Neural Networks 50 (2014) 110–119

Contents lists available at ScienceDirect

Neural Networks

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

Semi-supervised learning of class balance under class-prior change by distribution matching☆

Marthinus Christoffel du Plessis*, Masashi Sugiyama

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

ARTICLE INFO

Article history:
Received 24 December 2012
Received in revised form 8 July 2013
Accepted 13 November 2013

Keywords:
Class-prior change
Density ratio
f-divergence
Selection bias

ABSTRACT

In real-world classification problems, the class balance in the training dataset does not necessarily reflect that of the test dataset, which can cause significant estimation bias. If the class ratio of the test dataset is known, instance re-weighting or resampling allows systematical bias correction. However, learning the class ratio of the test dataset is challenging when no labeled data is available from the test domain. In this paper, we propose to estimate the class ratio in the test dataset by matching probability distributions of training and test input data. We demonstrate the utility of the proposed approach through experiments.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Most supervised learning algorithms assume that training and test data follow the same probability distribution (Bishop, 2006; Hastie, Tibshirani, & Friedman, 2001; Vapnik, 1998). However, this de facto standard assumption is often violated in real-world problems, caused by intrinsic sample selection bias or inevitable non-stationarity (Heckman, 1979; Quiñonero-Candela, Sugiyama, Schwaighofer, & Lawrence, 2009; Sugiyama & Kawanabe, 2012).

In classification scenarios, changes in class balance are often observed—for example, the male–female ratio is almost fifty–fifty in the real-world (test set), whereas training samples collected in a research laboratory tend to be dominated by male data. Such a situation is called a *class-prior change*, and the bias caused by differing class balances can be systematically adjusted by instance re-weighting or resampling if the class balance in the test dataset is known (Elkan, 2001; Lin, Lee, & Wahba, 2002).

However, the class ratio in the test dataset is often unknown in practice. A possible approach to mitigating this problem is to learn a classifier so that the performance for all possible class balances are improved, e.g., through maximization of the area under the ROC curve (Clémençon, Vayatis, & Depecker, 2009; Cortes & Mohri, 2004). Alternatively, in the minimax approach, a classifier is learned so as to minimize the worst-case performance for any change in the class prior (Duda, Hart, & Stork, 2001; Van Trees, 1968). The disadvantage of the minimax approach is that it is often overly pessimistic. A more direct approach is to estimate the class ratio in the test dataset and use this estimate for instance re-weighting or resampling. We focus on this scenario under a semi-supervised learning setup (Chapelle, Schölkopf, & Zien, 2006), where no labeled data is available from the test domain.

Saerens, Latinne, and Decaestecker (2001) is a seminal paper on this topic, which proposed to estimate the class ratio by the expectation–maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977)—alternately updating the test class-prior and class-posterior probabilities from some initial estimates until convergence. This method has been successfully applied to various real-world problems such as word sense disambiguation (Chan & Ng, 2006) and remote sensing (Latinne, Saerens, & Decaestecker, 2001).

In this paper, we first reformulate the algorithm in Saerens et al. (2001), and show that this actually corresponds to approximating the test input distribution by a linear combination of class-wise training input distributions under the Kullback–Leibler (KL) divergence (Kullback & Leibler, 1951). In this procedure, the class-wise input distributions are approximated via class-posterior estimation, for example, by kernel logistic regression (Hastie et al., 2001) or its squared-loss variant (Sugiyama, 2010).

Since indirectly estimating the divergence by estimating the individual class-posterior distributions may not be the best scheme, the above reformulation motivates us to develop a more direct approach: matching the mixture of class-wise training input densities to the test input distribution. Historically, non-parametric estimation of the mixing proportions by matching

☆ This paper is an extended version of an earlier conference paper (du Plessis & Sugiyama, 2012).

* Corresponding author. Tel.: +81 8040687860.
E-mail addresses: christo@sg.cs.titech.ac.jp (M.C. du Plessis), sugi@cs.titech.ac.jp (M. Sugiyama).

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.11.010

the empirical distribution functions was investigated in Hall (1981), and its variant based on kernel density estimation has been developed in Titterington (1983). However, these classical approaches do not perform well in high-dimensional problems (Sugiyama et al., 2013). Recently, KL-divergence estimation based on *direct density-ratio estimation* has been shown to be promising (Nguyen, Wainwright, & Jordan, 2010; Sugiyama et al., 2008). Furthermore, a squared-loss variant of the KL divergence called the Pearson (PE) divergence (Pearson, 1900) can also be approximated in the same way, with an analytic solution that can be computed efficiently (Kanamori, Hido, & Sugiyama, 2009). Note that the PE-divergence and the KL divergence both belong to the f-divergence class (Ali & Silvey, 1966; Csiszár, 1967), which share similar properties. In this paper, with the aid of this density-ratio based PE-divergence estimator, we propose a new semi-supervised method for estimating the class ratio in the test dataset. Through experiments, we demonstrate the usefulness of the proposed method.

2. Problem formulation and existing method

In this section, we formulate the problem of semi-supervised class-prior estimation and review an existing method (Saerens et al., 2001).

2.1. Problem formulation

Let $\boldsymbol{x} \in \mathbb{R}^d$ be the d-dimensional input data, $y \in \{1, \ldots, c\}$ be the class label, and c be the number of classes. We consider class-prior change, i.e., the class-prior probability for training data $p(y)$ and that for test data $p'(y)$ are different. However, we assume that the class-conditional density for training data $p(\boldsymbol{x}|y)$ and that for test data $p'(\boldsymbol{x}|y)$ are the same:

$$p(\boldsymbol{x}|y) = p'(\boldsymbol{x}|y). \tag{1}$$

Note that training and test joint densities $p(\boldsymbol{x}, y)$ and $p'(\boldsymbol{x}, y)$ as well as training and test input densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ are generally different under this setup.

