Neural Networks 103 (2018) 9–18

Contents lists available at ScienceDirect

Neural Networks

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

Learning from label proportions on high-dimensional data

Yong Shi [a,b,c,f], Jiabin Liu [d], Zhiquan Qi [c,a,b,*], Bo Wang [e]

[a] Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China
[b] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China
[c] School of Economics and Management, University of Chinese Academy of Sciences, Beijing, 100190, China
[d] School of Computer and Control Engineering, University of Chinese Academy Sciences, Beijing 100190, China
[e] School of Information Technology and Management, University of International Business and Economics, Beijing 100029, China
[f] College of Information Science and Technology, University of Nebraska at Omaha, NE 68182, USA

ARTICLE INFO

Article history:
Received 4 August 2017
Received in revised form 5 February 2018
Accepted 6 March 2018
Available online 20 March 2018

Keywords:
Optimization
High-dimensional data
Learning from label proportions (LLP)
Random forests

ABSTRACT

Learning from label proportions (LLP), in which the training data is in the form of bags and only the proportion of each class in each bag is available, has attracted wide interest in machine learning. However, how to solve high-dimensional LLP problem is still a challenging task. In this paper, we propose a novel algorithm called learning from label proportions based on random forests (LLP-RF), which has the advantage of dealing with high-dimensional LLP problem. First, by defining the hidden class labels inside target bags as random variables, we formulate a robust loss function based on random forests and take the corresponding proportion information into LLP-RF by penalizing the difference between the ground truth and estimated label proportion. Second, a simple but efficient alternating annealing method is employed to solve the corresponding optimization model. At last, various experiments demonstrate that our algorithm can obtain the best accuracies on high-dimensional data compared with several recently developed methods.

1. Introduction

In the era of big data, the data amount is rapidly increasing, which leads to many classification problems fail to be efficiently solved by the traditional machine learning algorithms (Adankon & Cheriet, 2002; Breiman, 2001; Liu & Yao, 1999; Schmidhuber, 2014). For example, the performance of supervised learning algorithms extremely depends on the amount of labeled training data. However, as the number of training data increases, it is becoming infeasible or quite labor-intensive to obtain labels of instances. In fact, compared to the ground-truth label of each instance, the proportion of each class can be obtained much more easily and cheaply by random sampling or other prior knowledge. As a result, in practice it is very meaningful to learn a classifier only from the information of class proportions and instances without labels.

In recent years, learning from label proportions (LLP) has been proposed as a solution to the above problem (Quadrianto, Smola, Caetano, & Le, 2008; Yu, Liu, Kumar, Jebara, & Chang, 2013). In detail, it is a learning task where the training data is provided in the form of bags and only the proportions of the labels in each bag are available. Fig. 1 provides an illustration of this classification problem. The "○" is the unlabeled data, which is partitioned into four bags and there is no intersection between the data of different bags. In each bag the sizes of red and green rectangles denote the amount of different classes and a proportion information is available according to different sizes. A classifier can be trained only by the proportion information and instances without labels. On the right, the red and green points, respectively, represent different classes and the blue line denotes the hyperplane generated by the classifier.

As increasingly practice applications can be abstracted to this problem, LLP has received amount of attention from the research community. Some real-world applications for LLP are given as follows.

In the case of political election (Rüping, 2010), the voters can be divided into two groups: always-favorable voters and swing voters where the latter will make their decision depending on what the candidates can offer them. If the candidates want to win the election, the proportions of always-favorable voters plus a set of swing voters should exceed 50%. However, due to the limit of finance, campaign time and ability to make election promises, the candidates would like to focus on their attention to the regions where they can achieve the largest gains. As a result, it is very important for them to identify which class each voter belongs to according to proportions information revealed by the previous elections.

* Corresponding author at: School of Economics and Management, University of Chinese Academy of Sciences, Beijing, 100190, China.
E-mail address: qizhiquan@foxmail.com (Z. Qi).

https://doi.org/10.1016/j.neunet.2018.03.004
0893-6080/

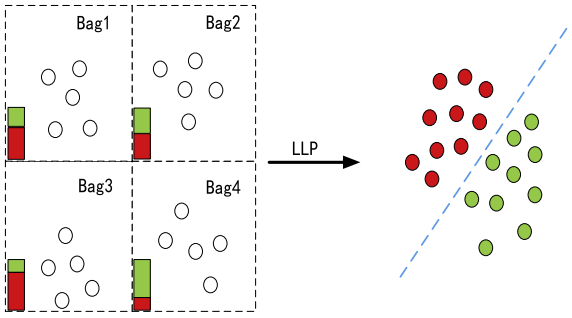

Fig. 1. The illustration of learning from class proportions. The "○" is the unlabeled data, which is partitioned into four bags, and there is no intersection between the data of different bags. In each bag the sizes of red and green rectangles denote the amount of different classes, and a proportion information is available according to different sizes. A classifier can be trained only by the proportion information and instances without labels. On the right the red and green points, respectively, represent different classes, and the blue line denotes the hyperplane generated by the classifier. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Similar problem can be found in spam filtering (Hernández-González, Inza, & Lozano, 2013). Spam is a kind of unsolicited and unwanted email from a stranger that is sent in bulk to large mailing lists. The user's inbox contains a mix of spam and non-spam and in most case it is difficult to obtain the actual labels of the spam and non-spam. However, the proportion of spam and non-spam in a user's inbox is usually cheaper to be achieved. Consequently, spam and non-spam in a user's box can be distinguished only using the inbox data and proportion information by LLP.

