Neural Networks 36 (2012) 73–82

Contents lists available at SciVerse ScienceDirect

Neural Networks

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

Improving importance estimation in pool-based batch active learning for approximate linear regression

Nozomi Kurihara, Masashi Sugiyama *

Department of Computer Science, Tokyo Institute of Technology, 2-12-1-W8-74 O-okayama, Meguro-ku, Tokyo 152-8552, Japan

ARTICLE INFO

Article history:
Received 10 February 2012
Received in revised form 4 June 2012
Accepted 3 September 2012

Keywords:
Pool-based batch active learning
Approximate linear regression
Covariate shift
Importance-weighted least-squares
P-ALICE
Inclusion probability

ABSTRACT

Pool-based batch active learning is aimed at choosing training inputs from a 'pool' of test inputs so that the generalization error is minimized. P-ALICE (Pool-based Active Learning using Importance-weighted least-squares learning based on Conditional Expectation of the generalization error) is a state-of-the-art method that can cope with model misspecification by weighting training samples according to the importance (i.e., the ratio of test and training input densities). However, importance estimation in the original P-ALICE is based on the assumption that the number of training samples to gather is small, which is not always true in practice. In this paper, we propose an alternative scheme for importance estimation based on the inclusion probability, and show its validity through numerical experiments.

1. Introduction

The objective of supervised learning is to find an input–output relationship behind training samples (Bishop, 2006; Hastie, Tibshirani, & Friedman, 2001). Once the input–output relationship is successfully learned, outputs for unseen inputs can be predicted, i.e., the learning machine can *generalize*.

When users are allowed to choose the location of training inputs, it is desirable to design the input locations so that the generalization error is minimized. Such a problem is called *active learning* (Settles, 2009) or *experiment design* (Fedorov, 1972; Pukelsheim, 1993), and has been shown to be useful in various application areas such as text classification (Lewis & Gale, 1994; McCallum & Nigam, 1998), age estimation from images (Ueki, Sugiyama, & Ihara, 2010), medical data analysis (Wiens & Guttag, 2010), chemical data analysis (Warmuth et al., 2003), biological data analysis (Liu, 2004), and robot control (Akiyama, Hachiya, & Sugiyama, 2010).

If users are allowed to locate training inputs at any position in the domain, the active learning setup is said to be *population-based* (Kanamori & Shimodaira, 2003; Sugiyama, 2006; Wiens, 2000). On the other hand, if users need to choose training input locations from a pool of finite candidate points, it is said to be *pool-based* (Kanamori, 2007; McCallum & Nigam, 1998; Sugiyama & Nakajima, 2009). Depending on the way training input locations are chosen, active learning is also categorized into *sequential* or *batch* approaches: Training inputs are selected one by one iteratively in the sequential approach (Box & Hunter, 1965; Sugiyama & Ogawa, 2000), while all training inputs are selected at once in the batch approach (Kiefer, 1959; Sugiyama & Ogawa, 2001). In this paper, we focus on pool-based batch active learning.

Active learning generally induces a *covariate shift* — a situation where training and test input distributions are different (Quiñonero-Candela et al., 2009; Shimodaira, 2000; Sugiyama & Kawanabe, 2012). When a model is correctly specified, covariate shifts do not matter in designing active learning methods. However, for a misspecified model, a covariate shift causes a strong estimation bias and thus classical active learning techniques that require a correct model become unreliable (Fedorov, 1972; Kiefer, 1959).

To cope with the bias induced by the covariate shift, active learning techniques that explicitly take model misspecification into account have been developed (Beygelzimer, Dasgupta, & Langford, 2009; Kanamori, 2007; Kanamori & Shimodaira, 2003; Sugiyama, 2006; Sugiyama & Nakajima, 2009; Wiens, 2000). The key idea of covariate shift adaptation is *importance weighting* — a loss function used for training is weighted according to the importance (i.e., the ratio of test and training input densities). Among the importance-weighted active learning methods, the pool-based batch active learning method for approximate linear regression called P-ALICE (Pool-based Active Learning using Importance-weighted least-squares learning based on Conditional Expectation of the generalization error) was demonstrated to be useful (Sugiyama & Nakajima, 2009).

However, in the original P-ALICE, the number of training samples to gather is assumed to be sufficiently smaller than the size of the sample pool. However, when this assumption is not satisfied, the importance weight used in P-ALICE is not reliable.

