Applied Soft Computing 59 (2017) 448–461

Contents lists available at ScienceDirect

Applied Soft Computing

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

Semi-supervised learning using hidden feature augmentation

Wenlong Hang [a,b], Kup-Sze Choi [c], Shitong Wang [a,*], Pengjiang Qian [a]

[a] School of Digital Media, Jiangnan University, Wuxi, Jiangsu, China
[b] Department of Computer and Technology, Nanjing University of Technology, Nanjing, Jiangsu, China
[c] School of Nursing, Hong Kong Polytechnic University, Hunghom, Hong Kong

ARTICLE INFO

Article history:
Received 23 March 2016
Received in revised form 6 May 2017
Accepted 8 June 2017
Available online 13 June 2017

Keywords:
Semi-supervised learning
Cluster assumption
Manifold assumption
Hidden features
Joint probability distribution

ABSTRACT

Semi-supervised learning methods are conventionally conducted by simultaneously utilizing abundant unlabeled samples and a few labeled samples given. However, the unlabeled samples are usually adopted with assumptions, e.g., cluster and manifold assumptions, which degrade the performance when the assumptions become invalid. The reliable hidden features embedded in both the labeled and the unlabeled samples can potentially be used to tackle this issue. In this regard, we investigate the feature augmentation technique to improve the robustness of semi-supervised learning in this paper. By introducing an orthonormal projection matrix, we first transform both the unlabeled and labeled samples into a shared hidden subspace to determine the connections between the samples. Then we utilize the hidden features, the raw features, and zero vectors determined to develop a novel feature augmentation strategy. Finally, a hidden feature transformation (HTF) model is proposed to compute the desired projection matrix by applying the maximum joint probability distribution principle in the augmented feature space. The effectiveness of the proposed method is evaluated in terms of the hinge and square loss functions respectively, based on two types of semi-supervised classification formulations developed using only the labeled samples with their original features and hidden features. The experimental results have demonstrated the effectiveness of the proposed feature augmentation technique for semi-supervised learning.

1. Introduction

Semi-supervised learning finds applications in various domains, such as machine learning, pattern recognition, image processing, computer vision and bioinformatics. The performance of semi-supervised learning is usually dependent on the availability of abundant labeled samples, without which promising model learning performance is difficult to achieve. In many real-world applications, collecting full classes of labeled samples is labor-intensive or even impracticable, whereas acquiring a large amount of unlabeled data is relatively feasible. Hence, semi-supervised learning using unlabeled data has received considerable attention [1,2,5,8]. In this approach, the intrinsic structure of the data is critical to the performance. For unlabeled data, it is necessary make assumptions on the intrinsic data structure but the validity can adversely affect the learning performance, as demonstrated in many empirical studies [30]. Specifically, the wide variation in modalities and distributions in different datasets, or even the variation of data distribution among different clusters within a dataset, makes it impractical to accurately model every dataset based on a few common and straightforward assumptions. For example, in manifold learning [3], Laplacian matrix is enlisted to depict the manifold structure existing in a dataset. Nevertheless, different structures could be disclosed with different choices of k-nearest-neighbors. Therefore, it is of significance to investigate more reliable and robust strategies for using unlabeled samples in semi-supervised learning. This is the motivation underlying the research in this paper.

It has been illustrated empirically that for a classifier, the negative influence incurred by feature errors is far less than that by incorrect labels [4]. That is, under a certain assumption, e.g. manifold preservation, the process of automated labeling may result in wrongly tagged labels which can propagate and seriously affect the performance of the classifier. Ideally, a robust semi-supervised learning method should take advantages of both labeled and unlabeled samples while avoiding the negative effects due to incorrect labels. However, achieving these goals simultaneously is difficult. In practice, classification approaches safeguarding invalid assumptions often produce little or even no improvement in classification performance. The need to make assumptions in existing semi-

* Corresponding author.
E-mail address: wxwangst@aliyun.com (S. Wang).

http://dx.doi.org/10.1016/j.asoc.2017.06.017
1568-4946/

supervised learning algorithms is indeed a major hurdle against effective leveraging of the unlabeled samples.

The adverse effect due to attribution error is essentially less than that caused by class label error [4]. Thus, it is more beneficial to make use of the hidden features based on the complete unlabeled samples information, rather than relying on brute-force assumptions in the model training process. In fact, hidden features also play an important role in the human cognition. The theory of adaptive control of thought (ACT) developed by John Anderson is a well-known theory in the community of cognitive psychology [21]. According to ACT, declarative knowledge and procedural knowledge are assumed to be two critical atomic components of human cognition. Explicit declarative knowledge is not always accessible. It is more common that based on the existing declarative knowledge, along with some unconscious data inference (hidden features) and retrieval of information, procedural knowledge gradually can generates new declarative knowledge. In other words, human can unconsciously abstract hidden features from the procedural knowledge to infer and create new declarative knowledge. Enlightened by the learning approach of human cognition, we adopt an analogy to propose the research in this paper.

