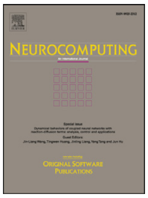
Contents lists available at ScienceDirect

Neurocomputing

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



Twin extreme learning machines for pattern classification

Yihe Wan[a,b], Shiji Song[a,*], Gao Huang[a], Shuang Li[a]

[a] *Department of Automation, Tsinghua University, Beijing 100084, China*
[b] *Naval Academy of Armament, Beijing 100161, China*

ARTICLE INFO

Article history:
Received 27 April 2016
Revised 21 April 2017
Accepted 22 April 2017
Available online xxx

Communicated by Huiyu Zhou

Keywords:
Twin extreme learning machine
Pattern classification
Extreme learning machine
Twin support vector machine
Nonparallel separating hyperplane

ABSTRACT

Extreme learning machine (ELM) is an efficient and effective learning algorithm for pattern classification. For binary classification problem, traditional ELM learns only one hyperplane to separate different classes in the feature space. In this paper, we propose a novel *twin extreme learning machine* (TELM) to simultaneously train two ELMs with two nonparallel classification hyperplanes. Specifically, TELM first utilizes the random feature mapping mechanism to construct the feature space, and then two nonparallel separating hyperplanes are learned for the final classification. For each hyperplane, TELM jointly minimizes its distance to one class and requires it to be far away from the other class. TELM incorporates the idea of twin support vector machine (TSVM) into the basic framework of ELM, thus TELM could have the advantages of the both algorithms. Moreover, compared to TSVM, TELM has fewer optimization constraint variables but with better classification performance. We also introduce a successive over-relaxation technique to speed up the training of our algorithm. Comprehensive experimental results on a large number of datasets verify the effectiveness and efficiency of TELM.

1. Introduction

Single hidden layer feedforward networks (SLFNs) have been extensively studied, and a large number of algorithms have been proposed based on SLFNs in the past few decades. These algorithms can be roughly divided into three categories. The first category adopts gradient methods to optimize the weights in the network, such as error back-propagation algorithm [1], Levenberg Marquardt algorithm [2], and neuron by neuron algorithm [3]. However, these gradient-based methods usually suffer from slow convergence or local minima issues.

The second one corresponds to the standard optimization based methods, such as the support vector machine (SVM) or its variants [4–8]. Due to its powerful classification and approximation capabilities, SVM is widely adopted in the classification and regression problems. To the specific, SVM constructs two parallel support hyperplanes which can separate the positive and negative data points very well. In order to reduce the generalization errors, SVM aims to find the solution of maximizing the width of two parallel support hyperplanes while minimizing the training errors by solving a quadratic programming problem (QPP).

The third category mainly contains the least square based methods, such as all kinds of extreme learning machines (ELMs) [9–12]. Because of the effectiveness and efficiency, ELM has been extensively studied recently. ELM has two remarkable features: (1) Unlike the conventional function approximation approaches, ELM randomly generates the input weights and hidden layer biases, and fixes them without tuning iteratively. Then the output weights can be determined analytically and efficiently. (2) ELM aims to minimize both training errors and the norm of output weights, which leads ELM to generalize well on the unseen testing data. Compared with the gradient-based algorithms, ELM spends much less time on training and tends to achieve much better generalization performance. Many studies [11] show that ELM can be comparable with or even better than the standard SVM in terms of the prediction accuracy. Moreover, since solving the least squares problem is faster than solving the QPP, ELM is usually much more efficient than SVM.

