Information Sciences 281 (2014) 310–320

Contents lists available at ScienceDirect

Information Sciences

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

MLSLR: Multilabel Learning via Sparse Logistic Regression

Huawen Liu [a,b,*], Shichao Zhang [c], Xindong Wu [d,e,*]

[a] College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, China
[b] NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China
[c] Department of Computer Science, Guangxi Normal University, China
[d] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China
[e] Department of Computer Science, University of Vermont, USA

ARTICLE INFO

Article history:
Received 2 February 2014
Received in revised form 21 April 2014
Accepted 10 May 2014
Available online 16 May 2014

Keywords:
Sparse learning
Logistic regression
Multilabel data
Elastic net
Variable selection

ABSTRACT

Multilabel learning, an emerging topic in machine learning, has received increasing attention in recent years. However, how to effectively tackle high-dimensional multilabel data, which are ubiquitous in real-world applications, is still an open issue in multilabel learning. Although many efforts have been made in variable selection for traditional data, little work concerns variable selection for multilabel data yet. In this paper, we propose a novel framework for multilabel learning, which can achieve the purposes of variable selection and classification learning simultaneously. Specifically, our method exploits logistic regression to train models on multilabel data for classification. Besides, an elastic net penalty is performed on the logistic regression model to handle ill-conditioned and over-fitting problems of high-dimensional data. To further improve the efficiency, we solve the convex optimization problem of logistic regression with the elastic net penalty by a quadratic approximation technique. The experimental results on seven multilabel data sets show that our method has achieved encouraging performance and is competitive with six popular multilabel learning algorithms in most cases.

1. Introduction

Supervised learning is a major task in data mining and machine learning and has been extensively studied during past four decades. Traditional supervised learning methods place emphases on the data (or samples), which are associated with class labels exclusively. However, in many real-world applications objects can often be tagged with two or more class labels simultaneously. For example, the movie '*avatar*' may be tagged with *action*, *science fiction* and *love* types; The journal of *Information Sciences* is associated with *journal*, *Elsevier*, *computer science* and so on. This kind of data is called multilabel data [34]. Multilabel data are ubiquitous in many domains, such as text categorization, image annotation and bioinformatics [30,34], and can be regarded as a special representation of multi-view data [36].

Multilabel learning has attracted significant attention from many interdisciplinary fields since it was introduced [30], because multilabel learning has a large number of potential applications in reality. By now, many multilabel learning algorithms have been developed and successfully applied in text categorization [21], images and video annotation [24], content annotation [23], music processing [18], bioinformatics [31], and so on. Generally, the multilabel learning algorithms can be categorized into two major groups, i.e., problem transformation and algorithm adaption [30]. Comparing to traditional

* Corresponding authors at: College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, China. Tel.: +86 431 85159375.
E-mail addresses: hwliu@zjnu.edu.cn (H. Liu), zhangsc@mailbox.gxnu.edu.cn (S. Zhang), xwu@uvm.edu (X. Wu).

http://dx.doi.org/10.1016/j.ins.2014.05.013
0020-0255/

learning methods, the characteristic of multilabel learning is that the outputs of multilabel learning methods are not mutually exclusive and they may even be correlated with each other in some cases.

A challenging issue for multilabel learning is the problem of '*curse of dimensionality*' raised from the high-dimensional multilabel data. With advancements of information techniques, data in many fields are turning to large in both size and dimensionality. As the dimensionality of data is getting larger, the so-called problem of multi-collinearity may occur with a high probability, resulting in multilabel learning more challenging and complicated. An effective solution of dealing with this troublesome issue is to perform dimension reduction on the multilabel data before constructing learning models. For example, Zhang and Zhou [35] projected the original data into a lower-dimensional space with the Hilbert–Schmidt Independence Criterion, while Ji et al. [12] extracted a common subspace shared among multiple labels by virtue of ridge regression. It is noticeable that this kind of work merely focuses on extracting common subspaces with fewer dimensions, but pays little attention on the interpretability of results. In fact, the derived subspaces are weighted combinations of the original ones by linear transformation, making the interpretation of results impossible [5].

In this paper, we investigate the problems of variable selection and classification learning for multilabel data, and then propose a joint multilabel learning framework. It can fulfill the purposes of multilabel classification and variable selection simultaneously. To the best of our knowledge, variable selection in the context of multilabel learning has not been fully exploited yet and is still an open issue, albeit it has been well investigated in the traditional learning methods [15]. For multilabel learning, there is a variety of reasons to perform variable selection, such as lessening computational cost, avoiding over-fitting, and improving prediction performance and interpretability [8,9]. To achieve variable selection, we impose an elastic net penalty on logistic regression for multilabel learning, where the ℓ_1-norm penalty of the elastic net aims at removing irrelevant variables, while the ℓ_2-norm penalty ensures that highly correlated variables have similar regression coefficients [37].

The main contributions of this paper are as follows:

- We propose a general framework of learning and variable selection for multilabel data, which can conduct classification learning and variable selection simultaneously. The purpose of classification learning is achieved via logistic regression, which is particularly capable of solving the problem of binary classification.
- We perform the elastic net penalty to the logistic regression model, where the ℓ_1-norm penalty provides a solution for variable selection, yielding a sparse model, while the ℓ_2-norm penalty offers a grouping effect of the correlated variables, and a unique solution when the number of variables is larger than samples.
- To further improve learning efficiency, we explore the convex optimization problem of the sparse logistic regression by virtue of a quadratic approximation technique.

Paper organization. The rest of this paper is organized as follows. Section 2 briefly reviews the state-of-the-art of dimension reduction techniques for multilabel data. The problem of multilabel learning is formulated in Section 3. Section 4 gives the model of our sparse logistic regression, and provides an analytical solution to this optimization problem. We report the experimental results on seven data sets in Section 5, and then conclude the paper in Section 6.

Notations. Throughout this paper, uppercase bold Roman letters denote matrices and lowercase ones denote column vectors. For a specific example, $\mathbf{x}$ is a column vector while $\mathbf{X}$ represents the matrix $[\mathbf{x}_1, \ldots, \mathbf{x}_d]$. $[\cdot]^T$ indicates the transpose of a matrix or vector. $\|\cdot\|_k$ denotes the Frobenius norm for matrices and k norm for vectors.

2. Related work

Generally speaking, multilabel data are often represented by variable and label spaces. Depending on the spaces performed by dimension reduction, the reduction methods for multilabel learning can be roughly divided into three major categories: *variable space*, *label space* and *hybrid space* reduction.

Variable space reduction, or dimension reduction, is conducted independently on the variable space of data without considering the label information. It is also known as unsupervised dimension reduction in the traditional learning methods, and has been extensively studied in machine learning. The merit of this kind of reduction methods is that the off-the-shelf reduction techniques are available for multilabel data without the necessity of any revisions. A typical example is principal component analysis (PCA), which projects high-dimensional data into a low-dimensional space while keeping the variance of the data as much as possible. However, since it does not exploit the label information completely, the final derived models often have relatively poor performance.

On the other hand, label space reduction places an emphasis merely on the class labels to lessen the correlations among them. For example, Hsu et al. [10] encoded and decoded the label space using the compressive sensing (CS) technique, which is widely used in image processing. Tai and Lin [28] transformed the label space into a small linear space by mapping all possible label sets to vertices of a hypercube, and then solved it with singular value decomposition. Bi and Kwok [1] represented the labels as a tree or directed acyclic graph, while Sun et al. [26] organized the labels and their relationships in the form of a hyper-graph, and then explored it as an eigenvalue problem. These reduction methods