In this paper, the problem of semi-supervised learning is investigated from the perspective of hidden feature augmentation. We firstly introduce an orthonormal projection matrix to transform both the labeled and unlabeled samples into a common subspace such that the connection of objects (to be classified) belonging to the same class can be maximized. According to the principle of maximum joint probability distribution between the labeled and unlabeled samples in the augmented feature space, we then propose the Hidden Feature Transformation (HFT) model to obtain the hidden feature projection matrix shared by them. Finally, semi-supervised SVM classification formulations are developed respectively based on two typical loss functions. To the best of our knowledge, up to date, no endeavor is taken to do semi-supervised classification study from the perspective of the maximum joint probability distribution between the labeled and unlabeled samples. The merits of the proposed approach are highlighted as follows.

1) With the hidden feature augmentation strategy, a new mechanism to effectively leverage unlabeled samples in semi-supervised learning is first proposed. Instead of brute-force assumptions, hidden features embedded in both the labeled and unlabeled samples are mined and utilized. The proposed approach avoids the propagation of labeling errors commonly existing in semi-supervised learning, which enables practical and reliable utilization of then unlabeled data by establishing connections between the labeled and unlabeled data.
2) The novel HFT model is proposed based on the maximum joint probability principle to extract hidden features in the samples. It guarantees the labeled and unlabeled data in the augmented feature space have similar data distributions. In addition, the desired hidden space projection matrix can be obtained analytically.
3) In the two proposed semi-supervised classification formulations, the augmented labeled instances and their original and hidden features are only involved, which make it easy to solve the classification problems.
4) Besides classification, the proposed hidden space augmentation based semi-supervised learning mechanism can also be conveniently used in other applications, such as semi-supervised clustering, regression, and fuzzy inference.

The remainder of this paper is organized as follows. A brief review of the related work is provided in Section 2. The proposed semi-supervised learning framework and the feature augmentation modality, as well as the estimation of the hidden space projection matrix and the semi-supervised classification formulations are presented in Section 3. Extensive experimental studies on the proposed method and the associated analyses are discussed in Section 4. Conclusions and the possible avenues for future research are given in the last section.

2. Related works

Semi-supervised classification methods attempt to improve the performance of classifier training by exploiting both the relatively ample yet unlabeled data (with potential to disclose the intrinsic data pattern), and a small amount of labeled samples available. Numerous semi-supervised classification approaches have been developed in the past decades [5–7]. Most of them rely on making two major types of assumptions on the unlabeled data, namely, the cluster assumption and the manifold assumption [1,5,8]. The cluster assumption

presumes that similar instances have the same class labels and the classification boundaries should pass through low density regions [9–12,17]. The manifold assumption supposes that the data are distributed in some low dimensional manifolds and similar instances have the same data feature. For example, data distribution and geometric structure are generally depicted by the Laplacian graph [3,13–15].

However, empirical studies have shown that in some cases, utilization of unlabeled data indeed decreases the learning performance since the assumptions made cannot be met [16,30,31]. This is evident from the example demonstrated in Fig. 1, where two categories of objects – round-shaped objects and quadrilateral objects – are presented in Fig. 1(a); and the classification results achieved with different assumptions on a specific dataset are shown in Fig. 1(b). On the one hand, because of the rough shape similarity between the rectangles and the ellipses and that between the circles and the squares, making the cluster assumption may result in the classification hyperplane shown with the blue dotted line in Fig. 1(b). On the other hand, based on the manifold assumption, another classification hyperplane, shown with the red dotted line in the figure, may be obtained according to the different data densities of the different classes. The figure illustrates intuitively the effect of assumptions on semi-supervised classification using unlabeled data.

In some assumption-based methods, the pseudo-label generation trick is typically applied to part of the unlabeled data to expand the labeled dataset [1,5,8,35]. However, if the pseudo-labels generated based on the assumptions are erroneous, information due to the wrong labels can propagate iteratively and eventually degrade the classification performance significantly. A data editing technique has been proposed as a counter measure [27]. However, this technique relies on the neighboring information and only works well in some dense-data scenarios.

Other semi-supervised learning methods concerning data leveraging as well as negative influence avoidance have also been investigated. For example, Wang et al. developed a safety-aware semi-supervised learning (SA-SSCCM) method [18] which is a compromise between the modified cluster assumption and least-square support vector machine (LS-SVM) [20] in order to tackle the sole dependence of the cluster assumption. The safe semi-supervised SVMs (S4VMs) was also proposed to only exploit the candidate low-density separators to assist model training, under the assumption that the ground-truth label could be attained by one of the low-density separators obtained [19]. This method attempts to achieve better performance based on the extrinsic data density property, which, to some extent, is equivalent to the assumption of density-based spatial cluster [32].