Neurocomputing 189 (2016) 33–42

Contents lists available at ScienceDirect

Neurocomputing

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

Semi-supervised support vector classification with self-constructed Universum

Yingjie Tian [a,b], Ying Zhang [c], Dalian Liu [d,*]

[a] Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190, China
[b] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China
[c] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100190, China
[d] Department of Basic Course Teaching, Beijing Union University, Beijing 100101, China

ARTICLE INFO

Article history:
Received 12 March 2015
Received in revised form
16 October 2015
Accepted 15 November 2015
Communicated by Yongdong Zhang
Available online 26 November 2015

Keywords:
Semi-supervised
Classification
Universum
Support vector machine

ABSTRACT

In this paper, we propose a strategy dealing with the semi-supervised classification problem, in which the support vector machine with self-constructed Universum is iteratively solved. Universum data, which do not belong to either class of interest, have been illustrated to encode some prior knowledge by representing meaningful concepts in the same domain as the problem at hand. Our new method is applied to seek more reliable positive and negative examples from the unlabeled dataset step by step, and the Universum support vector machine($\mathfrak{U}$-SVM) is used iteratively. Different Universum data will result in different performance, so several effective approaches are explored to construct Universum datasets. Experimental results demonstrate that appropriately constructed Universum will improve the accuracy and reduce the number of iterations.

1. Introduction

In many traditional supervised learning, we acquire the decision function only through learning labeled dataset, however, in some applications of machine learning, such as image retrieval [1], text classification [2], natural language parsing [3], abundant amounts of unlabeled data can be cheaply and automatically acquired. Even if we can label samples manually, it will be labor-intensive and very time consuming. In such situation, the traditional supervised learning usually goes worse with the lacking of enough supervised information. Semi-supervised learning (SSL) [4–9] has attracted an increasing amount of interests which addresses this problem by using large amount of unlabeled data, together with the labeled data, to build better classifier.

Semi-supervised learning problem: Given a training set

$$T=\{(x_1,y_1),\ldots,(x_l,y_l)\}\bigcup\{x_{l+1},\ldots,x_{l+q}\},\tag{1}$$

where $x_i\in\Re^n, y_i\in\{-1,1\}, i=1,\ldots,l$, $x_i\in\Re^n, i=l+1,\ldots,l+q$, and the set $x_{1+1},\ldots,x_{l+q}$ is a collection of unlabeled inputs known to belong to one of the classes, predict the outputs $y_{1+1},\ldots,y_{l+q}$ for $\{x_{l+1},\ldots,x_{l+q}\}$ and find a real function $g(x)$ in R^n such that the output y for any input x can be predicted by

$$f(x)=\mathrm{sgn}(g(x)).\tag{2}$$

The motivation of semi-supervised methods is to take advantage of the unlabeled data to improve the performance and there are roughly five kinds of methods for solving the semi-supervised learning problem such as Generative methods [10–13], Graph-based methods [14–16], Co-training methods [17,18], Low-density separation methods [19,20], and Self-training methods [21–23]. Self-training is probably the earliest idea about using unlabeled data classification is a commonly used technique. Self-training is also known as self-learning, self-labeling, or bootstrapping (not to be confused with the statistical procedure with the same name). This is a wrapper-algorithm that repeatedly uses a supervised method. First, only a small labeled examples are trained in a classifier to classify unlabeled data and select most confident unlabeled points which will be added into the training set. The classifier is re-trained with the new data and the process is repeated. The idea has been used in many applications [24–26]. Our method belongs to this ideology.

Universum, which is defined as a collection of unlabeled points known not belong to any class, was firstly proposed in [27]. It has captured a general backdrop against the problem of interest and is looked forward to represent meaningful information connected with the classification task at hand. Universum dataset is easy to acquire, since there is so few requirement for it. Additionally, it can catch some prior information of the ground-truth decision

* Corresponding author.
E-mail addresses: tyj@ucas.ac.cn (Y. Tian), zhangying112@mails.ucas.ac.cn (Y. Zhang), ldlluck@sina.com (D. Liu).

http://dx.doi.org/10.1016/j.neucom.2015.11.041
0925-2312/

boundary, because it need not to have the same distribution with the training set. So several algorithms about Universum have been proposed in machine learning. In [27], the authors proposed a new SVM framework called $\mathfrak{u}$-SVM which was used to handle supervised problem and their experimental results illustrated that $\mathfrak{u}$-SVM outperforms those SVMs without considering Universum data. An analysis of $\mathfrak{u}$-SVM was given by Sinz [28]. They also presented a Least Squares (LS) version of the $\mathfrak{u}$-SVM algorithm. Zhang [29] proposed a graph based semi-supervised algorithm in which the labeled data, unlabeled data and the Universum data were simultaneously utilized to improve the performance of classification. Qi et al. [30,31] used Universum to design new nonparallel Support Vector Machine in order to improve the classification performance. Other literatures can also be found in [32,33].

