Applied Soft Computing 69 (2018) 435–442

Contents lists available at ScienceDirect

Applied Soft Computing

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

An adapted surrogate kernel for classification under covariate shift

Hyun-Je Song [a], Seong-Bae Park [b,*]

[a] School of Computer Science and Engineering, Kyungpook National University, Daegu, South Korea
[b] Kyung Hee University, Department of Computer Science and Engineering, Yongin 17104, South Korea

ARTICLE INFO

Article history:
Received 25 July 2017
Received in revised form 10 March 2018
Accepted 30 April 2018
Available online 5 May 2018

Keywords:
Covariate shift
Adapted surrogate kernel
Nyström method
Large class discrepancy

ABSTRACT

A general solution to covariate shift is to minimize a distributional discrepancy between training and test data. To solve this problem, kernel-based learning methods minimize the discrepancy in a kernel-induced feature space by transforming training data to be similar to test data and learn the transformed data in the feature space. However, when they are applied to classification tasks, transformed training data points from different classes can be mixed up in the kernel-induced space since they ignore the class labels in training data. Therefore, we propose an *adapted surrogate kernel* that is able to manage the large class discrepancy. The proposed method incorporates the class discrepancy into the surrogate kernel which tries to minimize the discrepancy in a kernel-induced feature space. To do this, we interpret the surrogate kernel with the Nyström method which allows prior knowledge to be incorporated into kernel approximation. By using class discrepancy derived from training data as prior knowledge, the adapted surrogate kernel does not keep only the role of surrogate kernel, but also the large class discrepancy among classes. The experimental results on several classification tasks including text classification and WiFi localization prove that the proposed kernel results in significant improvement in classification performance under covariate shift.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

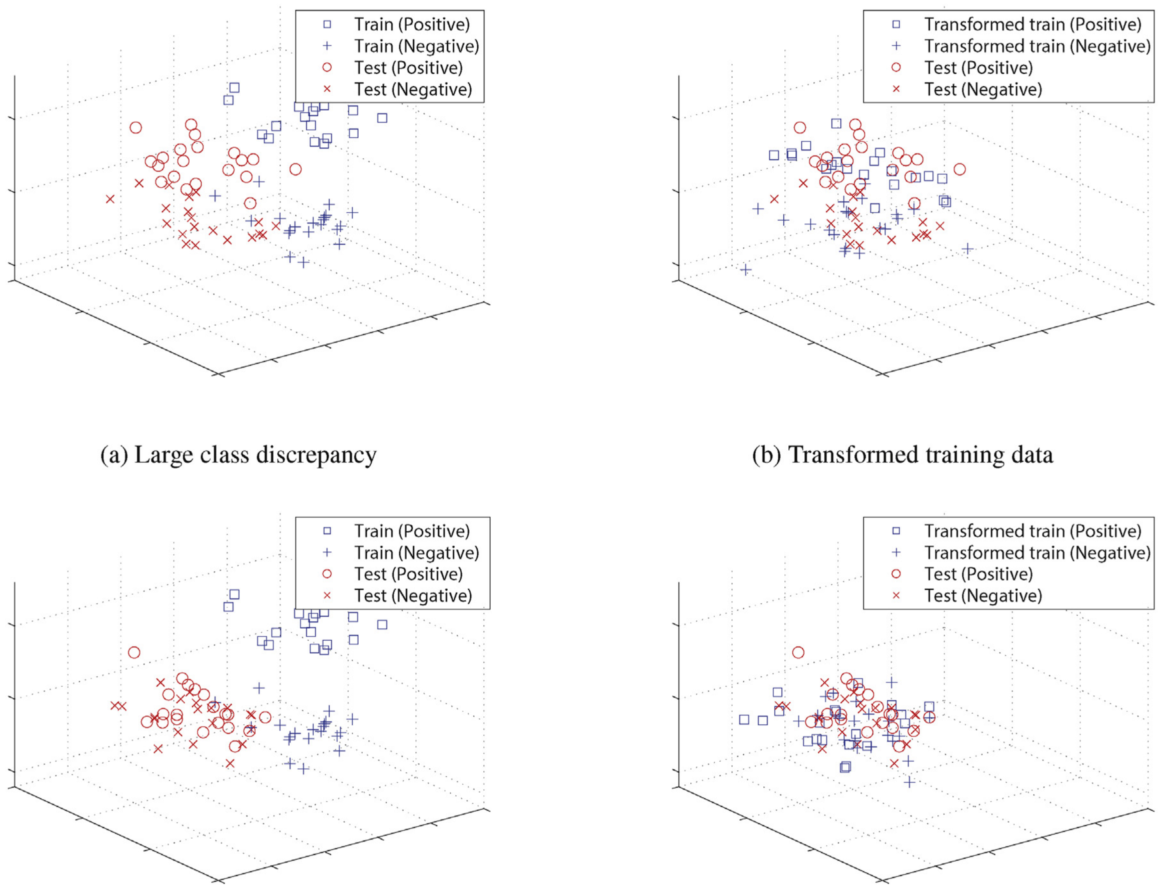
Covariate shift is a situation in which the distribution of test data is different from that of training data [1]. Most traditional machine learning techniques do not work properly under covariate shift due to their primary assumption that training and test data are drawn from an identical distribution. A general solution to this problem is to make the distribution of training data be similar to that of test data by minimizing the discrepancy of both distributions [2]. Thus, a number of studies have been proposed to minimize the discrepancy through discriminative model [3], multiple kernel learning [4], or transfer component analysis [5].

