Neural Networks 23 (2010) 44–59

Contents lists available at ScienceDirect

Neural Networks

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

Dimensionality reduction for density ratio estimation in high-dimensional spaces

Masashi Sugiyama [a,b,*], Motoaki Kawanabe [c,b], Pui Ling Chui [a]

[a] Department of Computer Science, Tokyo Institute of Technology, 2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan
[b] Mathematisches Forchungsinstitute Oberwolfach, Schwarzwaldstr. 9-11, 77709 Oberwolfach-Walke, Germany
[c] Fraunhofer Institute FIRST.IDA, Kekuléstr. 7, D-12489 Berlin, Germany

ARTICLE INFO

Article history:
Received 21 October 2008
Received in revised form 2 April 2009
Accepted 10 July 2009

Keywords:
Density ratio estimation
Dimensionality reduction
Local Fisher discriminant analysis
Unconstrained least-squares importance fitting

ABSTRACT

The ratio of two probability density functions is becoming a quantity of interest these days in the machine learning and data mining communities since it can be used for various data processing tasks such as *non-stationarity adaptation*, *outlier detection*, and *feature selection*. Recently, several methods have been developed for directly estimating the density ratio without going through density estimation and were shown to work well in various practical problems. However, these methods still perform rather poorly when the dimensionality of the data domain is high. In this paper, we propose to incorporate a dimensionality reduction scheme into a density-ratio estimation procedure and experimentally show that the estimation accuracy in high-dimensional cases can be improved.

1. Introduction

The ratio of two probability density functions (a.k.a. the *importance*; see Fishman, 1996) is attracting a great deal of attention these days in the machine learning and data mining communities since it can be used for various statistical data processing tasks such as *covariate shift adaptation* (Shimodaira, 2000; Sugiyama, Krauledat, & Müller, 2007; Zadrozny, 2004), *transfer learning* (Storkey & Sugiyama, 2007), *multi-task learning* (Bickel, Bogojeska, Lengauer, & Scheffer, 2008), *outlier detection* (Hido, Tsuboi, Kashima, Sugiyama, & Kanamori, 2008), *conditional density estimation* (Sugiyama, Takeuchi, Suzuki, Kanamori, & Hachiya, 2009), *variable selection* (Suzuki, Sugiyama, Kanamori, & Sese, 2009; Suzuki, Sugiyama, Sese, & Kanamori, 2008), *independent component analysis* (Suzuki & Sugiyama, 2009a), and *supervised dimensionality reduction* (Suzuki & Sugiyama, 2009b).

A naive approach to learning the density ratio is to estimate the two densities separately using a flexible technique such as *kernel density estimation* (Härdle, Müller, Sperlich, & Werwatz, 2004) and then take the ratio of the estimated densities. However, this two-step approach is not reliable in practice since kernel density estimation performs poorly in high-dimensional cases; furthermore, division by an estimated density tends to magnify the estimation error.

Thus it is important to avoid density estimation when learning the density ratio. Actually, estimating the densities is more general than estimating the density ratio since knowing the densities implies knowing the ratio but not vice versa. Such a statement is sometimes referred to as *Vapnik's principle* (Vapnik, 1998) and the *support vector machine* would be a successful example of this principle—instead of estimating the data generation model, it directly models the decision boundary which is simpler and sufficient for pattern recognition.

Following this spirit, various methods have been developed for directly estimating the density ratio without going through density estimation (Bickel, Brückner, & Scheffer, 2007; Cheng & Chu, 2004; Huang, Smola, Gretton, Borgwardt, & Schölkopf, 2007; Kanamori, Hido, & Sugiyama, 2009a; Qin, 1998; Sugiyama et al., 2008). These methods are shown to compare favorably with naive kernel density estimation through extensive experiments. However, these methods still perform rather poorly when the dimensionality of the data domain is high.

