Neurocomputing 139 (2014) 47–55

Contents lists available at ScienceDirect

Neurocomputing

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

Learning a generative classifier from label proportions

Kai Fan [a], Hongyi Zhang [a], Songbai Yan [a], Liwei Wang [a,*], Wensheng Zhang [b], Jufu Feng [a]

[a] *Key Laboratory of Machine Perception, MOE, School of Electronics Engineering and Computer Science, Peking University, China*
[b] *Institute of Automation, Chinese Academy of Sciences, China*

ARTICLE INFO

Article history:
Received 14 April 2013
Received in revised form
3 September 2013
Accepted 17 September 2013
Available online 31 March 2014

Keywords:
Proportion learning
Bayesian model
Restricted Boltzmann machine

ABSTRACT

Learning a classifier when only knowing the features and *marginal distribution of class labels* in each of the data groups is both theoretically interesting and practically useful. Specifically, we consider the case in which the ratio of the number of data instances to the number of classes is large. We prove sample complexity upper bound in this setting, which is inspired by an analysis of existing algorithms. We further formulate the problem in a density estimation framework to learn a generative classifier. We also develop a practical RBM-based algorithm which shows promising performance on benchmark datasets.

1. Introduction

With the advent of *big data* era, it is easy to collect millions or even billions of unlabeled data instances, each of which belongs to a specific data group. Many organizations would like to release some datasets to the public. In many cases, such datasets may contain some sensitive information which involves the privacy of users and should not be revealed. Examples include GIC, which collected health insurance data for Massachusetts state employees; AOL, which collected search log data from its users; and Netflix, which collected movie ratings from its customers [17]. While such a publishing could benefit both the society and organizations themselves (for example, some technical researchers could utilize some scientific tools to obtain a better understanding of medical data or use advanced collaborating filtering algorithms for recommendation in social networks), there still exits a serious concern that some individual information will be revealed. If we interpret such a situation with machine learning community, it means that the sensitive attribute (such as whether someone has a certain disease, 1 for yes and 0 for no), could be the label of some non-sensitive features respecting to someone.

Suppose we have collected data in a number of groups, the problem of learning a classifier when only the *label proportions* (i.e. marginal distribution of class labels) in each data group are available naturally arises in two scenarios. In the first scenario, the label statistics in each group is available or can be reliably estimated; however, it is too expensive (or not allowed for privacy concerns) to collect the label information for every instance, as a result purely supervised learning is not possible, and learning from label proportions provides an useful alternative. In the second scenario, the class-conditional distribution of features evolves with time, space, etc., making it usually inappropriate to use a classifier learned beforehand to deal with the classification problem at present, in which case we have to resort to new learning paradigm. To summarize, the question we are interested in is: can we learn a useful classifier using the label proportions only?

In this paper we give an affirmative answer to this question, arguing that it is possible to build a useful classifier given data groups and the label proportions in each group, even when *no* individual datum is labeled. We propose a framework to build generative classifiers in this situation by density estimation, which is particularly well-suited for *big data* problems. Specifically, we prove that when data groups grow large with respect to the number of groups, our approach gains considerable advantages over a previously proposed SVM-based discriminative method. We also show how to build an RBM-based generative classifier derived from this framework, using information of label proportions from each data group only. The algorithm estimates the group-conditional likelihood of a specific datum by exploiting the annealed importance sampling (AIS) technique [12] to reliably estimate the normalization constant of each RBM. In experiments, our algorithm shows favorable performance on various real-world datasets.

The rest of the paper is organized as follows. In Section 2 we formulate the problem of learning from label proportion and

[*] Corresponding author.
E-mail address: wanglw@cis.pku.edu.cn (L. Wang).

http://dx.doi.org/10.1016/j.neucom.2013.09.057
0925-2312/© 2014 Elsevier B.V. All rights reserved.

review a few important algorithms. Section 3 proves a sample complexity upper bound. We propose our learning framework and new algorithm in Section 4. Experimental results comparing the performances of our new method with existing algorithms are given in Section 5. Section 6 gives the conclusion. Appendix A contains the proof of some of the theorems. Appendix B discusses the sample complexity upper bound in more detail with some empirical verification.

2. Preliminary and background

In this section we first give a formal description of the learning problem and then briefly review some representative existing works, which inspire our theoretical results of learnability and sample complexity bound in Section 3.

2.1. Problem formulation

For the learning with label proportion problem, the information the learner has is quite limited compared to traditional supervised learning. Fig. 1 illustrates four settings: supervised learning, unsupervised learning, semi-supervised learning and learning with label proportion.

We first formally define learning with label proportion.

Definition 1 (*Learning with label proportion*). Assume that $\mathcal{X}$ is an instance space (features product space) and Y is the set of labels. Let $P(X,Y)$ be a fixed but unknown probability distribution and Y be some discrete values, i.e. $Y=\{y_1,\ldots,y_l\}$. Given a set of unlabeled observations $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$, which is drawn i.i.d. from P and divided into n disjunct subsets $A_1,\ldots,A_n$, where we denote the size of $|A_i|=m_i, \forall i\in\{1,2,\ldots,n\}$. In addition, we do not know the label of each observation but know the label proportions π_{ij} for each subset, where π_{ij} denotes the proportion of label y_j in subset A_i. The goal of learning with label proportion is to design an algorithm which is able to predict y for each observation $x\in\mathcal{X}$ (or construct an estimator of $P(Y|X)$).

