Expert Systems With Applications 107 (2018) 243–254

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Safety-aware Graph-based Semi-Supervised Learning

Haitao Gan[a,b,*], Zhenhua Li[a], Wei Wu[a], Zhizeng Luo[a], Rui Huang[c]

[a] School of Automation, Hangzhou Dianzi University, Hangzhou, China
[b] MOE Key Laboratory of Image Processing and Intelligence Control, Wuhan, China
[c] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, China

ARTICLE INFO

Article history:
Received 21 October 2017
Revised 8 March 2018
Accepted 25 April 2018
Available online 27 April 2018

Keywords:
Semi-supervised learning
Graph composite
Safety mechanism
Laplacian support vector machine

ABSTRACT

In machine learning field, Graph-based Semi-Supervised Learning (GSSL) has recently attracted much attention and many researchers have proposed a number of different methods. GSSL generally constructs a k nearest neighbors graph to explore manifold structure which may improve learning performance of GSSL. If one uses an inappropriate graph to learn a semi-supervised classifier, the performance of the classifier may be worse than that of supervised learning (SL) only trained by labeled samples. Hence, it is worthy to design a safe version to broaden the application area of GSSL. In this paper, we introduce a Safety-aware GSSL (SaGSSL) method which can adaptively select the good graphs and learn a safe semi-supervised classifier simultaneously. The basic assumption is that a graph has a high quality if the sample margin obtained by GSSL with the graph is larger than that obtained by SL. By identifying the high-quality graphs and setting the corresponding weights large, the predictions of our algorithm will approach to those of GSSL with the graphs. Meanwhile, the weights of the low-quality graphs should be small and the predictions of our algorithm will be close to those of SL. Hence the degeneration probability will be reduced and our algorithm is expected to realize the goal of safe exploitation of different graphs. Experimental results on several datasets show that our algorithm can simultaneously implement the graph selection and safely exploit the unlabeled samples.

1. Introduction

Along with the development of machine learning community, past decades have witnessed the success of Semi-Supervised Learning (SSL) (Ashfaq, Wang, Huang, Abbas, & He, 2017; Dai, Yang, Yang, Cohen, & Salakhutdinov, 2017; Gao, Ma, & Yuille, 2017; Zhu, 2010) in both theory and application field. In many practical applications, labeled samples are often hard to be collected and unlabeled ones are easy to be gathered. In this situation, SSL is designed and aims to improve learning performance using both labeled and unlabeled samples in comparison to Supervised Learning (SL) only using the labeled ones. Various SSL methods are proposed by different researchers. A core idea in SSL is how to make full use of the unlabeled samples to improve the SSL performance. Among the various SSL methods, Graph-based SSL (GSSL) has become an interest topic since it explores manifold structure of the unlabeled samples to improve learning performance. The manifold structure is generally revealed by constructing a k nearest neighbors (k-NN) graph in GSSL. Over the past years many related studies have been proposed, such as Gaussian fields and harmonic functions (Zhu, Ghahramani, & Lafferty, 2003), graph mincuts (Blum & Chawla, 2001), manifold regularization (Belkin, Niyogi, & Sindhwani, 2006), learning with ℓ_1 graph (Yan & Wang, 2009), and so on.

In different GSSL methods, parameter setting plays an important role for the learning performance, such as k in the k-NN graph, kernel parameter in computing the weight of graph. Nevertheless, how to set the parameters is a difficult task. An inappropriate parameter setting may degenerate the performance of GSSL which means GSSL performs worse than SL. Hence, it is meaningful to design safe GSSL which focuses on safe exploitation of the unlabeled samples. Recently, Wang, Wang, and Li (2016) proposed a version of safe GSSL, named Graph Semi-Supervised Learning with Instance Selection (GssIs). GssIs constructed multiple GSSL classifiers to identify the risky unlabeled samples and tried to reduce the degeneration probability. Meanwhile, Li, Wang, and Zhou (2016) proposed LargE margin grAph quality juDgement (LEAD) for safe GSSL. LEAD attempted to judge the graph quality through Semi-Supervised SVM (S3VM) which used the predictions of multiple GSSL classifiers as the input. A low-quality graph with small mar-

* Corresponding authors.
E-mail addresses: htgan@hdu.edu.cn (H. Gan), 907880745@qq.com (Z. Li), ww@hdu.edu.cn (W. Wu), luo@hdu.edu.cn (Z. Luo), ruihuang@cuhk.edu.cn (R. Huang).

https://doi.org/10.1016/j.eswa.2018.04.031
0957-4174/

gin was rarely exploited. Therefore, LEAD realized the safe exploitation of different graphs.

However, it is independent between the samples selection or graph quality judgement and semi-supervised classifier learning in GsslS or LEAD. The obtained classifier may not be the optimal one. Furthermore, recent studies (Gao & Zhou, 2013; Zhou & Zhou, 2016) conjectured that the margin distribution is crucial to the generalization performance. Zhou and Zhou (2016) proposed Cost Interval Semi-supervised Large margin Distribution Machine (cisLDM) by optimizing the margin distribution on the labeled and unlabeled samples. The results demonstrated that the margin distribution was helpful to improve the generalization of cisLDM compared to SVM and semi-supervised SVM.