The purpose of this paper is to develop a new method that can mitigate this problem. Our basic assumption behind the proposed method is that the difference of the two distributions (i.e., the distributions corresponding to the denominator and numerator of the density ratio) does not spread over the entire data domain, but is confined in a *subspace*—which we refer to as the *hetero-distributional subspace*. Once the hetero-distributional subspace can be identified, the density ratio is estimated only within this subspace, which leads to more stable and reliable estimation of the density ratio. We experimentally show that the proposed method—which we refer to as *Direct Density-ratio estimation with*

* Corresponding author at: Department of Computer Science, Tokyo Institute of Technology, 2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan. Tel.: +81 3 5734 2699; fax: +81 3 5734 2699.
E-mail addresses: sugi@cs.titech.ac.jp (M. Sugiyama), motoaki.kawanabe@first.fraunhofer.de (M. Kawanabe), pauline@sg.cs.titech.ac.jp (P.L. Chui).
URL: http://sugiyama-www.cs.titech.ac.jp/~sugi/ (M. Sugiyama).

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.07.007

Dimensionality reduction (D^3; pronounced as 'D-cube')—improves the accuracy of density ratio estimation in high-dimensional cases, while the computational cost is still kept moderate.

The rest of this paper is organized as follows. In Section 2, we formulate the problem of density ratio estimation and illustrate how the density ratio could be utilized in various data processing tasks. In Section 3, the basic idea of the proposed method D^3 is explained; the details of the method are explained in Sections 4–6. Numerical examples are presented in Section 7 and concluding remarks are given in Section 8.

2. Formulation of density ratio estimation problem

In this section, we formulate the problem of density ratio estimation and briefly summarize possible usage of the density ratio in various data processing tasks.

2.1. Problem formulation

Let $\mathcal{D}(\subset \mathbb{R}^d)$ be the data domain and suppose we are given independent and identically distributed (i.i.d.) samples $\{\boldsymbol{x}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ from a distribution with density $p_{\mathrm{de}}(\boldsymbol{x})$ and i.i.d. samples $\{\boldsymbol{x}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$ from another distribution with density $p_{\mathrm{nu}}(\boldsymbol{x})$. We assume that the first density $p_{\mathrm{de}}(\boldsymbol{x})$ is strictly positive, i.e.,

$$p_{\mathrm{de}}(\boldsymbol{x}) > 0 \quad \text{for all } \boldsymbol{x} \in \mathcal{D}.$$

The problem we address in this article is to estimate the density ratio (also called the *importance* depending on the context)

$$r(\boldsymbol{x}) := \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})} \tag{1}$$

from samples $\{\boldsymbol{x}_i^{\mathrm{de}}\}_{i=1}^{n_{\mathrm{de}}}$ and $\{\boldsymbol{x}_j^{\mathrm{nu}}\}_{j=1}^{n_{\mathrm{nu}}}$. The subscripts 'nu' and 'de' denote 'numerator' and 'denominator', respectively.

2.2. Usage of density ratio in data processing

We are interested in estimating the density ratio since it is useful in various data processing tasks. Here we briefly review possible usage of the density ratio.

2.2.1. Covariate shift adaptation

Covariate shift (Shimodaira, 2000) is a situation in supervised learning where the input distributions change between the training and test phases but the conditional distribution of outputs given inputs remains unchanged. Under covariate shift, standard learning techniques such as maximum likelihood estimation are biased; the bias caused by covariate shift can be asymptotically canceled by weighting the loss function according to the importance (Shimodaira, 2000; Sugiyama et al., 2007; Sugiyama & Müller, 2005; Zadrozny, 2004). The basic idea of covariate shift adaptation is summarized in the following importance sampling identity:

$$\begin{aligned}\mathbb{E}_{p_{\mathrm{nu}}(\boldsymbol{x})}[g(\boldsymbol{x})] &= \int g(\boldsymbol{x}) p_{\mathrm{nu}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= \int g(\boldsymbol{x}) r(\boldsymbol{x}) p_{\mathrm{de}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \mathbb{E}_{p_{\mathrm{de}}(\boldsymbol{x})}[g(\boldsymbol{x}) r(\boldsymbol{x})],\end{aligned}$$

where $r(\boldsymbol{x})$ is defined by Eq. (1). That is, the expectation of a function $g(\boldsymbol{x})$ over $p_{\mathrm{nu}}(\boldsymbol{x})$ can be computed by the importance-weighted expectation over $p_{\mathrm{de}}(\boldsymbol{x})$. Similarly, standard model selection criteria such as cross-validation or Akaike's information criterion lose their unbiasedness due to covariate shift; proper unbiasedness can be recovered by modifying the methods based on importance weighting (Huang et al., 2007; Qin, 2009; Shimodaira, 2000; Sugiyama et al., 2007; Sugiyama & Müller, 2005; Zadrozny, 2004). Furthermore, the performance of active learning or the experiment design – the training input distribution is designed by the user to enhance the generalization performance – could also be improved by the use of the importance (Kanamori & Shimodaira, 2003; Sugiyama, 2006; Sugiyama & Nakajima, 2009; Wiens, 2000).

