Related work

2.1. Unsupervised object matching

Unsupervised object matching is a task that involves finding the correspondence between objects in different domains without correspondence information. For example, kernelized sorting (Quadrianto et al., 2010) finds the correspondence by permuting a set to maximize the dependence between two domains where the Hilbert Schmidt Independence Criterion (HSIC) is used as the dependence measure. Kernelized sorting requires the two domains have the same number of objects. Convex kernelized sorting (Djuric, Grbovic, & Vucetic, 2012) is a convex formulation of kernelized sorting. Matching canonical correlation analysis (MCCA) (Haghighi et al., 2008) is another unsupervised object matching method based on a probabilistic model, where bilingual translation lexicons are learned from two monolingual corpora. MCCA simultaneously finds latent variables that represent correspondences and latent vectors so that the latent vectors of corresponding objects exhibit the maximum correlation. Tripathi, Klami, Orešič, and Kaski (2011) also proposed a method for unsupervised object matching that is related to MCCA. These methods assume one-to-one matching of objects. On the other hand, the proposed model can find many-to-many matching, and is applicable to objects in more than two domains. Bayesian solution for MCCA (BMCCA) has been proposed (Klami, 2012; 2013). BMCCA assumes that latent vectors are generated from a Gaussian distribution, and finds one-to-one matching by inferring a permutation matrix. In contrast, the proposed model assumes that latent vectors are generated from an infinite Gaussian mixture model (Rasmussen, 2000), and finds many-to-many matching by inferring cluster assignments.

Manifold alignment is related to the proposed model because they both find latent vectors of multiple sets in a joint latent space. The unsupervised manifold alignment method (Wang & Mahadevan, 2009) finds latent vectors of different domains in a joint latent space in an unsupervised manner. The method first identifies all possible matches for each example by leveraging its local geometry, and then finds an embedding in the latent space. The method requires permutations of the order of the factorial of the size of neighborhoods to match the local geometry. Note that the method does not explicitly find correspondences.