Information Fusion 38 (2017) 43–54

Contents lists available at ScienceDirect

Information Fusion

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

Full Length Article

Multi-view learning overview: Recent progress and new challenges

Jing Zhao[a], Xijiong Xie[a], Xin Xu[b], Shiliang Sun[a,*]

[a] Department of Computer Science and Technology, East China Normal University, 3663 North Zhongshan Road, Shanghai 200062, PR China
[b] College of Mechatronics and Automation, National University of Defense Technology, Changsha, 410073, PR China

ARTICLE INFO

Article history:
Received 2 December 2016
Revised 14 February 2017
Accepted 20 February 2017
Available online 21 February 2017

Keywords:
Multi-view learning
Statistical learning theory
Co-training
Co-regularization
Margin consistency

ABSTRACT

Multi-view learning is an emerging direction in machine learning which considers learning with multiple views to improve the generalization performance. Multi-view learning is also known as data fusion or data integration from multiple feature sets. Since the last survey of multi-view machine learning in early 2013, multi-view learning has made great progress and developments in recent years, and is facing new challenges. This overview first reviews theoretical underpinnings to understand the properties and behaviors of multi-view learning. Then multi-view learning methods are described in terms of three classes to offer a neat categorization and organization. For each category, representative algorithms and newly proposed algorithms are presented. The main feature of this survey is that we provide comprehensive introduction for the recent developments of multi-view learning methods on the basis of coherence with early methods. We also attempt to identify promising venues and point out some specific challenges which can hopefully promote further research in this rapidly developing field.

1. Introduction

Multi-view data are very common in real world applications. Many data are often collected from different measuring methods as particular single-view data cannot comprehensively describe the information of all examples. For instance, for images and videos, color information and texture information are two different kinds of features, which can be regarded as two-view data. In web page classification, there are often two views for describing a given web page: the text content of the web page itself and the anchor text of any web page linking to this web page. It is significant to make good use of the information from different views. A well designed multi-view learning strategy may bring performance improvements.

Multi-view learning aims to learn one function to model each view and jointly optimizes all the functions to improve the generalization performance. A naive solution for multi-view learning considers concatenating all multiple views into one single view and applies single-view learning algorithms directly. However, the drawbacks of this method are that the over-fitting problem will arise on comparatively small training sets and the specific statistical property of each view is ignored. A noteworthy merit for multi-view learning is that performance on a natural single view could still be improved by using manually generated multiple views. It is important and promising to study multi-view learning methods.

Since our last review paper on multi-view machine learning [1] that was published in early 2013, multi-view learning has made great progress and developments. No matter from the perspective of utilizing data information from multiple views or from the perspective of the machine learning branches being applied to, the newly proposed multi-view learning methods show advantages to some extent. These multi-view learning methods may inspire methodological research and practical applications as well. Therefore, it is necessary to introduce the recent developments of multi-view learning, and analyze their characteristics as well as promising applications. Compared with the previous review paper, the content and structure in this paper are brand new. First, we provide comprehensive introduction for the more recent developments of multi-view learning methods on the basis of coherence with early methods. Further, in order to show a clear structure of the multi-view learning methods, the multi-view learning methods are summarized through a new kind of categorization from a relatively high level. In addition, many additional useful datasets and software packages are introduced to offer helpful advice. Finally, we discuss several latest open problems and challenges which may provide promising venues for future research.

Specifically, in this paper, multi-view learning methods are divided into three major categories: co-training style algorithms, co-regularization style algorithms and margin-consistency style algorithms. 1) Co-training style algorithms are enlightened by

* Corresponding author.
E-mail addresses: jzhao2011@gmail.com (J. Zhao), shiliangsun@gmail.com, slsun@cs.ecnu.edu.cn (S. Sun).

http://dx.doi.org/10.1016/j.inffus.2017.02.007
1566-2535/

co-training [2]. Co-training is one of the earliest methods for multi-view learning for which learners are trained alternately on two distinct views with confident labels for the unlabeled data. For example, co-EM [3], co-testing [4], and robust co-training [5] belong to this co-training style algorithm. 2) For co-regularization style algorithms, the disagreement between the discriminant or regression functions of two views is regarded as a regularization term in the objective function. Sparse multi-view SVMs [6], multi-view TSVMs [7], multi-view Laplacian SVMs [8] and multi-view Laplacian TSVMs [9] are representative algorithms. 3) Besides the two conventional style algorithms, margin-consistency style algorithms are recently proposed to make use of the latent consistency of classification results from multiple views [10–13]. They are realized under the framework of maximize entropy discrimination (MED) [14]. Different from the co-regularization style algorithms which make restrictions on the discriminant or regression functions from multiple views, margin-consistency style algorithms model the margin variables of multiple views to be as close as possible, and constrain that the product of every output variable and discriminant function should be greater than every margin variable. Particularly, in the margin-consistency style algorithms, the values of multiple views' discriminant functions may have large difference.

