Knowledge-Based Systems 000 (2016) 1–11

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Robust label compression for multi-label classification

Ju-Jie Zhang, Min Fang*, Jin-Qiao Wu, Xiao Li

School of Computer Science and Technology, Xidian University, No. 2, South Taibai Road, Xi'an, Shaanxi, 710071, P.R. China

ARTICLE INFO

Article history:
Received 19 October 2015
Revised 24 May 2016
Accepted 25 May 2016
Available online xxx

Keywords:
Multi-label classification
Label compression
Encoding loss
Dependence loss
Outliers
$l_{2,1}$-norm

ABSTRACT

Label compression (LC) is an effective strategy to reduce time cost and improve classification performance simultaneously for multi-label classification. One main limitation of existing LC methods is that they are prone to outliers. Here outliers include outliers in the feature space and outliers in the label space. Outliers in the feature space are obtained due to data acquisition devices. Outliers in the label space refer to label vectors that are inconsistent with the regular label correlations. In this paper, we propose a new LC method, termed *robust label compression* (RLC), based on $l_{2,1}$-norm to deal with outliers in the feature space and label space. The objective function of RLC consists of two losses: the encoding loss to measure the compression error and the dependence loss to measure the relevance between the instances and the obtained code vectors after compressing the label vectors. To achieve robustness to outliers, we utilize the $l_{2,1}$-norm on both losses. We propose an efficient optimization algorithm for it and present theoretical analysis. Experiments across six data sets validate the superiority of our proposed method to state-of-art LC methods for multi-label classification.

1. Introduction

In the past decade, multi-label classification has attracted the attention from a lot of research fields, such as image annotation [1,2], gene functional analysis [3], music understanding [4] and text categorization [5,6,7]. In multi-label context, an instance (object) can be associated with several labels simultaneously, while in the single-label (binary or multi-class) context, an instance is only allowed to be associated with one label. Due to the demand of real-world applications, a lot of multi-label classification methods have been proposed [8]. A consensus on multi-label classification is to make use of label correlations for performance improvement [8,9].

In many real-world applications, there are a lot of possible labels, which poses a great challenge to multi-label classification. On one hand, it usually costs a lot of time to conduct training and testing even on the basis of intuitive one-versus-all strategy (OVA) [1,10,11], and more time on the basis of exploiting pairwise [12,13] or higher-order label correlation [14,15]. Although it saves a lot of time with the help of pruning techniques [9,14,16], the time complexity is still prohibitive. On the other hand, it has been a consensus that the classification performance can be improved by utilizing label correlations [17,18]. How to use label correlations effectively is still an open question. Therefore, some researchers have resorted to *label compression* (LC) on multi-label classification problems to reduce the training and testing time while improving classification performance by exploiting label correlations to [19,20,21]. This paper focuses on label compression methods for multi-label classification.

There are only a few LC methods. The state-of-art LC one, e.g. *principal label space transformation* (PLST) [21], has been shown to be inferior to feature-aware ones, e.g. *conditional PLST* (CPLST) [20], which attempt to exploit feature information in the label compression process. However, they suffer from the limitation that they are prone to outliers. The outliers may exist in instances or label vectors. Outliers in instances may affect the compression quality via its relationship with the code vectors as done in CPLST. Some label vectors work as outliers since they are inconsistent with the main label correlations which play the key role in label compression. These label vectors may have negative impact on the label compression. Therefore, it is necessary to take these outliers into account when compressing the label space.

In this paper, we propose a new LC method, termed *robust label compression* (RLC), which suppresses the effect of outliers by using $l_{2,1}$-norm. We assume that the objective function consists of two losses: encoding loss and dependence loss. The encoding loss measures the compression quality of the label matrix, while the dependence loss measures the dependence between the code vectors of label matrix and the instances. To combat the negative influence of outliers, we propose to use $l_{2,1}$-norm for both encoding and dependence losses, in which we also need to find an optimal mean of the label matrix besides the projection matrix. We then propose an efficient optimization algorithm for it, which is proven

* Corresponding author.
E-mail address: mfang@mail.xidian.edu.cn, fanglabtg@163.com (M. Fang).

http://dx.doi.org/10.1016/j.knosys.2016.05.051
0950-7051/

Table 1
The general process of label compression methods for multi-label classification.

