Neurocomputing 223 (2017) 45–53

Contents lists available at ScienceDirect

Neurocomputing

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

Error analysis for the semi-supervised algorithm under maximum correntropy criterion

Ling Zuo[a,b,*], Yulong Wang[c]

[a] School of Science, Hubei University of Technology,Wuhan 430068, China
[b] Hubei Key Laboratory of Applied Mathematics, Faculty of Mathematics and Statistics, Hubei University,Wuhan 430062, China
[c] Faculty of Science and Technology, University of Macau, China

ARTICLE INFO

Communicated by Zidong Wang

Keywords:
Semi-supervised learning
Correntropy
Excess generalization error
Manifold error

ABSTRACT

As a similarity measure, correntropy has been increasingly employed in machine learning research. While numerous experimental results have shown the effectiveness of correntropy based methods, the theoretical analysis in this area is still poorly understood. In this paper, we propose a novel semi-supervised algorithm under the maximum correntropy criterion, and present an elaborate error analysis for it. An excess generalization error bound is established, which demonstrates that the proposed method is consistent, and converges at a faster rate compared with the related studies. Moreover, experiments are implemented to show the efficiency of the proposed method.

1. Introduction

Semi-supervised learning (SSL), as a powerful tool to learn from a limited number of labeled data and a large number of unlabeled data, has attracted more and more attention in the machine learning research [1–4]. In this paper, we consider the semi-supervised regression (SSR) problem [5–8]. Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact metric space and $\mathcal{Y} = [-M, M]$ with a positive constant M. Assume that ρ is an underlying probability measure on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. And the data set consists of l labeled samples $\mathbf{z} := \{(x_i, y_i)\}_{i=1}^{l}$ drawn independently from ρ, together with a typically much larger collection of u unlabeled data $\mathbf{x} := \{x_j\}_{j=l+1}^{l+u}$ generated according to the marginal distribution $\rho_{\mathcal{X}}$ of ρ.

Suppose the data generation model is given by

$$Y = f_\rho(X) + E. \tag{1}$$

In Eq. (1), $X \in \mathcal{X}$ is the explanatory variable, E is a noise process, and $Y \in \mathcal{Y}$ stands for the response variable which is the label of X. The regression function $f_\rho(X)$ is defined as $\int_{\mathcal{X}} Y d\rho(Y|X)$, where $\rho(\cdot|X)$ is the conditional distribution of ρ. The aim of the SSR is to predict $f_\rho(X)$ from the labeled and unlabeled data generalized by (1).

Usually the prediction accuracy is measured by the mean square error (MSE). For two variables X and Y, MSE is defined as $\mathbf{E}_{XY}[(X-Y)^2]$, where $\mathbf{E}_{XY}$ is the expected value over the joint space. Under MSE, the best estimator for $f_\rho(X)$ is found by minimizing the loss quantities $(f(X)-Y)^2$ over the training samples for any measurable function $f: \mathcal{X} \to \mathbb{R}$. The drawback is that the quadratic greatly increases for values away from the $f(X) = Y$ line [9,10], which leads to amplify the contribution of samples that are far away from the mean value of the noise distribution. Since the heavy-tailed distributions usually have a few very large values compared to the other values of the data set, the learning solution for regression will be slanted by these samples. Therefore, MSE is optimal for Gaussian distributed residuals while it does not work well for noise distributions that have outliers or with peaks in the tails.

As a generalized similarity measure, correntropy has attracted increasing attention in machine learning fields [11–13]. Correntropy between two random variables X and Y involves all the even moments of $(X-Y)$ (see Section 2.1 for details), while MSE only concerns the second order statistics which depends heavily on the assumption of Gaussianity.

In this paper, a novel SSR algorithm is introduced, which contains a loss function and a manifold regularizer based on correntropy. The theoretical understanding is investigated for the proposed approach. Our main contribution is that we establish an excess generalization error bound, which shows that the proposed formulation is consistent and has a fast convergence rate with $O(l^{\epsilon-1})$ decay. Here l is the labeled data number and ϵ is a small parameter tending to zero. For two sequences $(e_v)_v$ and $(h_v)_v$, $v = 1, 2, \ldots$, we write $e_v = O(h_v)$ if there exists a constant $D > 0$ such that $e_v \le Dh_v$.

