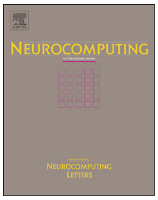
Neurocomputing 160 (2015) 85–92

Contents lists available at ScienceDirect

Neurocomputing

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



Regularized maximum correntropy machine

Jim Jing-Yan Wang[a], Yunji Wang[b], Bing-Yi Jing[c], Xin Gao[a,*]

[a] Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

[b] Electrical and Computer Engineering Department, The University of Texas at San Antonio, San Antonio, TX 78249, USA

[c] Department of Mathematics, Hong Kong University of Science and Technology, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 20 December 2013
Received in revised form
24 July 2014
Accepted 15 September 2014
Available online 12 February 2015

Keywords:
Pattern classification
Label noise
Maximum Correntropy Criteria
Regularization

ABSTRACT

In this paper we investigate the usage of regularized correntropy framework for learning of classifiers from noisy labels. The class label predictors learned by minimizing transitional loss functions are sensitive to the noisy and outlying labels of training samples, because the transitional loss functions are equally applied to all the samples. To solve this problem, we propose to learn the class label predictors by maximizing the correntropy between the predicted labels and the true labels of the training samples, under the regularized Maximum Correntropy Criteria (MCC) framework. Moreover, we regularize the predictor parameter to control the complexity of the predictor. The learning problem is formulated by an objective function considering the parameter regularization and MCC simultaneously. By optimizing the objective function alternately, we develop a novel predictor learning algorithm. The experiments on two challenging pattern classification tasks show that it significantly outperforms the machines with transitional loss functions.

1. Introduction

The classification machine design has been a basic problem in the pattern recognition field. It tries to learn an effective predictor to map the feature vector of a sample to its class label [1–9]. We study the supervised multi-class learning problem with L classes. Suppose we have a training set denoted as $\mathcal{D}=\{(x_i,y_i)\}, i=1,\dots,N$, where $x_i=[x_{i1},\dots,x_{iD}]^\top\in\mathbb{R}^D$ is the D dimensional feature vector of the i-th training sample, and $y_i\in\{1,\dots,L\}$ is the class label of i-th training sample. Moreover, we also denote the label indicator matrix as $Y=[Y_{li}]\in\mathbb{R}^{L\times N}$, and $Y_{li}=1$ if $y_i=l$, and -1 otherwise. We try to learn L class label predictors $\{f_\theta^l(x)\}, l=1,\dots,L$, for the multi-class learning problem, where $f_\theta^l(x)$ is the predictor for the l-th class and θ is its parameter. Given a sample x_i, the output of the l-th predictor is denoted as $f_\theta^l(x_i)$, and we further denote the prediction result matrix as $F_\theta=[F_{\theta li}]\in\mathbb{R}^{L\times N}$, and $F_{\theta li}=f_\theta^l(x_i)$. To make the prediction as precise as possible, the target of predictor learning is to learn parameter θ, so that the difference between true class labels of the training samples in Y and the prediction results in F_θ could be minimized, while keeping the complexity of the predictor as low as possible. To measure how well the prediction results fit the true class label indicator, several loss functions $L(F_\theta,Y)$ could be considered to compare the prediction results in F_θ against the true class labels of the training samples in Y, such as the 0–1 loss function, the square loss function, the hinge loss function, and the logistic loss function. We summarize various loss functions in Table 1.

These loss functions introduced in Table 1 have been used widely in various learning problems. One common feature of these loss functions is that a sample-wise loss function is applied to each training sample equally and then the losses of all the samples are summed up to obtain the final overall loss. The sample-wise loss functions are of exactly the same form with the same parameter (if they have parameters). The basic assumption behind this loss function is that the training samples are of the same importance. However, due to the limitation of the sampling technology and noises occurred during the sampling procedure, there are some noisy and outlying samples in real-world applications. If we use the transitional loss functions listed in Table 1, the noisy and outlying training samples will play more important roles even than the good samples. Thus the predictors learned by minimizing the transitional loss functions are not robust to the noisy and outlying training samples, and could bring a high error rate when applied to the prediction of test samples.

Recently, regularized correntropy framework has been proposed for robust pattern recognition problems [10–13]. In [14], He et al. argued that the classical mean square error (MSE) criterion is sensitive to outliers, and introduced the correntropy to improve the robustness of the presentation. Moreover, the l_1 regularization scheme is imposed

[*] Corresponding author. Tel.: +966 12 8080323.
E-mail addresses: jimjywang@gmail.com (J.-Y. Wang), yunjiwang@gmail.com (Y. Wang), majing@ust.hk (B.-Y. Jing), xin.gao@kaust.edu.sa (X. Gao).

http://dx.doi.org/10.1016/j.neucom.2014.09.080
0925-2312/© 2015 Elsevier B.V. All rights reserved.

Table 1
Various empirical loss functions for predictor learning.

