Neural Networks 53 (2014) 146–164

Contents lists available at ScienceDirect

Neural Networks

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

Safe semi-supervised learning based on weighted likelihood

Masanori Kawakita *, Jun'ichi Takeuchi

Graduate School of Information Science and Electrical Engineering, Kyushu University, 744 Motooka, Nishi-Ku, Fukuoka, 819-0395, Japan

ARTICLE INFO

Article history:
Received 22 July 2013
Received in revised form 10 December 2013
Accepted 24 January 2014

Keywords:
Semi-supervised learning
Weighted likelihood
Estimating function
Statistical paradox
Geometrical interpretation
Density ratio

ABSTRACT

We are interested in developing a safe semi-supervised learning that works in any situation. Semi-supervised learning postulates that n' unlabeled data are available in addition to n labeled data. However, almost all of the previous semi-supervised methods require additional assumptions (not only unlabeled data) to make improvements on supervised learning. If such assumptions are not met, then the methods possibly perform worse than supervised learning. Sokolovska, Cappé, and Yvon (2008) proposed a semi-supervised method based on a weighted likelihood approach. They proved that this method asymptotically never performs worse than supervised learning (i.e., it is safe) without any assumption. Their method is attractive because it is easy to implement and is potentially general. Moreover, it is deeply related to a certain statistical paradox. However, the method of Sokolovska et al. (2008) assumes a very limited situation, i.e., classification, discrete covariates, $n' \to \infty$ and a maximum likelihood estimator. In this paper, we extend their method by modifying the weight. We prove that our proposal is safe in a significantly wide range of situations as long as $n \leq n'$. Further, we give a geometrical interpretation of the proof of safety through the relationship with the above-mentioned statistical paradox. Finally, we show that the above proposal is asymptotically safe even when $n' < n$ by modifying the weight. Numerical experiments illustrate the performance of these methods.

1. Introduction

Semi-supervised learning (SSL) has been shown to be effective in various fields. In the usual supervised learning (SL), we are given complete pairs of feature vector x and a label $y\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} | i = 1, 2, \ldots, n\}$ generated from $p(x)p(y|x)$. These data are called *labeled data*. In the SSL setting, additional *unlabeled data* $\{x'_j | j = 1, 2, \ldots, n'\}$ generated from $p'(x)$ are available. Though various settings have been considered for SSL, we define SSL as the case of $p'(x) \equiv p(x)$ throughout this paper. The goal of SSL is to improve the performance of SL by using unlabeled data. Almost all of the existing SSL methods have assumptions. For example, Nigam, McCallum, Thrun, and Mitchell (2000), Singh, Nowak, and Zhu (2008), and Sinha and Belkin (2008) have a cluster assumption, Bennett and Demiriz (1999) a low-density assumption, Belkin, Niyogi, and Sindhwani (2006), and Melacci and Belkin (2011) a manifold assumption, and co-training (Blum & Mitchell, 1998) a feature-split assumption. We shall refer to these assumptions as SSL assumptions. If the assumptions are not satisfied, SSL may perform worse than SL. This sounds somewhat strange because we have additional information from unlabeled data in the SSL setting. Thus, we conjecture that even without such assumptions, there is a form of SSL that never performs worse than SL. Furthermore, if the SSL assumptions are satisfied, SL can also exploit them. Several previous literatures compared SSL with SL that does not exploit SSL assumptions. In such cases, the comparison may be unfair because the effect of the unlabeled data can be confused with the effect of the SSL assumptions. We are interested in SSL that requires no assumptions to improve SL and our goal is to develop a SSL method that is always as good as or better than SL with as few assumptions as possible. We say that such an SSL is "safe" throughout the paper.

