Pattern Recognition Letters 89 (2017) 18–24

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

A probabilistic multi-label classifier with missing and noisy labels handling capability

Amirhossein Akbarnejad, Mahdieh Soleymani Baghshah*

Department of Computer Engineering, Sharif University of Technology, Azadi Avenue, Tehran 1458889694, Iran

ARTICLE INFO

Article history:
Received 15 May 2016
Available online 3 February 2017

MSC:
41A05
41A10
65D05
65D17

Keywords:
Multi-label classification
Probabilistic model
Missing labels
Label-space dimension reduction

ABSTRACT

Multi-label classification with a large set of labels is a challenging task. Label-Space Dimension Reduction (LSDR) is the most popular approach that addresses this problem. LSDR methods project the high-dimensional label vectors onto a low-dimensional space that can be predicted from the feature space. Many LSDR methods assume that the training data provide complete label vector for all training samples while this assumption is usually violated particularly when label vectors are high dimensional. In this paper, we propose a probabilistic model that has an effective mechanism to handle missing and noisy labels. In the proposed Bayesian network model, a set of auxiliary random variables, called *experts*, are incorporated to provide robustness to missing and noisy labels. Variational inference is utilized to find the desired probabilities in this model. The proposed approximate inference is highly parallelizable and can be implemented efficiently. Experiments on real-world datasets show that our method outperforms state-of-the-art multi-label classifiers by a large margin.

1. Introduction

Unlike the traditional single-label classification in which each instance can have only one label, instances can be associated with a set of labels in multi-label classification tasks. For example, in automated image annotation, an image can have the tags '*jungle*', '*mountain*' and *'lake'*, or in the text categorization, a text document can be associated with both of the tags '*scientific*' and '*biomedical*'. Many real-world problems can be formulated as a multi-label classification, e.g., multi-label text classification [14], automated image annotation [15], and protein function prediction [1].

A simple approach for multi-label classification is to learn an independent classifier for each label called *Binary Relevance (BR)* [11] that is a fast and simple method. However, in the multi-label classification, there is a useful property that BR does not utilize and thus it can't tackle with challenges of multi-label classification properly. In fact, although some labels are highly correlated, BR assumes that labels are independent. Thus, it cannot tackle with the challenge referred in the literature as *missing labels, weak labels*, or *partially labeled data*. This challenge is related to the incompleteness and/or incorrectness of the set of assigned labels to the training data. In fact, in multi-label datasets, labels are often obtained through crowd sourcing or crawling the web pages. Thus, the associated labels to an instance can be incomplete or incorrect. Indeed, although an image or a text document can be associated with many labels, labelers may provide only a subset of them and perhaps incorrectly assign a label to an instance. We refer to the former problem, i.e. incomplete label assignments, as missing labels and the latter problem, i.e. incorrect label assignments, as noisy labels. Many of the state-of-the-art multi-label classification methods have no explicit mechanism to handle missing or noisy labels and assume that datasets provide flawless label assignments [8,18,20]. Note that the linear dimensionality reduction of the label space used in the LSDR methods like FaIE [18], PLST [20], and CPLST [8] makes these methods almost robust to noisy labels. On the other hand, some of the existing methods [3,7,15] can handle missing labels, but they assume that the training set provides correct label assignments for an instance. The proposed probabilistic method takes advantage of LSDR methods to model label correlations and to deal with a large set of labels, while having an effective mechanism, called Expert Ensemble with an Overriding Expert (EEOE), to handle missing labels, as well as noisy ones. In fact, the proposed method handles missing and noisy labels by introducing auxiliary random variables, named *experts*. EEOE is a general Bayesian network framework that seems suitable to also other settings in which the problem of missing and/or noisy tags, rates, or etc arises (e.g., in recommender systems). In the proposed method, an approximate inference is done by mean-field variational inference. Update iterations obtained by the variational inference on our model are highly parallelizable and thus the inference

* Corresponding author.
E-mail address: soleymani@sharif.edu (M.S. Baghshah).

http://dx.doi.org/10.1016/j.patrec.2017.01.022

can be implemented efficiently. Experiments on some real-world datasets show that the proposed method outperforms state-of-the-art multi-label classifiers by a large margin.

1.1. Notation

Having N instances in the training set, we denote representation of the nth instance in the feature space, latent space, and label space by X_n, C_n and Y_n respectively, where $X_n \in \mathbb{R}^F$, $C_n \in \mathbb{R}^L$, and $Y_n \in \{0, 1\}^K$ (F, L, and K show the number of dimensions of the input, the latent, and the label space respectively). Let $\mathbf{X} \in \mathbb{R}^{N\times F}$, $\mathbf{C} \in \mathbb{R}^{N\times L}$, and $\mathbf{Y} \in \{0, 1\}^{N\times K}$ be the three matrices that the nth row of them is shown by X_n, C_n and Y_n respectively . We denote the maximum value of an index by capital letters (e.g. N, F and L) and show the index of a sample by lower case letters. For example, we show the ℓth dimension of X_n by $X_{n\ell}$. Let $\mathbf{D} \in \mathbb{R}^{K\times L}$ be a matrix that maps the latent space to the label space, i.e. $\mathbf{D}C_n \approx Y_n$. We denote the kth row of the matrix $\mathbf{D}$ by D_k. The sigmoid function is shown as $\sigma(\bullet)$.