For binary classification problem, comparing ELM with SVM from the perspective of classification mechanism, ELM and its variants focus on finding a separating hyperplane, which passes through the origin of the ELM random feature space [13]. Whereas SVM aims to learn two parallel support hyperplanes to separate the testing data apart. Though ELM and SVM become more and more popular in a wide range of domains recently. In some cases, it is still difficult to achieve satisfying performance only by exploiting one separating hyperplane

[*] Corresponding author.
E-mail addresses: wanyh12@mails.tsinghua.edu.cn (Y. Wan), shijis@mail.tsinghua.edu.cn (S. Song), huang-g09@mails.tsinghua.edu.cn (G. Huang), l-s12@mails.tsinghua.edu.cn (S. Li).

http://dx.doi.org/10.1016/j.neucom.2017.04.036
0925-2312/© 2017 Published by Elsevier B.V.

or two parallel separating hyperplanes, for such as the cross data.

To expand the applicability of SVM, a famous twin support vector machine (TSVM) [14] is proposed, which targets at deriving two nonparallel separating hyperplanes. Each hyperplane tends to reach the smallest distance to one of the two classes and makes its distance to the other class as large as possible. By solving two reduced-sized QPPs, TSVM could learn faster than the standard SVM. The variants of TSVM [15–17] have been widely studied and exploited in the classification fields.

In this paper, motivated by TSVM, we propose a novel twin extreme learning machine (TELM) algorithm, which aims to learn two nonparallel separating hyperplanes in the ELM feature space for data classification. For each hyperplane, TELM minimizes its distance to one of the two classes and requires it to be far away from the other class. In order to alleviate over-fitting problems, TELM allows an acceptable training error by minimizing a regularization term jointly. Specifically, TELM tries to minimize both the training error and the sum of squares of the distance from one hyperplane to one of the two classes. Thus, TELM simultaneously trains two ELMs based on the optimization method, and has inherited the merits of ELM and TSVM.

It is worth noting that TELM is different from the methods proposed by Ning et al. [18], which aim to construct a prediction interval. In [18], two twin ELM models: regularized asymmetric least squares ELM (RALS-ELM) and asymmetric Bayesian ELM (AB-ELM) are proposed. In RALS-ELM, the asymmetric least squared error loss function with different weights is used, and AB-ELM exploits asymmetric Gaussian distribution as the likelihood function of the model output. RALS-ELM and AB-ELM could obtain two similar ELMs by setting a pair of weights. Thus, an upper-bound and a lower-bound regression curve can be deduced by RALS-ELM and AB-ELM, respectively, which leads to calculate the prediction interval. However, TELM mainly deal with the classification problem.

Similar to TSVM, the proposed TELM generates two nonparallel separating hyperplanes by solving a pair of QPPs. However, they are different in several aspects: (1) for the objective function and the corresponding constraints, the bias b is not required in TELM, whereas in TSVM, the bias b is an important optimization item; (2) in terms of two nonparallel separating hyperplanes in TELM, they pass through the origin, however, in TSVM they are basically not passing through the origin; (3) random feature mapping is adopted in the TELM, i.e., TELM initializes an network with random input weights and biases, and projects training sample into the random feature space; (4) in TELM, many nonlinear continuous functions can be utilized as activation functions, and thus TELM can be trained on diverse nonlinear feature mappings. We have evaluated the proposed algorithm on a large number of datasets comprehensively, and compare it with several related state-of-the-art algorithms. Experimental results manifest that TELM can outperform other algorithms in terms of the prediction accuracy and efficiency.

The rest of the paper is organized as follows: in Section 2, we give a brief review of SVM, TSVM and ELM. The proposed twin extreme learning machine algorithm is proposed in Section 3. In Section 4, the complete TELM and some remarks are described. Then, we give discussions about TELM in Section 5. Section 6 describes experimental details and results, and we conclude this paper in Section 7.

2. Background

In this section, we will briefly introduce SVM, TSVM and ELM.

2.1. Support vector machine

Suppose we have training data $\{\boldsymbol{x}_i, t_i\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the input pattern and $t_i \in \{+1, -1\}$ is the corresponding output. SVM aims to minimize the generalization error by maximizing the margin between two parallel support hyperplanes. When the input patterns are strictly linearly separable, to maximize the separating margin $2/\|\boldsymbol{w}\|$ is equivalent to:

$$\begin{aligned} \min_{\boldsymbol{w},b} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \\ s.t. \quad & t_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geq 1, \end{aligned} \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$.

When the input patterns are not linearly separable, we can utilize a mapping function $\boldsymbol{\psi}(\boldsymbol{x})$: $\boldsymbol{x}_i \to \boldsymbol{\psi}(\boldsymbol{x}_i)$ to project the input pattern $\boldsymbol{x}_i$ from the original input space to a SVM feature space Ψ. Here, the relax variable ξ_i associated with the ith input pattern can be introduced, and we can rewritten (1) as:

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^N \xi_i \\ s.t. \quad & t_i(\boldsymbol{w}\cdot\boldsymbol{\psi}(\boldsymbol{x}_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \end{aligned} \tag{2}$$

where C is a penalty coefficient. The dual function of (2) is

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N t_i t_j \alpha_i \alpha_j \boldsymbol{\psi}(\boldsymbol{x}_i)\cdot\boldsymbol{\psi}(\boldsymbol{x}_i) - \sum_{i=1}^N \alpha_i \\ s.t. \quad & \sum_{i=1}^N t_i\alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, \end{aligned} \tag{3}$$

where α_i is the Lagrange multiplier corresponding to the input data $\boldsymbol{x}_i$. Then the decision function of SVM can be calculated as:

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{s=1}^{N_s} \alpha_s t_s \boldsymbol{\psi}(\boldsymbol{x})\cdot\boldsymbol{\psi}(\boldsymbol{x}_s) + b\right), \tag{4}$$

where α_s corresponds to support vectors $\boldsymbol{x}_s$, N_s is the number of support vectors.

2.2. Twin support vector machine

Assume that the matrix $\boldsymbol{A} \in \mathbb{R}^{m_1\times d}$ and $\boldsymbol{B} \in \mathbb{R}^{m_2\times d}$ represent the input data belonging to class $+1$ and class -1, respectively, where m_1, m_2 are the numbers of the input data in the corresponding class.

TSVM searches for two nonparallel separating hyperplanes

$$\boldsymbol{f}_1(\boldsymbol{x}) = \boldsymbol{w}_1^T\boldsymbol{x} + b_1 = 0 \quad \text{and} \quad \boldsymbol{f}_2(\boldsymbol{x}) = \boldsymbol{w}_2^T\boldsymbol{x} + b_2 = 0, \tag{5}$$

where $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$, and $b_1, b_2 \in \mathbb{R}$.

The TSVM constructs two primal problems:

$$\begin{aligned} \min_{\boldsymbol{w}_1,b_1,\xi} \quad & \frac{1}{2}\|(\boldsymbol{A}\boldsymbol{w}_1 + \boldsymbol{e}_1 b_1)\|_2^2 + c_1\boldsymbol{e}_2^T\boldsymbol{\xi} \\ s.t. \quad & -(\boldsymbol{B}\boldsymbol{w}_1 + \boldsymbol{e}_2 b_1) + \boldsymbol{\xi} \geq \boldsymbol{e}_2 \\ & \boldsymbol{\xi} \geq 0, \end{aligned} \tag{6}$$

and

$$\begin{aligned} \min_{\boldsymbol{w}_2,b_2,\boldsymbol{\eta}} \quad & \frac{1}{2}\|(\boldsymbol{B}\boldsymbol{w}_2 + \boldsymbol{e}_2 b_2)\|_2^2 + c_2\boldsymbol{e}_1^T\boldsymbol{\eta} \\ s.t. \quad & (\boldsymbol{A}\boldsymbol{w}_2 + \boldsymbol{e}_1 b_2) + \boldsymbol{\eta} \geq \boldsymbol{e}_1 \\ & \boldsymbol{\eta} \geq 0, \end{aligned} \tag{7}$$