To the best of our knowledge, there have been two proposals for safe SSL so far. Sokolovska et al. (2008) proposed the first safe semi-supervised classification. Li and Zhou (2011) also proposed a safe semi-supervised support vector machine. We will focus on the results of Sokolovska et al. (2008) in this paper because their method is simpler and potentially general. Their method employs a weighted likelihood approach with an estimated ratio $p'(x)/p(x)$ as its weight. Sokolovska et al. (2008) proved that this weighted likelihood approach is asymptotically safe without any SSL assumptions. Their method is also theoretically interesting in that it is deeply related to a statistical paradox: even if you know the true value of nuisance parameter, you should estimate it using data in order to attain a more accurate interest parameter estimation. The relationship with this paradox will be discussed in Section 4. Note that Sokolovska et al. (2008) only concerns a considerably simple setting: the problem is restricted to only classification, the covariate space $\mathcal{X}$ is finite, $n' \to \infty$, the

* Corresponding author. Tel.: +81 92 802 3617.
E-mail address: kawakita@inf.kyushu-u.ac.jp (M. Kawakita).

http://dx.doi.org/10.1016/j.neunet.2014.01.016
0893-6080/

maximum likelihood estimator is used. It is not a trivial question to ask how widely their method and theory can be generalized by removing these restrictions.

In this paper, we propose an extension of Sokolovska et al. (2008)'s method that is free of these restrictions. A major difficulty in proving its safety arises in the continuity of $\mathcal{X}$. In this case, p'/p is a density ratio, so it may diverge to infinity. There are several ways to avoid this problem. Our choice is to modify the weights as follows:

$$\frac{p'(x)+\epsilon}{p(x)+\epsilon} \tag{1}$$

where ϵ is a small positive constant. With this modified weight, we prove that the weighted likelihood estimator is asymptotically safe. We call this estimator DRESS I (Density Ratio Estimation-based Semi-Supervised estimator). The proof of safety is free from all the above restrictions but requires only two major assumptions: the statistical model for $p(x)$ is correctly specified and $n \leq n'$. Many SSL methods assume the second assumption $n \leq n'$, because the SSL setting is supposed to have many unlabeled samples. In general, the assumptions of the previous SSL methods concern the structure between two distributions $p(x)$ and $p(y|x)$. In contrast, the first assumption says that if we prepare a sufficiently rich model $g(x;\eta)$ such that it contains $p(x)$, DRESS I is guaranteed to be safe.

Furthermore, we give a geometrical interpretation of the above asymptotical results. As described before, this weighted likelihood approach is deeply related to the statistical paradox mentioned above. Henmi and Eguchi (2004) analyzed its structure and gave a geometrical interpretation. Using similar techniques, we also give a graphical interpretation of the theoretical result. This graphical interpretation implies the existence of a better estimator than DRESS I. We show that such an estimator can be obtained by simply modifying the weights again. We call this estimator DRESS II. DRESS II is always safe as long as the statistical model for $p(x)$ is correctly specified. In other words, DRESS II is safe even if there are fewer unlabeled data than labeled data. In this sense, DRESS II is an almost perfect answer to our conjecture. We also illustrate the performance of DRESS I and II by numerical experiments.

DRESS I and II have two drawbacks. The first is if the model for $p(y|x)$ is correctly specified, both DRESS I and II perform asymptotically equally to SL (i.e., no improvement is obtained). In numerical experiments (i.e., n is finite), DRESS performs slightly worse than SL. However, the model misspecification of $p(y|x)$ often occurs in SSL. The SSL setting assumes only a few labeled data. Therefore, if you apply the conventional supervised model selection methods, they tend to select simpler models than the true model. Thus, the first drawback does not matter much. The second drawback is that DRESS needs a condition to guarantee the improvement on supervised learning. This condition requires that the statistical model of $p(x)$ is correctly specified. Exactly speaking, this condition is too strong because it is a sufficient condition to guarantee the improvement. A mitigated condition will be discussed in Section 4.1. However, it is still not trivial to find when this mitigated condition holds. This drawback can be easily overcome by applying density ratio estimation method (e.g., Kanamori, Hido & Sugiyama, 2009) in practice because we do not need to know the knowledge about the true density itself but only the density ratio $p'(x)/p(x)$. Actually, the author showed that DRESS I with parametric density ratio estimation is also asymptotically safe and performs better than SL on many real-world data sets (Kawakita & Kanamori, 2013). This is because specifying a model of the density ratio is much easier than specifying the model for density itself. In this sense, the proposals in this paper are not always good enough for real-world applications. However, this paper's work is conceptually basic and provides a mathematically firm basis of Kawakita and Kanamori (2013). Furthermore, it is intuitively more understandable and has a prominent mathematical structure. Our work is important in this sense.