Inspired by the success of $\mathfrak{u}$-SVM, in this paper, we propose an iterative support vector machine with self-constructed Universum for semi-supervised classification. It has the following advantages:

- We simultaneously utilize Universum data and iterative method to improve performance. Universum data is applied to catch some prior knowledge information of training set and the iterative method is applied to seek more reliable positive and negative examples from unlabeled dataset step by step. We train on labeled points using $\mathfrak{u}$-SVM in each step and most confident unlabeled points are added into the training set. The experiment results show that it performs better than other methods.
- Different Universum data will lead to different results, so it is crucial to construct appropriate Universum. In this paper, we only use the dataset itself to generate Universum data instead of constructing from other new dataset by considering that there would be more useful information in dataset itself. For example, considering the classification of '5' and '8' in handwritten digits recognition, we only use '5' and '8' to generate Universum data instead of '1','2','3','4','5','6','7','8','9'. Moreover, several methods to construct the appropriate Uninversum using dataset itself are also compared and suggested.

This paper is organized as follows. Section 2 dwells on the Transductive support vector machine (TSVM) [2], LapSVM (Laplacian SVM) [6] and $\mathfrak{u}$-SVM. Section 3 proposes our new method, the $\mathfrak{u}$-SVM with self-constructed Universum, termed as $\mathfrak{u}_s$-SVM. Section 4 deals with experimental results and Section 5 contains concluding remarks.

2. Background

In this section, we briefly introduce the TSVM and LapSVM for semi-supervised classification problem, and $\mathfrak{u}$-SVM for the Universum classification problem.

2.1. TSVM

TSVM aims to identify the classification model following the framework of maximum margin for both label and unlabeled examples. The popular version of TSVM [2] is to solve the following primal problem with the training set (1):

$$\begin{aligned}\min_{w,b,\xi,\xi^*,y^*}\ & \frac{1}{2}\|w\|^2+C\sum_{i=1}^{l+q}\xi_i+C^*\sum_{i=1}^{l+q}\xi_i^*,\\ \text{s. t.}\ & y_i((w\cdot x_i)+b)\geq 1-\xi_i,\quad i=1,\ldots,l,\\ & y_i^*((w\cdot x_i)+b)\geq 1-\xi_i^*,\quad i=l+1,\ldots,l+q,\\ & \xi_i\geq 0,\quad i=1,\ldots,l,\\ & \xi_i^*\geq 0,\quad i=l+1,\ldots,l+q,\end{aligned}\tag{3}$$

we can see that the above problem is a non-convex optimization problem due to the product term $y_i(w_i\cdot x_i)$ in the constraints. In order to find the approximate solutions to TSVM, extensive research efforts have been devoted [19,20,34,35]. For example, a label-switching-retraining procedure is proposed in [2] to speed up the computation.

2.2. LapSVM

The regularization framework of Laplacian support vector machine has been extended in SSL field by [6]. With the training set (1) and a kernel function $K(\cdot,\cdot)$ applied, the decision function can be obtained by

$$f=\min_{f\in H_k}\frac{1}{l}\sum_{i=1}^{l}(1-y_if(x_i))_+ +\gamma_{\mathcal{H}}\|f\|_{\mathcal{H}}^2+\gamma_{\mathcal{M}}\|f\|_{\mathcal{M}}^2\tag{4}$$

where f is the decision function,

$$f(x)=\sum_{i=1}^{l+q}\alpha_iK(x_i,x)+b,\tag{5}$$

the regularization term $\|f\|_{\mathcal{H}}^2$ can be expressed as

$$\|f\|_{\mathcal{H}}^2=\|w\|^2=(\Phi\alpha)^T(\Phi\alpha)=\alpha^TK\alpha,\tag{6}$$

and the manifold regularization is written by

$$\|f\|_{\mathcal{M}}^2=\frac{1}{(l+q)^2}\sum_{i,j=1}^{l+q}W_{ij}(f(x_i)-f(x_j))^2=f^TLf,\tag{7}$$

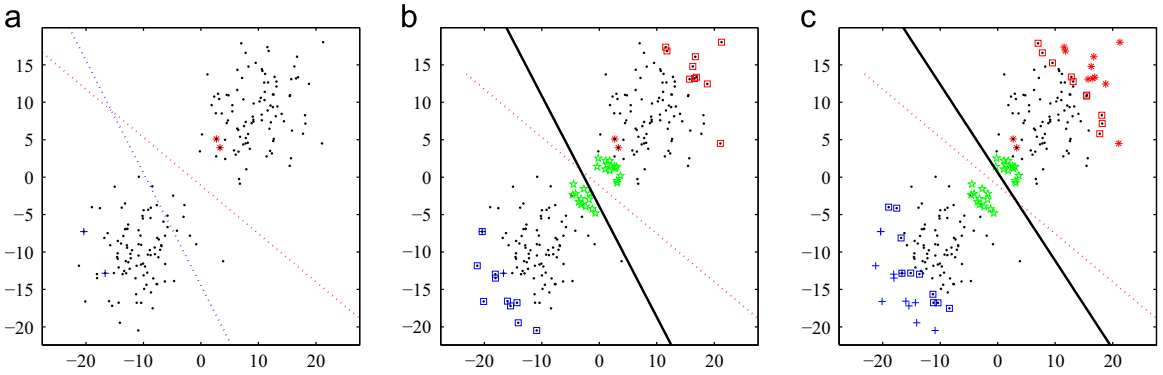

Fig. 1. Positive points (marked by "+"), negative points (marked by "∗"), unlabeled points (marked by "."), Universum points (marked by "⋆"), the ideal decision boundary (real dotted line), the decision boundary of standard SVM (blue dotted line), the decision boundary of $\mathfrak{u}$-SVM (black solid line). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)