Title	Formula of $L(F_\theta, Y)$	Notes
0–1 loss	$\sum_{i,l}\mathbb{I}[F_{\theta li}Y_{li} < 0]$, where $\mathbb{I}(\cdot)$ is the indicator function and $\mathbb{I}(\cdot) = 1$ if $(\cdot)$ is true, 0 otherwise.	The 0–1 loss function is NP-hard to optimize, non-smooth and non-convex.
Square loss	$\sum_{i,l}[F_{\theta li} - Y_{li}]^2 = \|F_\theta - Y\|_2^2$	The square loss function is a convex upper bound on the 0–1 loss. It is smooth and convex, thus easy to optimize
Hinge loss	$\sum_{i,l}[1 - F_{\theta li}Y_{li}]_+ = \mathbf{1}_N^\top[\mathbf{1}_{N\times L} - F_\theta \circ Y]_+ \mathbf{1}_L$ where $[x]_+ = \max(0, x)$, $\mathbf{1}_N \in \mathbb{R}^N$ is a columnvector with all ones, and $\circ$ denotes the elementwise product of two matrices	The hinge loss function is not smooth but subgradient descent can be used to optimize it. It is the most common loss function in SVM
Logistic loss	$\sum_{i,l}\ln[1 + e^{-F_{\theta li}Y_{li}}] = \mathbf{1}_N^\top \ln\left[\mathbf{1}_{N\times L} + e^{-F_\theta \circ Y}\right]\mathbf{1}_L$	This loss function is also smooth and convex, and is usually used in regression problem.

on the correntropy to learn robust and sparse representations. Inspired by their work, we propose to use the regularized correntropy as a criterion to compare the prediction results and the true class labels. We use correntropy to compare the predicted labels and the true labels, instead of comparing the feature of test sample and its reconstruction from the training samples in He et al.'s work. Moreover, an l_2 norm regularization is introduced to control the complexity of the predictor. In this way, the predictor learned by maximizing the correntropy between prediction results and the true labels will be robust to the noisy and outlying training samples. The proposed classification Machine Maximizing the Regularized CorrEntropy, which is called RegMaxCEM, is supposed to be less sensitive to outlining samples than those with transitional loss functions. Yang et al. [15] also proposed to use correntropy to compare predicted class labels and true labels. However, in their framework, the target is to learn the class labels of the unlabeled samples in a transductive semi-supervised manner, while we try to learn the parameters for the class label predictor in a supervised manner.

The rest of this paper is structured as follows: in Section 2, we propose the regularized maximum correntropy machine by constructing an objective function based on the maximum correntropy criterion (MCC) and developing an expectation- maximization (EM) based alternative algorithm for its optimization. In Section 3, the proposed methods are validated by conducting extensive experiments on two challenging pattern classification tasks. Finally, we give the conclusion in Section 4.

2. Regularized maximum correntropy machine

In this section we will introduce the classification machine maximizing the correntropy between the predicted class labels and the true class labels, while keeping the solution as simple as possible.

2.1. Objective function

To design the predictors $f_\theta^l(x)$, we first represent the data sample x as $\tilde{x}$ in the linear space and the kernel space as

$$\tilde{x} = \begin{cases} x & \text{(linear)}, \\ K(\cdot, x) & \text{(kernel)}, \end{cases} \tag{1}$$

where $K(\cdot, x) = [K(x_1, x), \ldots, K(x_N, x)]^\top \in \mathbb{R}^N$ and $K(x_i, x_j)$ is a kernel function between x_i and x_j. Then a linear predictor $f_\theta^l(x)$ will be designed to predict whether the sample belongs to the l-th class as

$$f_\theta^l(x) = w_l^\top \tilde{x} + b_l, \quad l = 1, \ldots, L, \tag{2}$$

where $\theta = \{(w_l, b_l)\}_{l=1}^L$ is the parameters of the predictors, $w_l \in \mathbb{R}^D$ is the linear coefficient vector and $b_l \in \mathbb{R}$ is a bias term for the l-th predictor. The target of predictor designing is to find the optimal parameters to have the prediction result $f_\theta^l(x_i)$ of the i-th sample to fit its true class label indicator Y_{li} as well as possible, while keeping the solution as simple as possible. To this end, we consider the following two problems simultaneously when designing the objective function:

Prediction accuracy criterion based on correntropy: To consider the prediction accuracy, we could learn the predictor parameters by minimizing a loss function listed in Table 1 as

$$\min_\theta L(F_\theta, Y) \tag{3}$$

However, as we mentioned in Section 1, all these loss functions are applied to all the training samples equally, which is not robust to the noisy samples and outlying samples. To handle this problem, instead of minimizing a loss function to learn the predictor, we use the MCC [10] framework to learn the predictor by maximizing the correntropy between the predicted results and the true labels.

Remark 1. In previous studies, it has been demonstrated that the MCC is robust to outliers, for example, see [10]. Based on this, we assume that the predictors developed by MCC should also be insensitive to outliers.

Correntropy is a generalized similarity measure between two arbitrary random variables A and B. However, the joint probability density function of A and B is usually unknown, and only a finite number of samples of them are available as $\{(a_i, b_i)\}_{i=1}^d$. It leads to the following sample estimator of correntropy:

$$V(A, B) = \frac{1}{d}\sum_{i=1}^d g_\sigma(a_i - b_i), \tag{4}$$

where $g_\sigma(a_i - b_i) = \exp\left(-(a_i - b_i)^2/2\sigma^2\right)$ is a Gaussian kernel function, and σ is a kernel width parameter. For a learning system, MCC is defined as

$$\max_\vartheta \frac{1}{d}\sum_{i=1}^d g_\sigma(a_i - b_i) \tag{5}$$

where ϑ is the parameter to be optimized in the criterion so that B is as correlated to A as possible.

Remark 2. ϑ is usually a parameter to define B, but not the kernel function parameter σ. In the learning system, we try to learn ϑ so that with the learned ϑ, B is correlated to A. For example, in this case, A is the true class label matrix while B is the predicted class label matrix, and ϑ is the predictor parameter to define B.

To adapt the MCC framework to the predictor learning problem, we let A be the prediction result matrix F_θ parameterized by θ, and B be the true class label matrix Y, and we want to find the predictor