Kernel-based learning aims to solve non-linear problems by mapping data into a high-dimensional feature space. It has been also used to solve covariate shift by minimizing the distributional discrepancy between training and test data in a kernel-induced feature space. Schölkopf et al. [6] showed that the discrepancy between training and test data in the kernel-induced features space can be eliminated by aligning kernel matrices of training and test data. However, this alignment cannot be executed directly since the two matrices have different rows and columns. In order to solve this problem, several studies proposed a mediator to align the matrices such as surrogate kernel [7] or domain-invariant kernel [8]. Through the kernels, it becomes possible to minimize the data discrepancy in a kernel-induced feature space. However, in classification tasks, these methods reveal yet another problem that training data from different classes are mixed up in the kernel-induced space.

Fig. 1 illustrates the solution of covariate shift by minimizing the distributional discrepancy between training and test data in a kernel-induced feature space. When the class discrepancy among test data is large (Fig. 1(a)), the training data transformed by minimizing the distributional discrepancy are easily classified (Fig. 1(b)). However, when the class discrepancy is small (Fig. 1(c)), the transformed data points are apt to be lumped (Fig. 1(d)) even if the original training data have a large discrepancy between positive and negative class. This is because previous studies focus only on minimizing the distributional discrepancy between training and test data. As a result, it is difficult to construct a good classifier with the transformed data. In general, according to the margin theory [9] a classifier has a good discriminating ability when it is trained with the data of which class discrepancy is large. Therefore, the classification tasks under covariate shift should achieve two goals at the same time. The first goal is to minimize the discrepancy between

* Corresponding author.
E-mail address: sbpark71@khu.ac.kr (S.-B. Park).

https://doi.org/10.1016/j.asoc.2018.04.060
1568-4946/© 2018 Elsevier B.V. All rights reserved.



(a) Large class discrepancy (b) Transformed training data

(c) Small class discrepancy (d) Transformed training data

Fig. 1. An illustration of minimizing the distributional discrepancy between training and test data in a kernel-induced feature space.

training and test data, and the other is to have a large discrepancy among classes in the transformed data.

In this paper, we propose an *adapted surrogate kernel* which forces training data not only to be transformed to be similar to test data, but also to have a large discrepancy among classes after being transformed. We first adopt the Nyström method [10] to interpret the surrogate kernel which helps minimizing the distributional discrepancy between training and test data. The Nyström method is an efficient method for kernel matrix approximation, and allows prior knowledge to be incorporated into kernel approximation. Thus, the class discrepancy derived from labels of training data can be used as prior knowledge of the Nyström method. As a result, the resultant kernel, the adapted surrogate kernel, does not preserve only the role of the original surrogate kernel, but also the large discrepancy among classes.

The effectiveness of the adapted surrogate kernel has been verified through various kinds of classification tasks in single- and multi-instance learning. In single-instance learning, text classification and WiFi localization tasks are solved which are publicly used to evaluate covariate shift. Since many multi-instance learning tasks also suffer from covariate shift [11], three multi-instance learning datasets of benchmark, Reuster-21578, and natural scene images are used to check the performance of the adapted surrogate kernel. In all these tasks, the kernel-based learning with the adapted surrogate kernel achieves a competitive performance to the state-of-the-art method of each task. Especially in WiFi localization and scene image classification, the proposed method outperforms all previous methods. These results prove empirically that the consideration of large class discrepancy improves classification performance under covariate shift and the proposed surrogate kernel reflects the class discrepancy effectively into solving covariate shift of classification tasks.

The rest of the paper is organized as follows. Section 2 is devoted to related studies on covariate shift. Section 3 introduces the solution of covariate shift in kernel-based classification. Then, the proposed method, *adapted surrogate kernel*, which tries to minimize the data discrepancy and maximize the large class discrepancy is presented in Section 4. Section 5 explains the experimental settings and the results obtained by the proposed method. Finally, Section 6 draws the conclusion.

2. Related work

2.1. Covariate shift

Several approaches have been proposed to minimize the discrepancy between training and test distributions. One well-known approach is importance weighting on training data to generalize test data well [1]. It estimates an importance weight of each training instance with respect to test data. Then, learning algorithms use the weights as a cost factor of its loss function. Previous studies such as KLIEP [12] or uLSIF [13] have focused on how to estimate the weights. The main problem of these studies is that estimating data weights is carried out independently from learning the data. As a

Download English Version:

https://daneshyari.com/en/article/6903509

Download Persian Version:

https://daneshyari.com/article/6903509

Daneshyari.com