* Corresponding author.
E-mail addresses: sugi@cs.titech.ac.jp, sugi@sg.cs.titech.ac.jp (M. Sugiyama).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.09.002

In this paper, we propose a new method to set the importance weight that does not rely on this assumption. Our new weighting scheme is based on the *inclusion probability* (Horvitz & Thompson, 1952), which allows us to precisely capture the relation between the training and test input distributions. Through experiments, we show that the active learning performance of P-ALICE can be improved by the proposed weighting method when the training sample size is relatively large.

The rest of this paper is structured as follows. In Section 2, we formulate the problem of pool-based active learning and give an overview of P-ALICE. In Section 3, we point out a limitation of importance estimation in P-ALICE, and propose an alternative method. In Section 4, experimental results on toy and benchmark datasets are reported. Finally, concluding remarks are given in Section 5.

2. Problem formulation

In this section, we formulate the problem of pool-based active learning and briefly review the P-ALICE method.

2.1. Pool-based active learning for linear regression

Let us consider a regression problem of learning a real-valued function $f(\boldsymbol{x})$ defined on $\mathcal{D} \subset \mathbb{R}^d$. For training input–output samples

$$\{(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}}) \mid y_i^{\mathrm{tr}} = f(\boldsymbol{x}_i^{\mathrm{tr}}) + \epsilon_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}},$$

where $\{\epsilon_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ are i.i.d. noise with mean zero and unknown variance σ^2, let us use the following linear regression model:

$$\widehat{f}(\boldsymbol{x}) = \sum_{\ell=1}^{t} \theta_\ell \varphi_\ell(\boldsymbol{x}), \tag{1}$$

where $\{\varphi_\ell(\boldsymbol{x})\}_{\ell=1}^{t}$ are fixed linearly independent basis functions and $\{\theta_\ell\}_{\ell=1}^{t}$ are parameters to be learned. Let us denote the vector of parameters by $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_t)^\top$, where $^\top$ denotes the transpose.

The parameter $\boldsymbol{\theta}$ of the regression model is learned by *Weighted Least-Squares (WLS)* with weight function $w(\boldsymbol{x})$ (>0 for all $\boldsymbol{x} \in \mathcal{D}$), i.e.,

$$\widehat{\boldsymbol{\theta}}_{\mathrm{W}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n_{\mathrm{tr}}} w(\boldsymbol{x}_i^{\mathrm{tr}}) \left(\widehat{f}(\boldsymbol{x}_i^{\mathrm{tr}}) - y_i^{\mathrm{tr}}\right)^2 \right], \tag{2}$$

where the subscript 'W' denotes 'Weighted'. Let $\boldsymbol{X}$ be the $n_{\mathrm{tr}} \times t$ matrix with the (i, ℓ)-th element

$$X_{i,\ell} = \varphi_\ell(\boldsymbol{x}_i^{\mathrm{tr}}), \tag{3}$$

and let $\boldsymbol{W}$ be the $n_{\mathrm{tr}} \times n_{\mathrm{tr}}$ diagonal matrix with the i-th diagonal element

$$W_{i,i} = w(\boldsymbol{x}_i^{\mathrm{tr}}). \tag{4}$$

Then $\widehat{\boldsymbol{\theta}}_{\mathrm{W}}$ is given in a closed form as

$$\widehat{\boldsymbol{\theta}}_{\mathrm{W}} = \boldsymbol{L}_{\mathrm{W}} \boldsymbol{y}^{\mathrm{tr}}, \tag{5}$$

where[1]

$$\boldsymbol{L}_{\mathrm{W}} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W}, \tag{7}$$

$$\boldsymbol{y}^{\mathrm{tr}} = (y_1^{\mathrm{tr}}, y_2^{\mathrm{tr}}, \ldots, y_{n_{\mathrm{tr}}}^{\mathrm{tr}})^\top. \tag{8}$$

[1] Although we can obtain the analytic solution for Eq. (2), we often face with numerical instability when computing the inverse of the matrix $\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$ in Eq. (7). To avoid this problem, we practically employ a regularization technique (Hoerl & Kennard, 1970; Poggio & Girosi, 1990; Tikhonov & Arsenin, 1977), i.e., Eq. (7) is replaced with

$$\boldsymbol{L}_{\mathrm{W}} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} + \gamma \boldsymbol{I}_t)^{-1} \boldsymbol{X}^\top \boldsymbol{W}, \tag{6}$$

where γ is a small positive scalar called the regularization parameter and $\boldsymbol{I}_t$ is the $t \times t$ identity matrix. In our experiments, we set $\gamma = 10^{-10}$.