Training
1. Label compression: get the code vectors $\{\mathbf{z}_i\}_{i=1}^N$ via $\mathbf{z}_i = \phi(\mathbf{y}_i)$;
2. Learning: obtain a model $g(\mathbf{x})$ mapping $\{\mathbf{x}_i\}_{i=1}^N$ to $\{\mathbf{z}_i\}_{i=1}^N$;
Testing
3. Prediction: make prediction for an unseen instance $\mathbf{x}$: $g(\mathbf{x})$;
4. Reconstruction: get the final prediction $\varphi(g(\mathbf{x}))$

to converge to a local minimum. In experiments, we evaluate our proposed method on 6 benchmark data sets and the results verify its superiority to several state-of-art LC methods.

The main contributions of this paper are:

- To the best of our knowledge, this paper is the first one that takes outliers into account in label compression methods for multi-label classification.
- In this paper, those label vectors that are not in accordance with the main label correlations can be viewed as outliers in the label space.
- To combat the effect of outliers in the instances and label vectors, we propose a new LC method for multi-label classification based on $l_{2,1}$-norm, termed *robust label compression* (RLC).
- We propose an efficient optimization algorithm for RLC, whose convergence analysis is also discussed.
- The proposed method needs to find the optimal mean for the label matrix under $l_{2,1}$-norm, which is automatically found by the optimization algorithm.

The remainder of this paper is organized as follows. Section 2 briefly reviews existing LC methods. Section 3 presents the proposed method in details. Section 4 presents and analyzes the experimental results. Section 5 concludes this paper and points out some issues for future work.

2. Related works

Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^N\}$ be a multi-label data set with N examples. The ith example $(\mathbf{x}_i, \mathbf{y}_i)$ consists of an instance (or feature vector) $\mathbf{x}_i \in R^d$ of d dimensions and its corresponding label vector $\mathbf{y}_i \in \{0, 1\}^L$ of L possible labels. When $\mathbf{x}_i$ is relevant to the lth label, $\mathbf{y}_{il} = 1$ (or $\mathbf{y}_i(l) = 1$); otherwise $\mathbf{y}_{il} = -1$ (or $\mathbf{y}_i(l) = -1$). We also introduce some notations and definitions used in this paper. The l_p-norm of the vector $\mathbf{m} \in R^d$ is defined as $\|\mathbf{m}\|_p = (\sum_{j=1}^d |\mathbf{m}_j|^p)^{\frac{1}{p}}$. Thus the l_2-norm of $\mathbf{m}$ is $\|\mathbf{m}\|_2 = \sqrt{\mathbf{m}^T\mathbf{m}}$. Given a matrix $\mathbf{M} = \{\mathbf{M}_{ij}\}$, denote its ith row and jth column as $\mathbf{M}^i$ and $\mathbf{M}_j$ respectively. The $l_{2,1}$-norm of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_{2,1} = \sum_{i=1} \|\mathbf{M}^i\|_2$ and the l_1-norm of $\mathbf{M}$ is defined as $\|\mathbf{M}\|_1 = \sum_{i=1} \|\mathbf{M}^i\|_1$. Moreover, let $\mathbf{I}$ be an identity matrix of appropriate size; let $\mathbf{1}$ be a vector of appropriate length with each element being one; denote by $tr(\cdot)$ the trace operator of a square matrix.

We present the general process of LC method for multi-label classification in Table 1. The task of label compression is to learn a compression function ϕ: $R^L \to R^t$ and a reconstruction function φ: $R^t \to R^L$ to compress the label matrix efficiently and facilitate the subsequent learning, where $t \le L$ is the number of desired dimensions of the reduced label space. Since this paper focuses on label compression, we do not involve concrete learning algorithms. The label compression phase is independent of the subsequent learning algorithm. For convenience of presentation, denote $\mathbf{X} \in R^{N \times d}$ the instance matrix with the ith row being $\mathbf{x}_i^T$ and denote $\mathbf{Y} \in \{0, 1\}^{N \times L}$ be the label matrix with the ith row being $\mathbf{y}_i^T$.