In addition, predicting income based census data (Yu, Choromanski, Kumar, Jebara, & Chang, 2014), image classification (Wang & Feng, 2013; Yu, Cao et al., 2014), video event detection (Lai, Yu, Chen, & Chang, 2014) and privacy preserving (Yu, Choromanski et al., 2014) can be also efficiently solved by LLP.

1.1. Related work

The learning problem was first addressed by Kuck and de Freitas (2012). They employed an Markov Chain Monte Carlo algorithm to handle the problem. However, the efficiency of this method is extremely limited by its complexity. Quadrianto et al. (2008) first gave a formal description of LLP problem and proposed a method called MeanMap which modeled the conditional class probability $p(y|x, \theta)$ using the conditional exponential models. Although this model performs well, the key assumption of MeanMap is that the class-conditional distribution of data is independent of the bags, which fail to hold for many real world applications. For example, the data distribution of voting behaviors can be highly dependent on the bags.

Rüping (2010) proposed a parametric LLP called Inverse Calibration that treated the mean of each bag as an instance. The classifier is learned from the group probabilities based on support vector regression and the idea of inverting a classifier calibration process. In detail, the objective function is modeled as a standard support vector regression problem:

$$\begin{aligned} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \quad &\rightarrow min \\ s.t. \forall_{i=1}^{m} : \xi_i, \xi_i^* \quad &\geq 0 \\ \forall_{i=1}^{m} : \frac{1}{|S_i|}\sum_{j\in S_i}(\mathbf{w}x_j + b) \quad &\geq y_i - \epsilon_i - \xi_i \\ \forall_{i=1}^{m} : \frac{1}{|S_i|}\sum_{j\in S_i}(\mathbf{w}x_j + b) \quad &\leq y_i + \epsilon_i + \xi_i^*, \end{aligned} \tag{1}$$

where the mean of each bag is treated as an instance. More specifically, $\mathbf{w}$, b are the model parameters of SVR and ξ_i, ξ_i^* are two slack variables. The constant C determines the penalty between the difference of real and predicted values, where the maximum tolerable error is defined by ϵ_i. Furthermore, S_i is the number of total instances in ith bag and $\frac{1}{|S_i|}\sum_{j\in S_i}(\mathbf{w}x_j + b)$ represents the ith bag mean. One limitation of Inverse Calibration is that $p_i = (1 + exp(-\mathbf{w}^T m_k + b))^{-1}$ is not a good way of measuring the proportion predicted and the performance can be extremely terrible in some uncommon cases argued by Yu et al. (2013).

Recently, a new method based on large-margin framework was proposed by Yu et al. (2013) by jointly optimizing the unknown instance labels and the known label proportions. The hyperplane was obtained by reaching a systemic compromise between large margin principle (LMP) and empirical proportion risk minimization principle (EPRMP). In detail, this LLP algorithm can be expressed as follows:

$$\begin{aligned} \min_{\mathbf{y},\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} L(y_i, \mathbf{w}^T\varphi(\mathbf{x}_i) + b) + C_p\sum_{k=1}^{K} L_p(p_k(y), p_k) \\ s.t. \quad & \forall_{i=1}^{N}, y_i \in \{-1, 1\}, \end{aligned} \tag{2}$$

where $L()$ is the loss function from SVM and $L_p()$ is a loss function to penalize the difference between the ground-truth and estimated proportion.

This algorithm outperforms the former known methods in most situations, and alleviates the need for making restrictive assumptions on the data. However, there also exist several limitations about this method. On one hand, this algorithm leads to a non-convex integer programming problem, which is always difficult to be solved. On the other hand, the performance of this algorithm is obtained on low-to-medium dimensional data. It is uncertain whether it has a similar performance when dealing with high-dimensional data.

Besides, Stolpe & Morik (2011) proposed a method based on clustering, which suffered from the extremely high computing complexity. Fan, Zhang, Yan, Wang, Zhang, and Feng (2014) formulated the learning from label proportion problem in a density estimation framework and proved sample complexity upper bound of this setting. Wang, Chen, and Qi (2015) proposed a new classification model based on twin SVM, which is in a large-margin framework and only needs to solve two smaller problems. Chen, Qi, Wang, Cui, Meng, and Shi (2017) and Qi, Wang, Meng, and Niu (2017) tried to address the LLP problem via NPSVM, where the former solved it by a generalized classifier instead of a transductive learning framework, and the latter incorporated the label proportion information with the unknown labels into one optimization model under a large-margin framework. Ding, Li, and Yu (2017) presented an efficient SVM-based classification framework for high-resolution SAR images. More specifically, it extended the method of Yu et al. (2013) by implementing a reweighting strategy for the training data. Shi, Cui, Chen, and Qi (2017) proposed a new SVM-based method by replacing hinge loss with pinball loss, and the model is effective to eliminate the impact of noise. Fish and Reyzin (2017) solved foundational questions which consider the computational complexity of LLP. In detail, they compared the computational complexity of learning in this model to classical PAC learning, and also gave an algorithm that demonstrated the feasibility of learning under well-behaved distributions.

1.2. Motivation

In the era of big data, as the dimension of data has increased substantially, high-dimensional data has become a trend for many machine learning problems. Meanwhile, how to solve