Pattern Recognition Letters 38 (2014) 132–141

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Coupled dimensionality reduction and classification for supervised and semi-supervised multilabel learning ☆

Mehmet Gönen *

Sage Bionetworks, 1100 Fairview Ave N, M1-C108, Seattle, 98109 WA, USA

ARTICLE INFO

Article history:
Received 22 July 2013
Available online 8 December 2013

Keywords:
Multilabel learning
Dimensionality reduction
Supervised learning
Semi-supervised learning
Variational approximation
Automatic relevance determination

ABSTRACT

Coupled training of dimensionality reduction and classification is proposed previously to improve the prediction performance for single-label problems. Following this line of research, in this paper, we first introduce a novel Bayesian method that combines linear dimensionality reduction with linear binary classification for supervised multilabel learning and present a deterministic variational approximation algorithm to learn the proposed probabilistic model. We then extend the proposed method to find intrinsic dimensionality of the projected subspace using automatic relevance determination and to handle semi-supervised learning using a low-density assumption. We perform supervised learning experiments on four benchmark multilabel learning data sets by comparing our method with baseline linear dimensionality reduction algorithms. These experiments show that the proposed approach achieves good performance values in terms of hamming loss, average AUC, macro F_1, and micro F_1 on held-out test data. The low-dimensional embeddings obtained by our method are also very useful for exploratory data analysis. We also show the effectiveness of our approach in finding intrinsic subspace dimensionality and semi-supervised learning tasks.

1. Introduction

Multilabel learning considers classification problems where each data point is associated with a set of labels simultaneously instead of just a single label (Tsoumakas et al., 2009). This setup can be handled by training distinct classifiers for each label separately (i.e., assuming no correlation between the labels). However, exploiting the correlation information between the labels may improve the overall prediction performance. There are two common approaches for exploiting this information: (i) joint learning of the model parameters of distinct classifiers trained for each label (Boutell et al., 2004; Zhang and Zhou, 2007; Sun et al., 2008; Petterson and Caetano, 2010; Guo and Gu, 2011; Zhang, 2011; Zhang et al., 2011) and (ii) learning a shared subspace and doing classification in this subspace (Yu et al., 2005; Park and Lee, 2008; Ji and Ye, 2009; Rai and Daumé, 2009; Ji et al., 2010; Wang et al., 2010; Zhang and Zhou, 2010). In this paper, we are focusing on the second approach.

Dimensionality reduction algorithms try to achieve two main goals: (i) removing the inherent noise to improve the prediction performance and (ii) obtaining low-dimensional visualizations for exploratory data analysis. *Principal component analysis* (PCA) (Pearson, 1901) and *linear discriminant analysis* (LDA) (Fisher, 1936) are two well-known algorithms for unsupervised and supervised dimensionality reduction, respectively.

We can use any unsupervised dimensionality reduction algorithm for multilabel learning. However, the key idea in multilabel learning is to use the correlation information between the labels and we only consider supervised dimensionality reduction algorithms. As an early attempt, Yu et al. (2005) propose a supervised *latent semantic indexing* variant that makes use of multiple labels. Park and Lee (2008) and Wang et al. (2010) modify LDA algorithm for multilabel learning. Rai and Daumé (2009) propose a probabilistic *canonical correlation analysis* method that can also be applied in semi-supervised settings. Ji et al. (2010) and Zhang and Zhou (2010) formulate multilabel dimensionality reduction as an eigenvalue problem that uses input features and class labels together.

For supervised learning problems, dimensionality reduction and prediction steps are generally performed separately with two different target functions, leading to low prediction performance. Hence, coupled training of these two steps may improve the overall system performance. Biem et al. (1997) propose a multilayer perceptron variant that performs coupled feature extraction and classification. Coupled training of the projection matrix and the classifier is also studied in the framework of support vector machines by introducing the projection matrix into the optimization problem solved (Chapelle et al., 2002; Pereira and Gordon, 2006). Gönen and Alpaydın (2010) introduce the same idea to a localized

☆ This paper has been recommended for acceptance by J. Yang.

