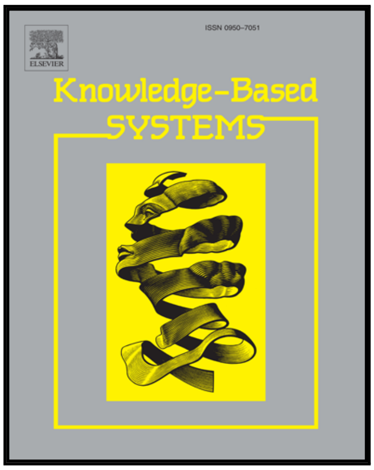
Accepted Manuscript



Revisiting Transductive Support Vector Machines with Margin Distribution Embedding

Yanchao Li, Yongli Wang, Cheng Bi, Xiaohui Jiang

PII: S0950-7051(18)30176-X
DOI: 10.1016/j.knosys.2018.04.017
Reference: KNOSYS 4299

To appear in: *Knowledge-Based Systems*

Received date: 10 June 2017
Revised date: 6 April 2018
Accepted date: 8 April 2018

Please cite this article as: Yanchao Li, Yongli Wang, Cheng Bi, Xiaohui Jiang, Revisiting Transductive Support Vector Machines with Margin Distribution Embedding, *Knowledge-Based Systems* (2018), doi: 10.1016/j.knosys.2018.04.017

Revisiting Transductive Support Vector Machines with Margin Distribution Embedding

Yanchao Li[a], Yongli Wang[a], Cheng Bi[b], Xiaohui Jiang[a]

[a] *Nanjing University of Science and Technology, Nanjing, China*
[b] *University of Edinburgh, Edinburgh, UK*

Abstract

Transductive Support Vector Machine (TSVM) is one of the most successful classification methods for semi-supervised learning (SSL). One challenge of TSVMs is that the performance degeneration is caused by unlabeled examples that are obscure or misleading for the discovery of the underlying distribution. To address this problem, we disclose the underlying data distribution and describe the margin distribution of TSVMs as the first-order (margin mean) and second-order (margin variance) statistics of examples. Since the optimization problems of TSVMs are not convex, we utilize the concave-convex procedure and variation of stochastic variance reduced gradient methods to solve them. Particularly, we propose two specific algorithms to optimize the margin distribution of TSVM via maximizing the margin mean and minimizing the margin variance simultaneously, which the generalization ability is improved and being robust to the outliers and noise. In addition, we derive a bound on the expectation of error according to the leave-one-out cross-validation estimate, which is an unbiased estimate of the probability of test error. Finally, to validate the effectiveness of the proposed method, extensive experiments are conducted on diversity datasets. The experimental results demonstrate that the performance of proposed algorithms are superior to the existing TSVMs and other semi-supervised learning methods.

Keywords: Semi-Supervised Learning, Transductive Support Vector Machine, Margin Distribution, Classification.

1. Introduction

It is quite easy to get a large number of unlabeled data in many practical settings (e.g., document classification, image classification, speech recognition), but labeled ones are fairly expensive because they require human effort. For example, people could create data and share with others on social media using digital equipments (e.g., camera, iPhone, video recorder) in the era of big data. However, plentiful of data are unlabeled. Clearly, if we only use a small amount of "expensive" labeled data, the learning system that modeled from training data cannot achieve strong generalization performance. Meanwhile, ignoring the large amount of "cheap" unlabeled data is a huge waste of data resources [1, 2, 3].

* Corresponding author
Email address: yongliwang@njust.edu.cn (Yongli Wang)