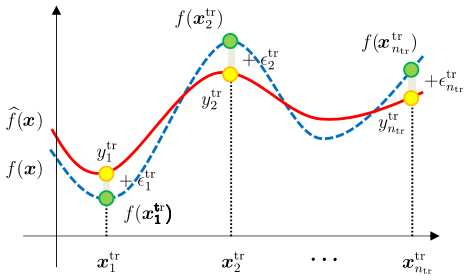

Fig. 1. Regression problem.

Note that the solution $\widehat{\boldsymbol{\theta}}_{\mathrm{W}}$ does not depend on the constant scaling of the weight function $w(\boldsymbol{x})$.

We adopt the squared-loss as the generalization error, i.e., the goodness of a learned function $\widehat{f}(\boldsymbol{x})$ is measured by

$$G = \int \left(\widehat{f}(\boldsymbol{x}^{\mathrm{te}}) - f(\boldsymbol{x}^{\mathrm{te}})\right)^2 p_{\mathrm{te}}(\boldsymbol{x}^{\mathrm{te}}) \mathrm{d}\boldsymbol{x}^{\mathrm{te}}, \tag{9}$$

where $p_{\mathrm{te}}(\boldsymbol{x})$ (>0 for all $\boldsymbol{x} \in \mathcal{D}$) is a probability density function of test input points. The above formulation is summarized in Fig. 1.

Below, we consider a pool-based active learning situation where we are given a 'pool' of test input points $\{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{n_{\mathrm{te}}}$ drawn independently from $p_{\mathrm{te}}(\boldsymbol{x})$ and choose training input points $\{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{n_{\mathrm{tr}}}$ from the pool so that the generalization error (9) is minimized.

2.2. P-ALICE

P-ALICE (Pool-based Active Learning using Importance-weighted least-squares learning based on Conditional Expectation of the generalization error; Sugiyama & Nakajima, 2009) is a pool-based active learning method that chooses training input points one by one (i.e., sampling *without* replacement), with probability proportional to a user-designed *resampling bias function* $b(\boldsymbol{x})$. Mathematically, the resampling bias function is a strictly positive function defined over the pool of test input samples. P-ALICE finds the resampling bias function that minimizes a generalization error estimator

$$J = \mathrm{tr}(\widehat{\boldsymbol{U}} \boldsymbol{L}_{\mathrm{W}} \boldsymbol{L}_{\mathrm{W}}^\top),$$

where $\widehat{\boldsymbol{U}}$ is the $t \times t$ matrix with the (ℓ, ℓ')-th element

$$\widehat{U}_{\ell,\ell'} = \frac{1}{n_{\mathrm{te}}} \sum_{j=1}^{n_{\mathrm{te}}} \varphi_\ell(\boldsymbol{x}_j^{\mathrm{te}}) \varphi_{\ell'}(\boldsymbol{x}_j^{\mathrm{te}}), \tag{10}$$

and

$$w(\boldsymbol{x}_j^{\mathrm{te}}) \propto \frac{1}{b(\boldsymbol{x}_j^{\mathrm{te}})} \tag{11}$$

is used as a weight in WLS (2).

A more detailed description of P-ALICE is given in Appendix.

3. Improving importance estimation in P-ALICE

In this section, we point out a weakness of P-ALICE and propose an alternative approach.

3.1. Weakness of P-ALICE

In P-ALICE, the importance weight $w(\boldsymbol{x}_j^{\mathrm{te}})$ is set as Eq. (11), which implies that the training input density $p_{\mathrm{tr}}(\boldsymbol{x}_j^{\mathrm{te}})$ is proportional to the product of the test input density $p_{\mathrm{te}}(\boldsymbol{x}_j^{\mathrm{te}})$ and a resampling bias function $b(\boldsymbol{x})$, i.e.,

$$p_{\mathrm{tr}}(\boldsymbol{x}_j^{\mathrm{te}}) \propto p_{\mathrm{te}}(\boldsymbol{x}_j^{\mathrm{te}}) b(\boldsymbol{x}_j^{\mathrm{te}}).$$