Neurocomputing 259 (2017) 112–118

Contents lists available at ScienceDirect

Neurocomputing

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

Selecting label-dependent features for multi-label classification

Lishan Qiao*, Limei Zhang, Zhonggui Sun, Xueyan Liu

School of Mathematics, Liaocheng University, Liaocheng 252000, China

ARTICLE INFO

Article history:
Received 19 January 2016
Revised 27 May 2016
Accepted 3 August 2016
Available online 7 February 2017

Keywords:
Multi-label classification
Feature selection
Label dependency
Nesterov's method

ABSTRACT

An instance is often represented from different aspects (views or modalities), which leads to high-dimensional features and even multiple labels. In this paper, we focus on the feature selection problem in multi-label classification, for which a trivial solution is handling the labels dividedly. Obviously, such a scheme may not work well by leaving the label relationship out of consideration. Recently, several research works conduct feature selection directly under a multi-label framework by implicitly or explicitly modeling label relationship. However, these works assume that all labels share the same feature subset or subspace, which is not reasonable enough for some scenarios since different labels tend to convey different semantics. To address this problem, we develop a novel approach in this paper to select *label-dependent* features for multi-label classification. Specifically, we (1) formulate a convex model based on a more general and practical assumption that different labels convey different semantics with specific features; (2) design an alternating optimization algorithm based on Nesterov's method and L_1-ball projection for efficiently finding the optimal solution, which can realize multi-label classification, feature selection, and label relationship estimation simultaneously. Finally, experiments on publicly available datasets show that the proposed algorithm achieves better performance than several related methods.

1. Introduction

In many real problems, data are often collected from multiple modalities, represented by multiple views, or analyzed for multiple tasks. Consequently, an instance is generally associated with high-dimensional features and even more than one label, especially when the data have semantic ambiguity. For example, a natural scene can be described by many kinds of features (e.g., color and texture) and with several labels such as *sky, cloud*, and *tree*. Therefore, how to select suitable features for multi-label classification is an increasingly important topic, due to its great value in practice.

A simple strategy for multi-label feature selection is handling the labels one by one. That is, one first decomposes the multi-label problem into several single-label problems, and then selects features for each sub-problem based on traditional feature selection methods. However, such a scheme is obviously not optimal, since it fails to model the relationship among different labels [1,2]. For example, a scene image would be labeled as *sky* with high possibility if it has been labeled with *cloud*.

Recently, researchers proposed to select features directly under a multi-label classification framework. To our best knowledge, there are two representative methods following this line. One is Ji et al.'s multi-label formulation based on the least squares (ML_{LS}) [3] which **implicitly** considers the label relationship by assuming that all labels share a common feature subspace, and the other is Gu et al.'s correlated multi-label feature selection (CMLFS) [4] which **explicitly** models the label relationship for feature selection based on label rank SVM.

A main issue involved in the above two methods is their assumption that all labels share the same features without distinguishing their differences. In practice, however, different labels usually have different semantics, and thus tend to be supported by specific features. This is important for both interpretability and discrimination of the selected features. Therefore, in this paper, we develop a new method to **s**elect **l**abel-d**e**pendent **f**eatures for **m**ulti-**l**abel classification ($SLEF_{ML}$) with label relationship learning simultaneously in a unified framework. In particular, the proposed method has the following characteristics:

(1) Novel assumption. Different from the existing multi-label feature selection methods that select the same features for all labels, the proposed method assumes different labels being supported by specific features. As a result, each label will have its personalized features, and meanwhile different labels might automatically share some common features, which provide a natural way to exploit label relationship (see Section 3.2 for more details).
(2) A unified model for classification, feature selection and label relationship learning. Based on the proposed framework, we implement three tasks (i.e., multi-label classification,

* Corresponding author
E-mail addresses: qlishan@163.com, qiaolishan@lcu.edu.cn (L. Qiao).

http://dx.doi.org/10.1016/j.neucom.2016.08.122
0925-2312/

label-dependent feature selection, and label relationship learning) simultaneously in a single model. Many typical feature selection methods in single-label field, such as LASSO and elastic net [5], can be considered as special cases of the proposed multi-label learning framework.

(3) Convex formulation and efficient learning algorithm. The proposed model is convex, and thus has a global optimal solution. We design an efficient alternating optimization algorithm by extending Nesterov's method with L_1-ball projection developed in [6] to a multi-label version for obtaining the optimal solution.

Finally, we conduct experiments including two illustrative examples and multi-label classification on several data sets used in related works [3,4]. Despite its simplicity, the proposed algorithm achieves competitive performances compared with the related methods.

The rest of the paper is organized as follows. In Section 2, we introduce several related works. In Section 3, we present our method including motivation, formulation, algorithm, and some interesting insights. In Section 4, we conduct experiments for validating the proposed method, and in Section 5 we conclude the paper.

2. Related works

In this paper, we focus on feature selection for multi-label classification. As pointed out previously, the simplest way is decomposing the multi-label problem into a set of single-label ones, and then employ existing feature selection strategies (including filter, wrapper and embedding) for each sub-problem [7,8]. For instance, the often-used decomposition is the so-called 1-vs-all scheme, or binary relevance [9]. Refer to recent survey [10] for more information about this topic.