Compared with the existing outcomes of SSL, the rate in our paper is more rapid than the sparse semi-supervised method in [14] with the order of $O(l^{-\frac{1}{2}})$, and rate (3.14) in [15]. For supervised methods which

* Corresponding author at: School of Science, Hubei University of Technology, Wuhan 430068, China.
E-mail addresses: zuolingcc@gmail.com (L. Zuo), wangyulong6251@gmail.com (Y. Wang).

http://dx.doi.org/10.1016/j.neucom.2016.10.023
Received 14 April 2016; Accepted 23 October 2016
Available online 26 October 2016
0925-2312/

learn from the labeled data, our convergence rate is also faster than the supervised algorithm under the maximum correntropy criterion (MCC) with the order of $O(l^{-\frac{2}{3}})$ in [16], and the supervised method with the minimum error entropy (MEE) criterion, e.g., $O(l^{-\frac{1}{2}})$ in [17]. In particular, our analysis does not require the interior cone condition used in [15].

The remainder of the paper is organized as follows. In Section 2, we introduce the novel SSR method with MCC. The excess generalization error bounds are provided in Section 3, along with the discussions and comparisons with related studies. Section 4 is dedicated to investigating the proposed algorithm in terms of the ℓ^2-empirical covering numbers, and giving proofs of theoretical results stated in Section 3. The experiments are implemented in Section 5. We conclude the paper in Section 6.

2. The new SSR algorithm under MCC

Firstly, we revisit the definition and some properties of correntropy [18].

2.1. Correntropy

Definition 2.1. Correntropy is a generalized similarity measure between two arbitrary scalar random variables X and Y defined by

$$\mathcal{V}(X, Y) = \mathbf{E}_{XY}[K(X, Y)]$$

where the expected value is over the joint space and $K(\cdot,\cdot)$ is any continuous positive definite kernel.

We study the special case $\mathcal{V}_\sigma(X, Y) = \mathbf{E}_{XY}[K_\sigma(X, Y)]$, where $K_\sigma(X, Y)$ is a Gaussian kernel given by $K_\sigma(X, Y) = \exp\left\{-\frac{(X-Y)^2}{\sigma^2}\right\}$ with a bandwidth parameter $\sigma > 0$. From the definition we know that correntropy is symmetric, positive, bounded, and it reaches its maximum if and only if X=Y. The following result could be directly derived from Property 3 in [10].

Property 2.2. *Correntropy $\mathcal{V}_\sigma(X, Y)$ involves all the even moments of the random variable $(X - Y)$:* $\mathcal{V}_\sigma(X, Y) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!}\mathbf{E}_{XY}\left[\frac{(X-Y)^{2n}}{\sigma^{2n}}\right]$.

It concludes that $\mathcal{V}_\sigma(X, Y)$ consists of higher moments of $(X - Y)$ while MSE(X, Y) only concerns the second order statistics which depends heavily on the assumption of Gaussianity. We use Fig. 1 to show that correntropy is a local similarity measure whereas MSE is global. By global, it means all the examples in the joint space will contribute appreciably to the value of the similarity measure while the locality means the value is primarily dictated along the X=Y line. So compared with MSE, correntropy has a strong outlier rejection ability.

2.2. The new semi-supervised algorithm under MCC

In this part, we propose the SSR method under MCC. Recall that in semi-supervised setting, l labeled data $\{(x_i, y_i)\}_{i=1}^{l}$ together with u unlabeled data $\{x_j\}_{j=l+1}^{l+u}$ are available. To learn from samples we frequently employ the Tikhonov regularization scheme [7]

$$\min_{f\in\mathcal{H}}\{\mathcal{E}_{\mathbf{z}}(f) + \lambda_1\Omega(f)\},$$

where $\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{l}\sum_{i=1}^{l}\ell(y_i, f(x_i))$ is the empirical risk with ℓ: $\mathbb{R}^2 \to \mathbb{R}_+$ a loss function. λ_1 is a nonnegative regularization parameter, Ω is a penalty functional, and $\mathcal{H}$ is the hypothesis space containing any function defined as f: $\mathcal{X} \to \mathbb{R}$.

