Expert Systems With Applications 91 (2018) 263–269

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Multi-label classification using hierarchical embedding

Vikas Kumar[a,*], Arun K. Pujari[a,b], Vineet Padmanabhan[a], Sandeep Kumar Sahu[a], Venkateswara Rao Kagita[a]

[a] Artificial Intelligence Lab, School of Computer & Information Sciences, University of Hyderabad, Hyderbad-500046, AndhraPradesh, India
[b] Central University of Rajasthan, Rajasthan, India

ARTICLE INFO

Article history:
Received 11 May 2017
Revised 25 August 2017
Accepted 9 September 2017
Available online 11 September 2017

Keywords:
Multi-label learning
Matrix factorization
Label correlation

ABSTRACT

Multi-label learning is concerned with the classification of data with multiple class labels. This is in contrast to the traditional classification problem where every data instance has a single label. Multi-label classification (MLC) is a major research area in the machine learning community and finds application in several domains such as computer vision, data mining and text classification. Due to the exponential size of the output space, exploiting intrinsic information in feature and label spaces has been the major thrust of research in recent years and use of parametrization and embedding have been the prime focus in MLC. Most of the existing methods learn a single linear parametrization using the entire training set and hence, fail to capture nonlinear intrinsic information in feature and label spaces. To overcome this, we propose a piecewise-linear embedding which uses maximum margin matrix factorization to model linear parametrization. We hypothesize that feature vectors which conform to similar embedding are similar in some sense. Combining the above concepts, we propose a novel hierarchical matrix factorization method for multi-label classification. Practical multi-label classification problems such as image annotation, text categorization and sentiment analysis can be directly solved by the proposed method. We compare our method with six well-known algorithms on twelve benchmark datasets. Our experimental analysis manifests the superiority of our proposed method over state-of-art algorithm for multi-label learning.

1. Introduction

The objective of multi-label classification is to build a classifier that can automatically tag an example with the most relevant subset of labels. This problem can be seen as a generalization of the *single label* classification where an instance is associated with a unique class label from a set of disjoint labels $\mathcal{L}$. Formally, given n training examples in the form of a pair of feature matrix $\mathcal{X}$ and label matrix $\mathcal{Y}$ where each example $x_i \in \mathbb{R}^d, 1 \leq i \leq n$, is a row of $\mathcal{X}$ and its associated labels $Y_i \in \{\pm 1\}^{\mathcal{L}}$ is the corresponding row of $\mathcal{Y}$, the task of multi-label classification is to learn a parameterization $h : \mathbb{R}^d \rightarrow \{\pm 1\}^{\mathcal{L}}$ that maps each instance to a set of labels. Recent years have witnessed extensive applications of multi-label classification in machine learning (Read, Pfahringer, Holmes, & Frank, 2009; Zhang & Zhou, 2006), in computer vision (Boutell, Luo, Shen, & Brown, 2004; Cabral, Torre, Costeira, & Bernardino, 2011), and in data mining (Schapire & Singer, 2000; Tsoumakas & Vlahavas, 2007).

Existing methods of multi-label classification can be broadly divided into two categories (Sorower, 2010; Zhang & Zhou, 2014) - methods based on problem transformation and methods based on algorithm adaptation. Former approach transforms the multi-label classification problem into single label classification problems so that existing single-label classification algorithms can be applied. During the last decade, a number of problem transformation techniques are proposed in the literature such as Binary Relevance (BR) (Boutell et al., 2004), Calibrated Label Ranking (Fürnkranz, Hüllermeier, Mencía, & Brinker, 2008), Classifier Chains (Read et al., 2009), Random k-labelsets (Tsoumakas & Vlahavas, 2007). On the other hand, methods based on algorithm adaption extend or adapt the learning techniques to deal with multi-label data directly. Representative algorithms include AdaBoost.MH and AdaBoost.MR (Schapire & Singer, 2000) which are two simple extensions of AdaBoost, ML-DT (Clare & King, 2001) adapting decision tree techniques, lazy learning techniques such as ML-kNN (Zhang & Zhou, 2007) and BR-kNN (Spyromitros, Tsoumakas, & Vlahavas, 2008) to name a few.