For the purposes of classification, we are generally interested in selecting a classifier that minimizes the expected loss (or the risk) with respect to the test distribution. We can rewrite the expected loss in terms of the training class-conditional density, $p(\boldsymbol{x}|y)$, as

$$\begin{aligned} R &= \sum_y \int L(f(\boldsymbol{x}), y) p'(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x} \\ &= \sum_y \int L(f(\boldsymbol{x}), y) p(\boldsymbol{x}|y) p'(y) \mathrm{d}\boldsymbol{x}, \end{aligned} \tag{2}$$

where $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is the loss function. Thus, if an estimate of the test class-priors is known, the expected loss can be calculated from the training class-conditional densities. The goal of this paper is to estimate $p'(y)$ from labeled training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ drawn independently from $p(\boldsymbol{x}, y)$ and unlabeled test samples $\{\boldsymbol{x}'_i\}_{i=1}^{n'}$ drawn independently from $p'(\boldsymbol{x})$.[1] Given test labels $\{y'_i\}_{i=1}^{n'}$, $p'(y)$ can be naively estimated by n'_y/n', where n'_y is the number of test samples in class y. Here, however, we would like to estimate $p'(y)$ *without* $\{y'_i\}_{i=1}^{n'}$.

[1] As we can confirm later, our proposed method does not actually require the independence assumption on $\{y_i\}_{i=1}^n$, but is valid for *deterministic* $\{y_i\}_{i=1}^n$ as long as $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ is drawn independently from $p(\boldsymbol{x}|y = y_i)$. However, for being consistent with other methods, we assume the independence condition here.

2.2. Existing method

We give a brief overview of an existing method for semi-supervised class-prior estimation (Saerens et al., 2001), which is based on the expectation–maximization (EM) algorithm (Dempster et al., 1977).

In the algorithm, test class-prior and class-posterior estimates $\widehat{p}'(y)$ and $\widehat{p}'(y|\boldsymbol{x})$ are iteratively updated as follows:

1. Obtain an estimate of the training class-posterior probability, $\widehat{p}(y|\boldsymbol{x})$, from training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, for example, by kernel logistic regression (Hastie et al., 2001) or its squared-loss variant (Sugiyama, 2010).
2. Obtain an estimate of the training class-prior probability, $\widehat{p}(y)$, from the labeled training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ as $\widehat{p}(y) = n_y/n$, where n_y is the number of training samples in class y. Set the initial estimate of the test class-prior probability equal to it: $\widehat{p}'_0(y) = \widehat{p}(y)$.
3. Repeat until convergence: $t = 1, 2, \ldots$
   (a) Compute a new test class-posterior estimate $\widehat{p}'_t(y|\boldsymbol{x})$ based on the current test class-prior estimate $\widehat{p}'_{t-1}(y)$ as
   $$\widehat{p}'_t(y|\boldsymbol{x}) = \frac{\widehat{p}'_{t-1}(y)\widehat{p}(y|\boldsymbol{x})/\widehat{p}(y)}{\sum_{y'=1}^{c} \widehat{p}'_{t-1}(y')\widehat{p}(y'|\boldsymbol{x})/\widehat{p}(y')}. \tag{3}$$
   (b) Compute a new test class-prior estimate $\widehat{p}'_t(y)$ based on the current test class-posterior estimate $\widehat{p}'_t(y|\boldsymbol{x})$ as
   $$\widehat{p}'_t(y) = \frac{1}{n'} \sum_{i=1}^{n'} \widehat{p}'_t(y|\boldsymbol{x}'_i). \tag{4}$$

Note that Eq. (3) comes from the Bayes formulae,

$$p(\boldsymbol{x}|y) = \frac{p(y|\boldsymbol{x})p(\boldsymbol{x})}{p(y)} \quad \text{and} \quad p'(\boldsymbol{x}|y) = \frac{p'(y|\boldsymbol{x})p'(\boldsymbol{x})}{p'(y)},$$

combined with Eq. (1):

$$p'(y|\boldsymbol{x}) \propto \frac{p'(y)}{p(y)} p(y|\boldsymbol{x}).$$

Eq. (4) comes from empirical marginalization of

$$p'(y) = \int p'(y|\boldsymbol{x})p'(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

It was suggested that this procedure may converge to a local optimal solution (Saerens et al., 2001). In the following section, we will show that the objective function is actually convex, but that the method suggested in Saerens et al. (2001) may fail to converge to the unique optimal value.

3. Reformulation of the EM algorithm as distribution matching

In this section, we show that the class priors can be estimated by matching the test input density to a linear combination of class-wise training input distributions under the Kullback–Leibler (KL) divergence (Kullback & Leibler, 1951). We show that the existing EM method performs this matching via an estimation of the class posterior. Furthermore, we show that this results in a convex problem, but that the existing EM method may not obtain the optimal result.

3.1. Class-prior estimation as distribution matching

Based on the assumption that the class-conditional densities for training and test data are unchanged (see Eq. (1)), let us model the

Download English Version:

https://daneshyari.com/en/article/404016

Download Persian Version:

https://daneshyari.com/article/404016

Daneshyari.com