For convenience, we combine all label proportions π_{ij} as an $n\times l$ matrix Π, which will be referred to as *Proportion Matrix*. Clearly, the elements in each row sum up to one. This matrix will be used frequently below.

2.2. Related works

Although learning from incomplete label information (e.g., semi-supervised learning) has been extensively studied in the past two decades, there are relatively few works on learning from label proportions [3,1,5,6]. Below we briefly review two algorithms. These algorithms make clever uses of the Proportion Matrix. Analyses of these algorithms shed light on the learnability of the learning with label proportion problem.

2.2.1. Mean map

The concept of Proportion Matrix was initially proposed in [6], in which the authors develop an algorithm Mean Map to learn a classifier in the label proportion setting. The Mean Map algorithm is based on modeling the conditional class probability $p(y|x,\theta)$ by a conditional exponential model with only one parameter θ.

$$p(y|\mathbf{x},\theta)=\exp(\langle\phi(\mathbf{x},y),\theta\rangle-g(\theta|\mathbf{x})) \qquad (1)$$

where g is a normalizing function, θ is a vector in new space and ϕ is a feature map into a Reproducing Kernel Hilbert

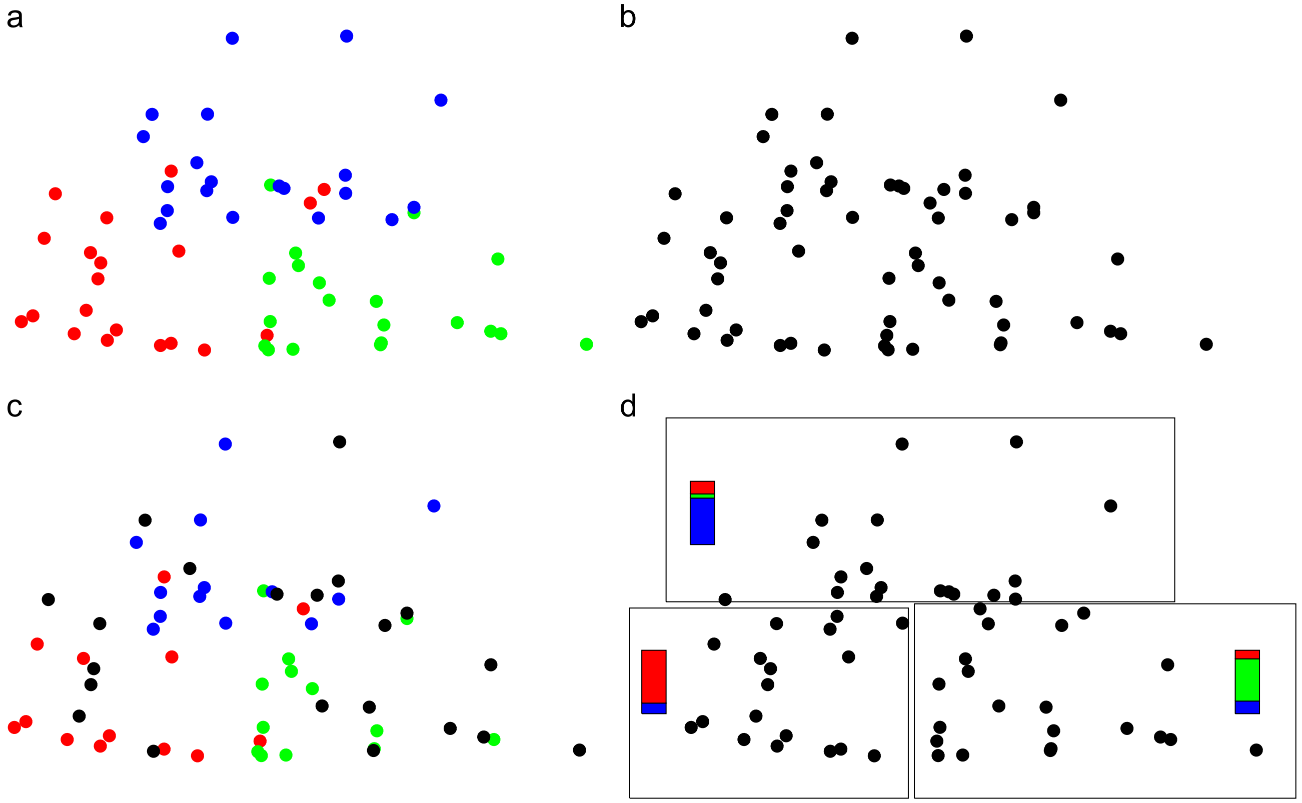

Fig. 1. Various cases of learning problem. From the colored pictures, we can apparently see that the information acquired in proportion case is between supervised case and unsupervised case. But it is difficult to determine which is more informative when the last case is compared semi-supervised one. (a) Supervised learning. (b) Unsupervised learning. (c) Semi-supervised learning. (d) Proportion learning. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)