Hence, we propose a safety-aware GSSL (SaGSSL) method which tries to simultaneously implement the graph composite and safe semi-supervised classifier learning. Our motivation is that the sample margin may be used to judge the graph quality. If the sample margin obtained by GSSL with some graph is larger than that obtained by SL, the graph may have a high quality and the graph weight in the graph composite should be large. Otherwise, the graph may be risky and the graph weight in the graph composite should be small. It is a reasonable assumption that a good graph may be helpful to improve the performance of GSSL. Hence, on the one hand, the predictions of SaGSSL may approach to that of GSSL with a good graph. On the other hand, the predictions of SaGSSL should approach to that of SL if the graph has a low quality. Both the graph composite and prediction constraints will be then incorporated into the objective function of GSSL as the regularization terms. Finally, the Optimization Problem (**OP**) can be solved by an alternating iterative method. In particular, we will employ SVM and Laplacian SVM (LapSVM) as the SL and SSL classifier, respectively.

The main contributions in this paper can be summarized as follows:

(1) We design a novel mechanism to develop safe GSSL and alleviate the performance degeneration of GSSL.
(2) The graph composite and safe semi-supervised classifier learning are unified in a framework and can be obtained simultaneously.
(3) The **OP** in our algorithm can be solved by an alternating iterative method.
(4) The empirical results show that our algorithm can achieve highly competitive performance compared to the GSSL methods.

The rest of the paper is organized as follows: Next section (i.e., Section 2) will present the related work. In Section 3, we will give our algorithm in detail. Section 4 will carry out a series of experiments on several datasets and analyze the results. Finally, we will give the conclusions and discuss some future work in Section 5.

2. Related work

SSL tries to achieve better learning performance than SL with the help of the unlabeled samples. The information of the unlabeled samples can be discovered through different assumptions, such as smoothness, cluster, manifold regularization, and disagreement. Based on these assumptions, different SSL methods are proposed by different researchers, such as Self-training (Gan, Sang, Huang, Tong, & Dan, 2013; Wu et al., 2018), disagreement-based methods (Zhan & Zhang, 2017; Zhou & Li, 2005), semi-supervised support vector machines (Ding, Zhu, & Zhang, 2017; Li, Kwok, & Zhou, 2009), generative models (Fujino, Ueda, & Saito, 2006; Nigam, 2001), and graph-based methods (Anis, Gamal, Avestimehr, & Ortega, 2017; Belkin et al., 2006; Zhuang et al., 2017), etc. More details can be found in Chapelle, Scholkopf, and Zien (2006), Zhu (2010), and Hady and Schwenker (2013).

However, some literatures (Singh, Nowak, & Zhu, 2009; Yang & Priebe, 2011) have verified that the unlabeled samples may degenerate the performance of SSL in terms of theory and experiment. Hence safe semi-supervised learning (Gan, Luo, Sun, et al., 2016b; Li & Zhou, 2011) has become an important topic in the SSL field. Li and Zhou (2011) firstly proposed an S3VM_*us* method where the helpful unlabeled samples were identified by a hierarchical clustering method. The helpful unlabeled samples were then classified by transductive SVM (TSVM) and the rest ones were classified by SVM. Experimental results showed that S3VM_*us* safely exploited the unlabeled samples. Kawakita and Takeuchi (2014) proposed a S3L method based on weighted likelihood which was expected to be safe in any situation. Experimental results on the regression and classification problems illustrated the effectiveness compared to SL. Gan, Luo, Meng, Ma, and She (2016a) and Gan, Luo, Sun, et al. (2016b) proposed two risk-based S3L methods, respectively. They tried to assign risk degrees to different unlabeled samples by analyzing different behaviors in SL and SSL. A risk-based regularization term was then embedded into SSL to reduce the risk. Wang, Meng, Fu, and Xue (2017) developed safe LS_S3VM based on Adjusted Cluster Assumption (ACA-S3VM). It investigated the negative effect of the inappropriate model assumption (e.g., cluster assumption). For semi-supervised regression, Li, Zha, and Zhou (2017) proposed SAFE semi-supervised Regression (SAFER). SAFER learned a safe prediction from multiple semi-supervised regressors and achieved the desired performance.

Except the above-mentioned learning methods which focus on the classification and regression problems, semi-supervised clustering (Gan, Sang, & Huang, 2015; Pei, Fern, Tjahja, & Rosales, 2016; Qian et al., 2017) is another important kind of learning. Unlike semi-supervised learning which utilizes the unlabeled samples to help train a classifier, semi-supervised clustering focuses on how to make use of prior knowledge to improve clustering performance. The common used prior knowledge includes sample labels and pair-wise constraints (i.e., must-link and cannot-link). Semi-supervised clustering can generally be divided into the following categories: (1) constraint-based approach; (2) distance-based approach. The constraint-based approach (Basu, Banerjee, & Mooney, 2002; Gan et al., 2015) mainly studies how to initialize the cluster centers or revise the objective function to guide the clustering process. The distance-based approach (Fukui, Ono, Megano, & Numao, 2013; Kalintha, Ono, Numao, & ichi Fukui, 2017; Yin, Shu, & Huang, 2012) studies how to learn a distance measure based on the given prior knowledge. For more details, please see the surveys presented in Bair (2013) and Grira, Crucianu, and Boujemaa (2005).

3. The details of our algorithm

In this section, we will discuss how to safely exploit the unlabeled samples in our algorithm.

3.1. Motivation

In the GSSL-like methods (e.g., LapSVM), graph construction plays a key role in the learning performance. The performance of GSSL heavily relies on the parameter setting, such as nearest neighbors k and Gaussian kernel width σ. GSSL with an inappropriate parameter setting may perform worse than SL. Moreover, since the actual data distribution is unknown, manifold structure revealed by the graph with the optimal parameters may not be inconsistent with the actual data distribution. The graph constructed by the labeled and unlabeled samples will degenerate the performance of GSSL. Fig. 1 shows a toy example on synthetic data. k is set to 5 and σ is selected from $\{2^{-4}, 2^{-2}, 1\}$. From this figure, one can