The paper is organized as follows. In Section 2, we describe the setup of the semi-supervised learning problem. By extending the method of Sokolovska et al. (2008) to more general setting, we proposed two semi-supervised estimators in Section 3. In Section 4, we analyze their performance by asymptotic theory. Section 5 provides numerical experiments to illustrate our result. Section 6 is the conclusion.

2. Problem formulation

Let $\mathcal{X} \subset \Re^d$ be a covariate space and $\mathcal{Y}$ be a label space. The label space $\mathcal{Y}$ can be discrete or $\Re$. Let $(\mathcal{X}\times\mathcal{Y}, \mathcal{A}, P)$ be a joint probability space and let $(\mathcal{X}, \mathcal{A}_x, P')$ be a marginal probability space. For two positive integers n and n', we will consider a direct product space $((\mathcal{X}\times\mathcal{Y})^n \times \mathcal{X}^{n'}, \mathcal{A}^n \times \mathcal{A}_x^{n'}, P^n P_x^{n'})$. Assume that the probability measures $P(x,y)$ and $P'(x)$ have Radon–Nikodym derivatives $p(x,y)$ and $p'(x)$. These indicate that we have n labeled data and n' unlabeled data

$$\begin{aligned} D_L &:= \{(X_i, Y_i) \in \mathcal{X}\times\mathcal{Y} | i = 1, 2, \ldots, n\}, \\ D_U &:= \{X'_j \in \mathcal{X} | j = 1, 2, \ldots, n'\} \end{aligned} \tag{2}$$

where the pairs $\{(X_i, Y_i)\}$ are i.i.d. samples from $p(x)p(y|x)$ and $\{x'_j\}$ is generated independently from $p'(x)$. For convenience, we write $D_L \cup D_U$ as D. Here, $p'(x)$ can be regarded as a test data distribution. In other words, the test data (x,y) are generated from $p'(x)p(y|x)$. Our interest is to estimate the conditional probability $p(y|x)$. The problem is called a classification problem when $\mathcal{Y}$ is finite. For $\mathcal{Y} = \Re$, the problem is called a regression problem. If only the labeled data D_L is available, the problem is called a supervised learning. Assume that both D_L and D_U are available in the sequel. If $p'(x) \neq p(x)$, this setting is said to be a *covariate shift* (Shimodaira, 2000). In contrast, we say that this setting is for *semi-supervised learning* if $p'(x) \equiv p(x)$ (Chapelle, Schölkopf, & Zien, 2006). Note that we do not need to assume $n' \gg n$ unlike the usual semi-supervised setting. To estimate $p(y|x)$, we use the model

$$\mathcal{M}_{y|x} := \left\{ p(y|x;\alpha) | \alpha \in \mathcal{A} \subset \Re^d \right\}. \tag{3}$$

In the usual supervised setting, this model suffices to estimate $p(y|x)$. To define our semi-supervised estimator, we need to prepare a model of $p(x)$ defined as

$$\mathcal{M}_x := \{ g(x;\eta) | \eta \in \mathcal{N} \subset \Re^k \}. \tag{4}$$

We say that the model $\mathcal{M}_x$ is correctly specified if it contains the true density $p(x)$. In this paper, $\mathcal{M}_x$ is assumed to be correctly specified, whereas $\mathcal{M}_{y|x}$ is not necessarily correctly specified. Let α^* be a parameter such that $p(y|x;\alpha^*)$ is the closest distribution to $p(y|x)$ in terms of a certain criterion. The exact definition of α^* in Section 3.2. The goal is to estimate the parameter α^* based on available data, i.e., D_L in supervised learning or D in semi-supervised learning.

Finally, we summarize the differences between our setting and that of Sokolovska et al. (2008). Sokolovska et al. (2008) assume that

1. the feature space $\mathcal{X}$ is restricted to be finite.
2. $\mathcal{Y}$ is restricted to be finite (classification).
3. only the maximum likelihood estimator is considered.
4. $n' \to \infty$ ($p'(x)$ is known).

In contrast, we place no such restrictions.