* Corresponding author.
E-mail addresses: vikas007bca@gmail.com, vikas@uohyd.ac.in (V. Kumar), akpujari@curaj.ac.in (A.K. Pujari), vineetcs@uohyd.ernet.in (V. Padmanabhan), uusandeepsahu@gmail.com (S.K. Sahu), venkateswar.rao.kagita@gmail.com (V.R. Kagita).

http://dx.doi.org/10.1016/j.eswa.2017.09.020
0957-4174/

To cope with the challenge of exponential-sized output space, modeling inter-label correlations has been the major thrust of research in the area of multi-label classification in recent years (Bi & Kwok, 2014; Huang, Zhou, & Zhou, 2012; Li et al., 2016) and for this, use of parametrization and embedding have been the prime focus (Cabral et al., 2011; Huang, Li, Huang, & Wu, 2015; Huang et al., 2012; Li et al., 2016; Yu, Jain, Kar, & Dhillon, 2014). There are two strategies of embedding for exploiting inter-label correlation. The first is to learn label-specific features for each class label. In Huang et al. (2015, 2016), a parametrized approach is suggested to transform data from original feature space to label-specific feature space with the assumption that each class label is associated with a sparse label specific feature. The second approach models inter-label correlation implicitly using low-rank parametrization (Cabral et al., 2011; Yu et al., 2014). The debate is going on as to whether it is the low-rank embedding or the label-specific sparse transformation that models the label correlation accurately when the size of label set is sufficiently large. It can be seen that both the approaches are essentially a process of parametrization to overcome the complexity of multi-label classification and most often it is proposed to adopt linear parametrization. Some researchers (Kimura, Kudo, & Sun, 2016; Li & Guo, 2015) suggest a natural extension of their proposal of linear parametrization to nonlinear cases but no detailed study is undertaken in this direction. Moreover, all these approaches do not yield results beyond a particular level of accuracy for problems with large data and large number of labels.

Our experimental and theoretical study of the recent approaches for multi-label classification reveals many important aspects of the problem. It is clear that a single linear embedding h may not take us very far in finding accurate multi-label classification. There are several reasons for this: the diversity of the training set, the correlation among labels, the feature-label relationship, and most importantly, the learning algorithm to determine the mapping h. Normally, h is determined by a process of nonlinear optimization. We conclude, from our experience with all the major algorithms proposed so far, that the use of the entire training set for training a single h is not appropriate and single embedding h for all instances is not suitable when inter-label correlation exists. Thus, a research question that naturally arises is whether there can be a parametrization which is piecewise-linear. In this paper, we investigate this aspect and propose a novel method that generates optimal embeddings for subsets of training examples. Our method is novel in the sense that it judiciously selects a subset of training examples for training and then it assigns a suitable subset of the training set to an embedding. Using multiple embeddings and their assigned training sets, a new instance is classified and we show that the proposed method outperforms all major algorithms on all major benchmark datasets.

The rest of the paper is organized as follows. Section 2 briefly reviews the earlier research on multi-label learning. The outline of the proposed method is described in Section 3. We introduce our proposed method, termed as MLC-HMF in Section 4. Experimental analysis of proposed method is reported in Section 5. Finally, Section 6 concludes and indicates several issues for future work.