* Tel.: +1 206 724 7461.
E-mail address: mehmet.gonen@sagebase.org

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.11.021

multiple kernel learning framework to capture multiple modalities that may exist in the data. There are also metric learning methods that try to transfer the neighborhood in the input space to the projected subspace in nearest neighbor settings (Goldberger et al., 2005; Globerson and Roweis, 2006; Weinberger and Saul, 2009). Sajama and Orlitsky (2005) use mixture models for each class to obtain better projections, whereas Mao et al. (2010) use them on both input and output data. The resulting projections found by these approaches are not linear and they can be regarded as manifold learning methods. Yu et al. (2006) propose a supervised probabilistic PCA and an efficient solution method, but the algorithm is developed only for real outputs. Rish et al. (2008) formulate a supervised dimensionality reduction algorithm coupled with generalized linear models for binary classification and regression, and maximize a target function composed of input and output likelihood terms using an iterative algorithm.

In this paper, we propose novel supervised and semi-supervised multilabel learning methods where the linear projection matrix and the binary classification parameters are learned together to maximize the prediction performance in the projected subspace. We make the following contributions: In Section 2, we give the graphical model of our approach for supervised multilabel learning called *Bayesian supervised multilabel learning* (BSML) and introduce a deterministic variational approximation for inference. Section 3 formulates our two variants: (i) BSML with *automatic relevance determination* (BSML + ARD) to find intrinsic dimensionality of the projected subspace and (ii) *Bayesian semi-supervised multilabel learning* (BSSML) to make use of unlabeled data. In Section 4, we discuss the key properties of our algorithms. Section 5 tests our algorithms on four benchmark multilabel data sets in different settings.

2. Coupled dimensionality reduction and classification for supervised multilabel learning

Performing dimensionality reduction and classification successively (with two different objective functions) may not result in a predictive subspace and may have low generalization performance. In order to find a better subspace, coupling dimensionality reduction and single-output supervised learning is previously proposed (Biem et al., 1997; Chapelle et al., 2002; Goldberger et al., 2005; Sajama and Orlitsky, 2005; Globerson and Roweis, 2006; Pereira and Gordon, 2006; Yu et al., 2006; Rish et al., 2008; Weinberger and Saul, 2009; Gönen and Alpaydın, 2010; Mao et al., 2010). We should consider the predictive performance of the target subspace while learning the projection matrix. In order to benefit from the correlation between the class labels in a multilabel learning scenario, we assume a common subspace and perform classification for all labels in that subspace using different classifiers for each label separately. The predictive quality of the subspace now depends on the prediction performances for multiple labels instead of a single one.

Fig. 1 illustrates the probabilistic model for multilabel binary classification with a graphical model and its distributional assumptions. The data matrix $\mathbf{X}$ is used to project data points into a low-dimensional space using the projection matrix $\mathbf{Q}$. The low-dimensional representations of data points $\mathbf{Z}$ and the classification parameters $\{\mathbf{b},\mathbf{W}\}$ are used to calculate the classification scores. Finally, the given class labels $\mathbf{Y}$ are generated from the auxiliary matrix $\mathbf{T}$, which is introduced to make the inference procedures efficient (Albert and Chib, 1993). We formulate a variational approximation procedure for inference in order to have a computationally efficient algorithm.

The notation we use throughout the manuscript is given in Table 1. The superscripts index the rows of matrices, whereas the subscripts index the columns of matrices and the entries of vectors. As short-hand notations, all priors in the model are denoted by $\boldsymbol{\Xi}=\{\boldsymbol{\lambda},\boldsymbol{\Phi},\boldsymbol{\Psi}\}$, where the remaining variables by $\boldsymbol{\Theta}=\{\mathbf{b},\mathbf{Q},\mathbf{T},\mathbf{W},\mathbf{Z}\}$ and the hyper-parameters by $\boldsymbol{\zeta}=\{\alpha_\lambda,\beta_\lambda,\alpha_\phi,\beta_\phi,\alpha_\psi,\beta_\psi\}$. Dependence on $\boldsymbol{\zeta}$ is omitted for clarity throughout the manuscript. $\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. $\mathcal{G}(\cdot;\alpha,\beta)$ denotes the gamma distribution with the shape parameter α and the scale parameter β. $\delta(\cdot)$ denotes the Kronecker delta function that returns 1 if its argument is true and 0 otherwise.