The proposed semi-supervised method is formulated as

$$f_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_K}\{\mathcal{E}^\sigma_{\mathbf{z}}(f) + \lambda_1\|f\|_K^2 + \lambda_2\mathcal{I}^\sigma_{\mathbf{z}}(f)\} \tag{2}$$

where λ_1 and λ_2 are nonnegative regularization parameters. λ is denoted as λ:=λ_1, λ_2. The first term in the right side of Eq. (2) is given by

$$\mathcal{E}^\sigma_{\mathbf{z}}(f) = \frac{1}{l}\sum_{i=1}^{l}\ell_\sigma(y_i, f(x_i)). \tag{3}$$

Here $\ell_\sigma(y, f(x)) = \sigma^2\left(1 - e^{-\frac{(y-f(x))^2}{\sigma^2}}\right)$ is the correntropy induced regression loss function [16], which is based on correntropy for Gaussian kernel. As a similarity measure, correntropy describes prediction errors of f over the labeled data. To minimize these loss quantities, we come to a maximum problem of correntropy, which consequently is referred to as MCC.

In Eq. (2), $\|\cdot\|_K$ is the norm restricted in the space $\mathcal{H}_K$, which is the reproducing kernel Hilbert space (RKHS) associated with a Mercer kernel K [7]. Recall that a Mercer Kernel is defined as K: $\mathcal{X}\times\mathcal{X} \to \mathbb{R}$, which is continuous, symmetric, and positive semi-definite. $\mathcal{H}_K$ is defined to be the closure of the linear span of the set of functions $\{K_x$:=$K(x, \cdot)$: $x \in \mathcal{X}\}$ with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_K} = \langle\cdot,\cdot\rangle_K$, satisfying $\langle K_x, K_y\rangle_K = K(x, y)$. For each $f \in \mathcal{H}_K$, the reproducing property takes the form of $f(x) = \langle f, K_x\rangle_K$.

The last term in the right side of (2) is the correntropy induced manifold regularizer

$$\mathcal{I}^\sigma_{\mathbf{z}}(f) = \frac{1}{2(l+u)^2}\sum_{i,j=1}^{l+u}\ell_\sigma(f(x_i), f(x_j))W_{ij} \tag{4}$$

where W_{ij}:=$W(x_i, x_j)$ is the weight given by a function $W(x, x')$, satisfying $0 \le W(x, x') \le \omega$ with a constant $\omega > 0$ for any $x, x' \in \mathcal{X}$. For example, the Gaussian kernel $K_\sigma(X, Y) = \exp\left\{-\frac{(X-Y)^2}{\sigma^2}\right\}$ with a bandwidth para-

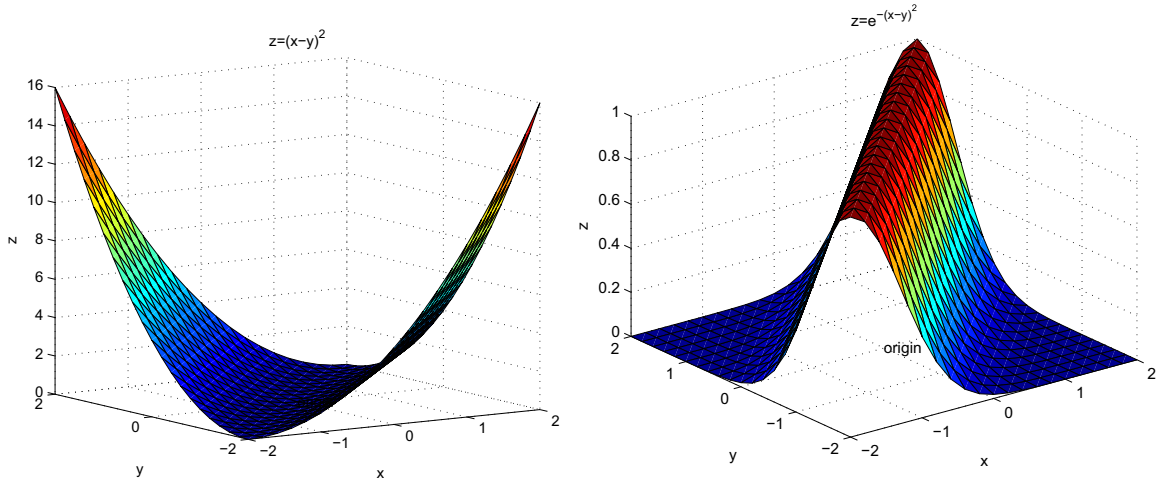

Fig. 1. The MSE and correntropy for the Gaussian kernel with $\sigma = 1$ in the joint space.