2. Related work

Given a feature matrix $\mathcal{X}$ and a label matrix $\mathcal{Y}$, the goal of linear parametrization is to learn the parameter W and a common formulation is the following optimization problem with regularized loss.

$$\min_{W} \ell(\mathcal{Y}, \mathcal{X}W) + \lambda R(W) \tag{1}$$

where $W \in \mathcal{R}^{d\times\mathcal{L}}$, $\ell(\cdot)$ is a loss function that measures how well $\mathcal{X}W$ approximates $\mathcal{Y}$, $R(\cdot)$ is a regularization function that promotes various desired properties in W (low-rank, sparsity, group-sparsity, etc.) and the constant $\lambda \geq 0$ is the regularization parameter which controls the extent of regularization. In Yu et al. (2014) a generic empirical risk minimization (ERM) framework is used with low-rank constraint on linear parametrization $W = UV^T$, where $U \in \mathcal{R}^{d\times k}$ and $V \in \mathcal{R}^{\mathcal{L}\times k}$ are of rank $k \ll d$. The problem can be restated as follows.

$$\min_{U,V} \ell(\mathcal{Y}, \mathcal{X}UV^T) + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2) \tag{2}$$

where $\|\cdot\|_F$ is Frobenius norm. The formulation in Eq. (2) can capture the intrinsic information of both feature and label space. It can also be seen as a joint learning framework in which dimensionality reduction and multi-label classification are performed simultaneously (Ji & Ye, 2009; Yu & Zhang, 2016).

Another approach to learn sparse label specific feature (Huang et al., 2015; 2016) is to constrain the ℓ_1 norms of the linear parametrization. The problem can be formulated as follows.

$$\min_{W} \ell(\mathcal{Y}, \mathcal{X}W) + \frac{\alpha}{2}\varphi(W) + \lambda\|W\|_1 \tag{3}$$

where $\|\cdot\|_1$ is ℓ_1 norm, the second term in Eq. (3) models inter-label correlation with the constant $\alpha \geq 0$ to control the extent of correlation.

These two approaches are limited to linear parametrization and hence, fail to capture nonlinear intrinsic information in feature and label spaces and may lead to severe information loss. To address this issue, nonlinear embedding methods for multi-label classification are proposed recently in Kimura et al. (2016); Li and Guo (2015). In Li and Guo (2015), a feature-aware nonlinear label space reduction method for multi-label classification is proposed. In Kimura et al. (2016), both training instances and labels are simultaneously embedded onto the same space to retain the existing relationships in original feature and label space.

3. Outline of the proposed approach

In this section we introduce the underlying principle of the proposed method. We start with the formulation given in Eq. (2). For exploiting correlations in the labels, one way is to factor the matrix $W = UV$ where $U \in \mathbb{R}^{d\times k}$ can be interpreted as a embedding of the features $\mathcal{X}$ into a k dimensional latent space and $V \in \mathbb{R}^{k\times L}$ is a linear classifier on this space. Regularization is provided by constraining the dimensionality of the latent space (k). The minimization in U and V is unfortunately non-convex, and Fazel et. al. (Fazel, Hindi, & Boyd, 2001) discovered the nuclear norm (sum of singular values) heuristic for matrix rank minimization, which is the convex relaxation of the rank minimization problem. Since $\mathcal{X}UV^T$ yields continuous values and $\mathcal{Y}$ is discrete, a natural choice is to use the well-known principle of maximum margin matrix factorization (MMMF) (Kumar, Pujari, Sahu, Kagita, & Padmanabhan, 2017; Rennie & Srebro, 2005). For a subset of training examples, the process of determining the embedding using principle of MMMF is described below.

Computing U, V**:** Let S $\subseteq$ {1, 2, ..., n} be the indices of current set $\mathcal{X}^S$ of training examples and the corresponding label vectors is submatrix $\mathcal{Y}^S$ of $\mathcal{Y}$. We use smooth hinge loss function to determine U^S and V^S for given training set $(\mathcal{X}^S, \mathcal{Y}^S)$. For sake of simplicity, we drop the suffix S. The problem is can be formulated as following minimization problem.

$$\min_{U,V} J(U,V) = \sum_{l=1}^{\mathcal{L}} \sum_{i\in S} h(y_{il}(x_i U V_l^T)) + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2) \tag{4}$$