Generally speaking, the objective function of existing label compression methods for multi-label classification can be formulated in the following form:

$$\begin{aligned} &\min \ell(\phi, \varphi) = \ell_e(\varphi(\phi(\mathbf{Y})), \mathbf{Y}) + \ell_d(\phi(\mathbf{Y}), \psi(\mathbf{X})) \\ &\text{s.t.} \quad \mathcal{C}(\phi, \varphi) \end{aligned} \tag{1}$$

where we assume the objective function $\ell(\cdot, \cdot)$ consists of two losses: $\ell_e(\cdot, \cdot)$ for the encoding loss and $\ell_d(\cdot, \cdot)$ for the dependence loss. $\psi(\mathbf{X})$ is a function dependent on the instances. $\mathcal{C}(\phi, \varphi)$ is the constraints on $\phi(\cdot)$ and $\varphi(\cdot)$. Note that for simple presentation, we denote the code matrix by $\phi(\mathbf{Y})$, the ith row of which is code vector $\phi(\mathbf{y}_i)$. Similar notations are taken for $\varphi(\phi(\mathbf{Y}))$ and $\psi(\mathbf{X})$. The encoding loss measures the encoding quality of label matrix $\mathbf{Y}$, while the dependence loss measures the dependence between the code vectors $\phi(\mathbf{Y})$ and $\psi(\mathbf{X})$. To the best of our knowledge, there are only a few papers devoted to label compression for multi-label classification.

Hsu, *et al.* proposed the CS method, which used compressive sensing for label compression in $\ell_e(\cdot, \cdot)$ and used the least square loss as $\ell_d(\cdot, \cdot)$ to obtain a classifier directly [19]. It has two limitations. Firstly, the projection matrix A ($\phi(\mathbf{Y}) = \mathbf{YA}$) is randomly generated and thus may fail to make use of label correlations. Secondly, although the label compression (Step 1) is quite efficient, its reconstruction process (Step 4) is time consuming as it needs to solve a complicated optimization problem. In fact, CS cannot be categorized into a LC method because it is designed to obtain a classification model and does not intend to solve for a compression function $\phi(\cdot)$ and a reconstruction function $\varphi(\cdot)$.

The LC methods proposed in [22] and [23] only focus on label combinations of high frequencies and use them to conduct label compression. They ignore those label combinations of low frequencies, which fails them to predict these label combinations for an unseen instance. Moreover, in order to achieve satisfactory predicting performance, they need complicated decoding procedures which are of high time complexity.

The LC method proposed in [24] used Hilbert-Schmidt Independence Criterion [25] to exploit the relationship between instances $\mathbf{X}$ and code vectors $\phi(\mathbf{Y})$ by maximizing their Hilbert-Schmidt norm. However, this method only has the dependence loss $\ell_d(\cdot, \cdot)$ but the encoding loss $\ell_e(\cdot, \cdot)$, which yields unsatisfactory performance.

PLST ignores the second part of (1) and its encoding loss function is $\|\mathbf{Z} - \mathbf{ZUU}^T\|_2^2$, where $\mathbf{Z}$ denotes the centered $\mathbf{Y}$, $\mathbf{U} \in R^{L \times t}$ is the projection matrix, $\phi(\mathbf{z}) = \mathbf{U}^T\mathbf{z}$ and $\varphi(\mathbf{z}) = \mathbf{Uz}$ [21]. The constraint $\mathcal{C}(\phi, \varphi)$ is $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. The encoding and reconstruction procedure of PLST is more efficient than CS. It is also proven to be more effective than CS [21]. However, since the code vectors $\phi(\mathbf{Y})$ are used for learning by a regression (or classification) algorithm, it is reasonable to take into account the correlation between $\phi(\mathbf{Y})$ and $\mathbf{X}$, as shown in [20].

CPLST deals with the limitation of PLST by assuming that there exists linear relationship between $\mathbf{Z} = \phi(\mathbf{Y})$ and $\mathbf{X}$, i.e. the dependence loss measures the linear mapping loss from $\mathbf{X}$ to $\mathbf{Z}$. Empirical results have demonstrated the superiority of CPLST to CS and PLST [20].

3. Robust label compression

Two state-of-art LC methods, including PLST and CPLST, are based on ℓ_2-norm, which are prone to outliers. In label compression, a label vector that is not in accordance with regular label correlations can be viewed as an outlier. They have negative impact on the encoding loss in PLST and CPLST. A distant instance may act as an outlier through the dependence loss in CPLST. It is necessary to take outliers into account when performing label compression. Therefore, we propose our method, termed *robust label compression*, on the basis of CPLST.

Please cite this article as: J.-J. Zhang et al., Robust label compression for multi-label classification, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.05.051