However, the *divide-and-conquer* strategy above cannot effectively encode the label relationship into the feature selection model. Recently, Zhang [11] proposed a LIFT algorithm to construct new features for each label by clustering analysis, but it emphasizes feature reconstruction instead of feature selection. In addition, it does not consider label relationship yet. As a result, researchers attempt to conduct feature selection and exploit label dependency simultaneously in a multi-label classification framework. In the subsections below, we briefly review two representative methods following this line.

2.1. Multi-label formulation based on least squares

Multi-label formulation based on least squares loss (ML_{LS}) learns a common feature subspace for all labels in a least squares classification framework. With a series of mathematical formulation, ML_{LS} can be rewritten as the following optimization problem:

$$\min_{W,\Theta} \ \frac{1}{n}\|XW-Y\|_F^2+\alpha\|W\|_F^2+\beta\left\|W-\Theta^T\Theta W\right\|_F^2$$
$$s.t. \quad \Theta\Theta^T=I \tag{1}$$

where $X=[x_1,\ldots\cdots,x_n]^T\in R^{n\times d}$ is a data matrix with n and d denoting sample size and dimension, respectively; $Y\in R^{n\times l}$ is the label matrix whose entry $y_{ij}=1$ if the ith sample has the jth label, and -1 otherwise; $W=[w_1,\ldots\cdots,w_l]^T\in R^{d\times l}$ is the model parameter matrix with $w_j,\ j=1,\ldots,l$, corresponding to the linear classifier for the jth label; $\Theta\in R^{r\times d}(r<d)$ spans the common feature subspace shared by all labels. As a result, ML_{LS} in fact exploits the label relationship **implicitly** since different labels share the same features. The optimal solution of Eq. (1) can be obtained by solving a generalized eigenvalue problem, though it is not jointly convex w.r.t. W and Θ [3].

2.2. Correlated multi-label feature selection

Similar to ML_{LS}, correlated multi-label feature selection (CMLFS) handles feature selection and label relationship in a regularized multi-label classification framework as follows.

$$\min_{W,\Omega,\xi,p} \ \sum_{i=1}^{n}\frac{1}{|y_i||\bar{y}_i|}\sum_{(j,k)\in y_i\times\bar{y}_i}\xi_{ijk}+\alpha\|W\|_F^2+\beta tr\left(W\Omega^{-1}W^T\right)$$
$$s.t. \quad \left\langle w_j^T-w_k^T,\, p\circ x_i\right\rangle\ge 1-\xi_{ijk},\quad (j,k)\in y_i\times\bar{y}_i$$
$$\xi_{ijk}\ge 0,\quad i=1,\ldots,n,\quad p\in\{0,1\}^d,\quad p^T1=r$$
$$\Omega\succ 0,\quad tr(\Omega)=1 \tag{2}$$

where y_i is the label set for the ith sample, and $\bar{y}_i$ is the complementary set of y_i. ξ_{ijk} is the slack variable used in label rank SVM [4]. Ω is a positive definite matrix for modeling the label relationship. From Eq. (2) we note that the main differences between CMLFS and ML_{LS} include: (1) CMLFS employs more complex loss (i.e., label ranking loss corresponding to the first term of the objective function); (2) CMLFS encodes label correlation in the third term of the objective function **explicitly** by imposing a matrix-variate Normal (MVN) prior [12] on the weight matrix W. In fact, MVN prior has been used in multi-task learning for modeling the relationship among different tasks [13]. Shortly, we will show that it is a special case of the prior used in our model, and we formulate the prior based on a different motivation. Furthermore, we note that the feature indicator parameter p in Eq. (2) is common for all labels, and thus CMLFS also select the same features for all labels like ML_{LS}.

3. The proposed method

In this section, we dwell on the proposed method, including its motivation (assumption), model, and algorithm.

3.1. Motivation and assumption

As discussed above, ML_{LS} and CMLFS are two representative methods for treating feature selection and label correlation in a single framework. Despite their encouraging performances on some public data sets, both methods suffer from a common issue that they assume different labels share the same features. In fact, such an assumption partially stems from multi-task learning [14] where the feature sharing scheme aims at transferring knowledge among different data sources. In our view, the feature sharing strategy successes in multi-task scenario mainly owes to the fact such a scheme increases the training sample size by pooling the samples with common features together, and then may help improve the generalization.

However, multi-label problem has some differences from multi-task learning, and the assumption above does not necessarily work well. We summarize several possible reasons as follows.

(1) Multiple labels are based on a single data source, and thus the training samples do not increase, even though using the same feature-sharing scheme as in multi-task learning.

(2) Different labels tend to have different semantics that are more likely to be supported by different features. In Section 4.1, we provide an example based on face images for illustrating this point.

(3) Of course, there are some common features among different labels, but the sharing mechanism may be exceedingly complex. For example, label 1 shares some features with label 2, while label 2 shares another group of features with label 3. A recent work [15] assumes that "highly-related outputs may share a common set of relevant features, whereas weakly related outputs are less likely to be affected by the