Thus the importance plays a central role in covariate shift adaptation and density-ratio estimation methods could be utilized for reducing the estimation bias under covariate shift. Examples of successful real-world applications include brain–computer interface (Sugiyama et al., 2007), robot control (Hachiya, Akiyama, Sugiyama, & Peters, in press), speaker identification (Yamada, Sugiyama, & Matsui, 2009), and natural language processing (Tsuboi, Kashima, Hido, Bickel, & Sugiyama, 2009). A similar importance-weighting idea also plays a central role in domain adaptation (Storkey & Sugiyama, 2007) and multi-task learning (Bickel et al., 2008).

2.2.2. Inlier-based outlier detection

Let us consider an outlier detection problem (Breunig, Kriegel, Ng, & Sander, 2000; Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001) of finding irregular samples in a dataset ('evaluation dataset') based on another dataset ('model dataset') that only contains regular samples. Defining the density ratio over the two sets of samples, we can see that the density-ratio values for regular samples are close to one, while those for outliers tend to be significantly deviated from one. Thus the density-ratio value could be used as an index of the degree of outlyingness (Hido et al., 2008). Since the evaluation dataset has a wider support than the model dataset, we regard the evaluation dataset as samples corresponding to $p_{\mathrm{de}}(\boldsymbol{x})$ and the model dataset as samples corresponding to $p_{\mathrm{nu}}(\boldsymbol{x})$. Then outliers tend to have smaller density-ratio values (i.e., close to zero). As such, density-ratio estimation methods could be employed in outlier detection scenarios.

A similar idea could be used for change-point detection in time-series (Brodsky & Darkhovsky, 1993; Kawahara & Sugiyama, 2009) and two-sample problems in hypothesis testing Henkel (1979).

2.2.3. Conditional density estimation

Suppose we are given n i.i.d. paired samples $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^n$ drawn from a joint distribution with density $q(\boldsymbol{x}, \boldsymbol{y})$. The goal is to estimate the conditional density $q(\boldsymbol{y}|\boldsymbol{x})$. When the domain of $\boldsymbol{x}$ is continuous, conditional density estimation is not straightforward since a naive empirical approximation cannot be used (Bishop, 2006; Takeuchi, Nomura, & Kanamori, 2009).

In the context of density ratio estimation, let us regard $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^n$ as samples corresponding to the numerator of the density ratio and $\{\boldsymbol{x}_k\}_{k=1}^n$ as samples corresponding to the denominator of the density ratio, i.e., we consider the density ratio defined by

$$r(\boldsymbol{x}, \boldsymbol{y}) := \frac{q(\boldsymbol{x}, \boldsymbol{y})}{q(\boldsymbol{x})} = q(\boldsymbol{y}|\boldsymbol{x}),$$

where $q(\boldsymbol{x})$ is the marginal density of $\boldsymbol{x}$. Thus a density-ratio estimation method directly gives an estimate of the conditional density.

2.2.4. Mutual information estimation

Suppose we are given n i.i.d. paired samples $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^n$ drawn from a joint distribution with density $q(\boldsymbol{x}, \boldsymbol{y})$. Let us denote the marginal densities of $\boldsymbol{x}$ and $\boldsymbol{y}$ by $q(\boldsymbol{x})$ and $q(\boldsymbol{y})$, respectively. Then mutual information $I(X, Y)$ between random variables X and Y is defined by

$$I(X, Y) := \iint q(\boldsymbol{x}, \boldsymbol{y}) \log \frac{q(\boldsymbol{x}, \boldsymbol{y})}{q(\boldsymbol{x}) q(\boldsymbol{y})} \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y},$$

which plays a central role in information theory (Cover & Thomas, 1991).