2.1. Inference using variational approximation

The variational methods use a lower bound on the marginal likelihood using an ensemble of factored posteriors to find the joint parameter distribution (Beal, 2003). Assuming independence between the approximate posteriors in the factorable ensemble can be justified because there is not a strong coupling between our model parameters. We can write the factorable ensemble approximation of the required posterior as

$$p(\boldsymbol{\Theta},\boldsymbol{\Xi}|\mathbf{X},\mathbf{Y})\approx q(\boldsymbol{\Theta},\boldsymbol{\Xi})=q(\boldsymbol{\Phi})q(\mathbf{Q})q(\mathbf{Z})q(\boldsymbol{\lambda})q(\boldsymbol{\Psi})q(\mathbf{b},\mathbf{W})q(\mathbf{T})$$

and define each factor in the ensemble just like its full conditional distribution:

$$q(\boldsymbol{\Phi})=\prod_{f=1}^{D}\prod_{s=1}^{R}\mathcal{G}\left(\phi_s^f;\alpha(\phi_s^f),\beta(\phi_s^f)\right),$$

$$q(\mathbf{Q})=\prod_{s=1}^{R}\mathcal{N}(\mathbf{q}_s;\mu(\mathbf{q}_s),\Sigma(\mathbf{q}_s)),$$

$$q(\mathbf{Z})=\prod_{i=1}^{N}\mathcal{N}(\mathbf{z}_i;\mu(\mathbf{z}_i),\Sigma(\mathbf{z}_i)),$$

$$q(\boldsymbol{\lambda})=\prod_{o=1}^{L}\mathcal{G}(\lambda_o;\alpha(\lambda_o),\beta(\lambda_o)),$$

$$q(\boldsymbol{\Psi})=\prod_{o=1}^{L}\prod_{s=1}^{R}\mathcal{G}\left(\psi_o^s;\alpha(\psi_o^s),\beta(\psi_o^s)\right),$$

$$q(\mathbf{b},\mathbf{W})=\prod_{o=1}^{L}\mathcal{N}\left(\begin{bmatrix}b_o\\ \mathbf{w}_o\end{bmatrix};\mu(b_o,\mathbf{w}_o),\Sigma(b_o,\mathbf{w}_o)\right),$$

$$q(\mathbf{T})=\prod_{o=1}^{L}\prod_{i=1}^{N}\mathcal{TN}\left(t_i^o;\mu(t_i^o),\Sigma(t_i^o),\rho(t_i^o)\right),$$

where $\alpha(\cdot)$, $\beta(\cdot)$, $\mu(\cdot)$, and $\Sigma(\cdot)$ denote the shape parameter, the scale parameter, the mean vector, and the covariance matrix for their arguments, respectively. $\mathcal{TN}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma},\rho(\cdot))$ denotes the truncated normal distribution with the mean vector $\boldsymbol{\mu}$, the covariance matrix $\boldsymbol{\Sigma}$, and the truncation rule $\rho(\cdot)$ such that $\mathcal{TN}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma},\rho(\cdot))\propto\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ if $\rho(\cdot)$ is true and $\mathcal{TN}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma},\rho(\cdot))=0$ otherwise.

We choose to model projected data instances explicitly (i.e., not marginalizing out them) and independently (i.e., assuming a distribution independent of other variables) in the factorable ensemble approximation in order to decouple the dimensionality reduction and classification parts. By doing this, we achieve to obtain update equations for $\mathbf{Q}$ and $\{\mathbf{b},\mathbf{W}\}$ independent of each other.

We can bound the marginal likelihood using Jensen's inequality: