Neural Networks 36 (2012) 112–119

Contents lists available at SciVerse ScienceDirect

Neural Networks

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

# Twin support vector machine with Universum data

Zhiquan Qi [a], Yingjie Tian [a,*], Yong Shi [a,b,**]

[a] Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190, China
[b] College of Information Science & Technology, University of Nebraska at Omaha, Omaha, NE 68182, USA

ARTICLE INFO

*Article history:*
Received 30 May 2012
Received in revised form 20 August 2012
Accepted 3 September 2012

*Keywords:*
Classification
Twin support vector machine
Universum

ABSTRACT

The Universum, which is defined as the sample not belonging to either class of the classification problem of interest, has been proved to be helpful in supervised learning. In this work, we designed a new Twin Support Vector Machine with Universum (called 𝔘-TSVM), which can utilize Universum data to improve the classification performance of TSVM. Unlike 𝔘-SVM, in 𝔘-TSVM, Universum data are located in a nonparallel insensitive loss tube by using two Hinge Loss functions, which can exploit these prior knowledge embedded in Universum data more flexible. Empirical experiments demonstrate that 𝔘-TSVM can directly improve the classification accuracy of standard TSVM that use the labeled data alone and is superior to 𝔘-SVM in most cases.



## 1. Introduction

Supervised learning problem with Universum samples is a new research subject in machine learning. The concept of Universum sample was firstly introduced by Weston, Collobert, Sinz, Bottou, and Vapnik (2006), owing its name to the intuition that the Universum captures a general backdrop against which a problem at hand is solved. It is defined as the sample not belonging to any of the classes the learning task concerns. For instance, considering the classification of '5' against '8' in handwritten digits recognition, '0', '1', '2', '3', '4', '5', '6', '7', '9' can be considered as Universum samples. Since it is not required to have the same distribution with the training data, the Universum is able to show some prior information for the possible classifiers. Several works have been done using the Universum samples in machine learning. In Weston et al. (2006) the authors proposed a new Support Vector Machine (SVM) framework, called 𝔘-SVM and their experimental results show that 𝔘-SVM outperforms those SVMs without considering Universum data. Sinz, Chapelle, Agarwal, and Schölkopf (2008) gave an analysis of 𝔘-SVM. Then they presented a Least Squares (LS) version of the 𝔘-SVM algorithm. Zhang, Wang, Wang, and Zhang (2008) proposed a graph based semi-supervised algorithm, which learns from the labeled data, unlabeled data and the Universum data at the same time. Other literatures also can be found in Chen and Zhang (2009), Cherkassky, Dhar, and Dai (2011) and Shen, Wang, Shen, and Wang (2011).

Recently, Jayadeva, Khemchandani, and Chandra (2007) proposed a twin support vector machine (TSVM) classifier for binary classification, motivated by GEPSVM[1] (Mangasarian & Wild, 2006). TSVMs generates two nonparallel planes such that each plane is closer to one of two classes and is at least one distance from the other. It is implemented by solving two smaller Quadratic Programming Problems (QPPs) rather than a single large QPP, which makes the learning speed of TSVM faster than that of a classical SVM. Experimental results in Jayadeva et al. (2007) and Kumar and Gopal (2008) show the TSVM outperforms both standard SVM and GEPSVM in the most case. Some extensions to the TSVM can be found in Khemchandani, Jayadeva, and Chandra (2009), Kumar and Gopal (2008, 2009), Shao and Deng (2011), Shao, Zhang, Wang, and Deng (2011) and Zhiquan Qi and Shi (2012, 2013).

Inspired by the success of TSVM, in this paper, we propose a new Twin Support Vector Machines with Universum (called 𝔘-TSVM). The proposed 𝔘-TSVM has the following compelling properties.

◇ Except for labeled data from two classes, 𝔘-TSVM exploits Universum data as well. All experiments in both Toy data, UCI datasets and TFDS datasets show that the classification accuracy of 𝔘-TSVM is better than conventional TSVM algorithms that do not use Universum data. In addition, to our knowledge, this

* Corresponding author.
** Corresponding author at: Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190, China.
*E-mail addresses:* qizhiquan@gucas.ac.cn (Z. Qi), tyj@gucas.ac.cn (Y. Tian), yshi@gucas.ac.cn (Y. Shi).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.09.004

[1] In this approach, data points of each class are proximal to one of two nonparallel planes. Each plane is generated such that it is closest to one of the two data sets and as far as possible from the other data set. Each of the two nonparallel proximal planes is obtained by a single MATLAB command as the eigenvector corresponding to a smallest eigenvalue of a generalized eigenvalue problem (Mangasarian & Wild, 2006).

is the first TSVM implementation with Universum data. We also show that the TSVM are the special cases of $\mathfrak{U}$-TSVM. This provides an alternative explanation for the success of $\mathfrak{U}$-TSVM.

◇ The area of Universum data is defined by using two Hinge Loss functions. The definition is more flexible than that of $\mathfrak{U}$-SVM, which can more fully exploit the information embedded in Universum data to construct the final classifier. Fig. 2 in Section 3 gave the intuitive geometric interpretations.

The remaining parts of the paper are organized as follows. Section 2 briefly introduces the background of SVM and TSVM; Section 3 describes the detail of $\mathfrak{U}$-TSVM; In the Section 4, we show experiments of $\mathfrak{U}$-TSVM on various data sets. We conclude this work in Section 5.

## 2. Background

### 2.1. Support Vector Classification (SVC) (Vapnik, 1995)

For classification about the training data

$$T = \{(x_1, y_1), \ldots, (x_l, y_l)\} \in (\Re^n \times \mathcal{Y})^l, \tag{1}$$

where $x_i \in \Re^n$, $y_i \in \mathcal{Y} = \{1, -1\}$, $i = 1, \ldots, l$. Linear SVM is to solve the following primal QPP

$$\begin{aligned} &\min_{w,b,\xi} \ \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{l}\xi_i, \\ &\text{s.t. } y_i((w \cdot x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, i = 1, 2, \ldots, l, \end{aligned} \tag{2}$$

where $C$ is a penalty parameter and $\xi_i$ are the slack variables. The goal is to find an optimal separating hyperplane

$$(w \cdot x) + b = 0, \tag{3}$$

where $x \in \Re^n$. The Wolf Dual of (2) can be expressed as

$$\begin{aligned} &\max_{\alpha} \ \sum_{j=1}^{l}\alpha_j - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j (x_i \cdot x_j)\alpha_i\alpha_j \\ &\text{s.t. } \sum_{i=1}^{l} y_i\alpha_i = 0, \quad 0 \leq \alpha_i \leq C, i = 1, \ldots, l, \end{aligned} \tag{4}$$

where $\alpha \in \Re^l$ are Lagrangian multipliers. The optimal separating hyperplane of (3) can be given by

$$w = \sum_{i=1}^{l}\alpha_i^* y_i x_i, \qquad b = \frac{1}{N_{sv}}\left(y_j - \sum_{i=1}^{N_{sv}}\alpha_i^* y_i (x_i \cdot x_j)\right), \tag{5}$$

where $\alpha^*$ is the solution of the dual problem (4), $N_{sv}$ represents the number of support vectors satisfying $0 < \alpha < C$. A new sample is classified as $+1$ or $-1$ according to the finally decision function $f(x) = \text{sgn}((w \cdot x) + b)$.

### 2.2. Twin Support Vector Machine (TSVM) (Jayadeva et al., 2007)

Consider a binary classification problem of $l_1$ positive points and $l_2$ negative points ($l_1 + l_2 = l$). Suppose that data points belong to positive class are denoted by $A \in \Re^{l_1 \times n}$, where each row $A_i \in \Re^n$ represents a data point. Similarly, $B \in \Re^{l_2 \times n}$ represents all of the data points belong to negative class. For the linear case, the TSVM (Jayadeva et al., 2007) determines two nonparallel hyperplanes:

$$\begin{aligned} &f_+(x) = (w_+ \cdot x) + b_+ = 0 \quad \text{and} \\ &f_-(x) = (w_- \cdot x) + b_- = 0, \end{aligned} \tag{6}$$

where $w_+, w_- \in \Re^n$, $b_+, b_- \in \Re$. Here, each hyperplane is closer to one of the two classes and is at least one distance from the other. A new data point is assigned to positive class or negative class depending upon its proximity to the two nonparallel hyperplanes. Formally, for finding the positive and negative hyperplanes, the TSVM optimizes the following two respective QPPs:

$$\begin{aligned} &\min_{w_+,b_+,\xi} \ \frac{1}{2}\|Aw_+ + e_+b_+\|^2 + c_1 e_-^\top \xi, \\ &\text{s.t. } -(Bw_+ + e_-b_+) + \xi \geq e_-, \quad \xi \geq 0, \end{aligned} \tag{7}$$

and

$$\begin{aligned} &\min_{w_-,b_-,\eta} \ \frac{1}{2}\|Bw_- + e_-b_-\|^2 + c_2 e_+^\top \eta, \\ &\text{s.t. } (Aw_- + e_+b_-) + \eta \geq e_+, \quad \eta \geq 0, \end{aligned} \tag{8}$$

where $c_1, c_2 \geq 0$ are the pre-specified penalty factors, $e_+, e_-$ are vectors of ones of appropriate dimensions. By introducing the Lagrangian multipliers, the Wolfe dual of QPPs (7) and (8) can be represented as follows:

$$\begin{aligned} &\max_{\alpha} \ e_-^\top\alpha - \frac{1}{2}\alpha^\top G(H^\top H)^{-1}G^\top\alpha \\ &\text{s.t. } 0 \leq \alpha \leq c_1 e_-, \end{aligned} \tag{9}$$

and

$$\begin{aligned} &\max_{\beta} \ e_+^\top\beta - \frac{1}{2}\beta^\top P(Q^\top Q)^{-1}P^\top\beta \\ &\text{s.t. } 0 \leq \beta \leq c_2 e_+, \end{aligned} \tag{10}$$

where $G = [B\ e_-], H = [A\ e_+], P = [A\ e_+]$ and $Q = [B\ e_-]$, $\alpha \in \Re^{m_2}$, $\beta \in \Re^{m_1}$ are Lagrangian multipliers.

The non-parallel hyperplanes (6) can be obtained from the solutions $\alpha$ and $\beta$ of (9) and (10) by

$$\begin{aligned} &v_1 = -(H^\top H)^{-1}G^\top\alpha, \quad \text{where } v_1 = [w_+^\top\ b_+]^\top, \\ &v_2 = -(Q^\top Q)^{-1}P^\top\beta, \quad \text{where } v_1 = [w_-^\top\ b_-]^\top. \end{aligned} \tag{11}$$

For the nonlinear case, we can refer to Jayadeva et al. (2007).

## 3. Universum-Twin Support Vector Machine (linear $\mathfrak{U}$-TSVM)

### 3.1. Linear case

We firstly give the formal representation of classification problem with Universum. Suppose that the training set $\tilde{T}$ consists of two parts:

$$\tilde{T} = T \bigcup U, \tag{12}$$

where the symbol $\bigcup$ means the union of sets;

$$\begin{aligned} &T = \{(x_1, y_1), \ldots, (x_l, y_l)\} \in (\Re^n \times \mathcal{Y})^l, \\ &U = \{x_1^*, \ldots, x_u^*\} \in \Re^n, \end{aligned} \tag{13}$$

with $x_i \in \Re^n$, $y \in \mathcal{Y} = \{-1, 1\}$, $i = 1, \ldots, l$ and $x_j^* \in \Re^n$, $j = 1, \ldots, u$. The goal is to induce a real-valued function

$$y = \text{sgn}(g(x)), \tag{14}$$

to infer the label $y$ corresponding to any sample $x$ in $\Re^n$ space.

$\mathfrak{U}$-SVM uses the $\varepsilon$-insensitive loss for Universum:

$$\frac{1}{2}\|w\|_2^2 + c\sum_{i=1}^{l}\varphi_\varepsilon[y_i f_{w,b}(x_i)] + d\sum_{j=1}^{u}\rho[f_{w,b}(x_j^*)], \tag{15}$$

where $\varphi_\varepsilon[t] = \max\{0, \varepsilon - t\}$ is the hinge loss function, prior knowledge embedded in the Universum

$$\rho[t] = \rho_{-\varepsilon}[t] + \rho_{-\varepsilon}[-t] \tag{16}$$