2. Related work

2.1. Handling missing and noisy labels

To tackle with missing labels, many methods have been proposed recently. Some of these methods [7,15] assume that the complete label-set for an instance is a latent variable that is related to both the feature space and the available incomplete label space. These methods make simplifying assumptions about the relationship between the complete and the incomplete label set. Although these assumptions make these methods fast, their prediction performance can be degraded accordingly. LCML [3] assumes that elements of the label matrix $\mathbf{Y}$ are of three types: 0, 1, and unknown. It considers unknown elements as latent variables. Although this method makes no simplifying assumption about the relationship between complete and incomplete label vectors, actually in many real-world datasets, the elements of the label matrix $\mathbf{Y}$ are of two types: unknown or 1. Some other methods [19] tend to generate all the ones in the label matrix $\mathbf{Y}$ and give less emphasis to the correct prediction of zero elements of the label matrix. This approach is identical to that of CTR [22], and is shown to be beneficial in the context of recommender systems. Another related method to ours is RELIAB [16] which models the multi-label predictor by a simple maximum entropy model. To handle missing and noisy labels, we propose a probabilistic framework that uses an ensemble of experts with one distinguished expert which can override the vote of other experts. A related method to ours is MPU [15] whose generative process is as follows: Given a representation of instances in the feature space (i.e. the feature matrix $\mathbf{X}$), MPU generates unobserved complete label-sets. Then, it removes some labels from the complete label-set to generate the observed (incomplete) label-set. However, since MPU [15] only removes some labels from the complete label-set, implicitly assumes that all the provided label assignments are correct. However, in the proposed method, we can handle both missing and noisy labels. Furthermore, instead of simply generating the complete label assignment using a back-projection from the latent space (e.g. $Y_{nk} \sim \mathcal{N}(D_k^T C_n, \bullet)$), we assume that more proper labels are missed with lower probability.

2.2. Handling many unique labels, modeling label correlations

The common approaches proposed to tackle the problem of dealing with a large set of labels are as follows: (1) ML-CSSP [2] selects a small subset of class labels that can approximately span the original label space. (2) Some methods divide the learning task into simpler independent multi-class problems [12]. (3) LSDR methods using a dimensionality reduction approach can address the problem of having a large number of labels and can also model label correlations as well. Therefore, in the proposed method, we also use the same dimensionality reduction approach.

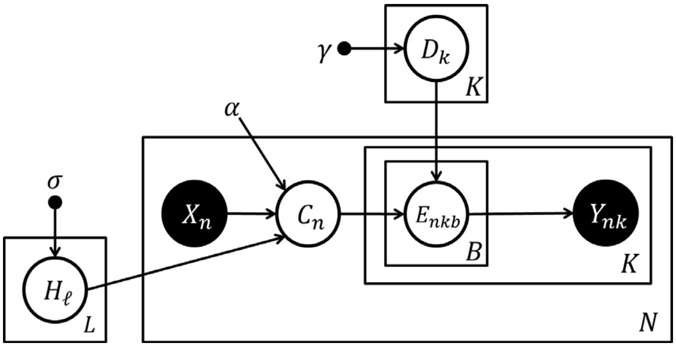

Fig. 1. A graphical model for the proposed method.

3. Proposed method

3.1. The model

In this section, we introduce the proposed Probabilistic Multi-Label Classification (PMLC) method. Fig. 1 illustrates the proposed graphical model for the multi-label classification problem. We denote representations of the nth instance in the feature space and the latent space by $X_n \in \mathbb{R}^F$ and $C_n \in \mathbb{R}^L$ respectively. The ℓth dimension of C_n can be determined by the input X_n and the regression function H_ℓ as $C_{n\ell} \approx H_\ell(X_n)$. Each regression function is generated from a Gaussian process prior. The matrix $\mathbf{D} \in \mathbb{R}^{K\times L}$ corresponds to the matrix that maps the latent space to the label space in the LSDR methods. In other words, we will use $\sigma(D_k^T C_n)$ as a measure showing suitability of the kth label for the nth instance. Inspired by the generalized cross-entropy loss function introduced in [10], we extend this measure of suitability, i.e. $\sigma(D_k^T C_n)$, and introduce a measure that is more proper when we encounter missing labels. In fact, we generate B experts $\{E_{nkb}\}_{b=1}^{B}$ where each expert independently votes on or against the association of the kth label with the nth instance. Note that E_{nkb} is a binary random variable. To aggregate the votes, we simply average these votes while letting the first expert (E_{nk1}) the ability to veto the association of the labels with this instance. More precisely, we aggregate the votes by the equation: $Y_{nk} = E_{nk1}\left(\sum_{b=1}^{B} E_{nkb}/B\right)$. As we will discuss in Section 3.2, this framework that uses an expert ensemble with the overriding capability for one of the experts can handle missing labels effectively. The generative process of the proposed model is as follows:

1. Draw L regression functions $H_\ell \sim GP(\mathbf{0}, \kappa(\bullet, \bullet; \sigma))$.
2. Draw rows of $\mathbf{D}$ as $D_k \sim \mathcal{N}(\mathbf{0}, \frac{1}{2\gamma} I_{L\times L})$.
3. For the nth instance and the kth label:
   a. Draw the latent space representations:
   $$C_{n\ell} \sim \mathcal{N}(H_\ell(X_n), \frac{1}{2\alpha})$$
   b. Draw B experts: $E_{nkb} \sim Bernoulli(\sigma(\lambda D_k^T C_n))$
   $$1 \le b \le B$$
   c. Generate Y_{nk} deterministically:
   $$Y_{nk} = E_{nk1} \frac{\sum_{b=1}^{B} E_{nkb}}{B}$$

Based on the range of values in the $\mathbf{D}$ and $\mathbf{C}$ matrices, it is better to use a scaled version of the sigmoid function. Thus, instead of