Besides the latest proposed multi-view learning strategies, some detailed multi-view learning algorithms are successively put forward for specific machine learning tasks. These algorithms can be summarized as multi-view transfer learning [15–17], multi-view dimensionality reduction [18–20], multi-view clustering [21–28], multi-view discriminant analysis [29,30], multi-view semi-supervised learning [8,9] and multi-task multi-view learning [31–35].

This overview aims to review key advancements in the field of multi-view learning on theoretical progress and the latest methodologies, and also point out future directions. The remainder of this paper is organized as follows. In Section 2, we introduce theoretical progress on multi-view learning, primarily focusing on PAC-Bayes bounds of multi-view learning. Section 3 surveys representative multi-view learning approaches in terms of three strategies of utilizing multi-view data information, and also provides the corresponding recent application progress. In Section 4, we describe widely used multi-view data sets and representative software packages which can provide supports for experimental purpose. In Section 5, we present some challenging problems which may be helpful for promoting further research of multi-view learning. Concluding remarks are given in Section 6.

2. Theoretical progress on multi-view learning

In order to understand the characteristics and performance of multi-view learning approaches, some generalization error analysis was successively provided, which is based on PAC-Bayes theory and Rademacher complexity theory. Here we introduce two kinds of recently proposed generalization error analysis, PAC-Bayes bounds and Rademacher complexity based generalization error bounds.

2.1. PAC-Bayes Bounds

Probably approximately correct (PAC) analysis is a basic and very general method for theoretical analysis in machine learning. It has been applied in co-training [36,37]. PAC-Bayes analysis is a related technique for data-dependent theoretical analysis, which often gives tight generation bounds [38]. Blum and Mitchell [39] presented the original co-training algorithm for semi-supervised classification and gave a PAC style analysis for justifying the effectiveness of co-training. They showed that when two prerequisite assumptions that (1) each view is sufficient for correct classification and (2) the two views of any example are conditionally independent given the class label are satisfied, PAC learning ability on semi-supervised learning holds with an initial weakly useful predictor trained from the labeled data. However, the second assumption of co-training tends to be too rigorous for many practical applications. Thus several weaker assumptions have been considered [40,41]. The PAC generalization bound for co-training provided by Dasgupta et al. [36] shows that the generalization error of a classifier from each view is upper bounded by the disagreement rate of the classifiers from the two views.

Recently, Sun et al. [42] proposed multiple new PAC-Bayes bounds for co-regularization style multi-view learning methods, which are the first application of PAC-Bayes theory to multi-view learning. They made generalization error analysis for both supervised and semi-supervised multi-view learning methods.

2.1.1. Supervised multi-view PAC-Bayes bounds

PCA-Bayes analysis for multi-view learning requires making assumptions for the distributions of weight parameters. The distribution on the concatenation of the two weight vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ is assumed as their individual product multiplied by a weight function which measures how well the two weights agree averagely on all examples. That is, the prior is $P([\mathbf{u}_1^\top, \mathbf{u}_2^\top]^\top) \propto P_1(\mathbf{u}_1)P_2(\mathbf{u}_2)V(\mathbf{u}_1, \mathbf{u}_2)$, where $P_1(\mathbf{u}_1)$ and $P_1(\mathbf{u}_2)$ are Gaussian distributions with zero mean and identity covariance, and $V(\mathbf{u}_1, \mathbf{u}_2) = \exp\left\{-\frac{1}{2\sigma^2}\mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)}(\mathbf{x}_1^\top\mathbf{u}_1 - \mathbf{x}_2^\top\mathbf{u}_2)^2\right\}$.

To specialize the PAC-Bayes bound for multi-view learning, they considered classifiers of the form $c(\mathbf{x}) = \text{sign}(\mathbf{u}^\top\phi(\mathbf{x}))$ where $\mathbf{u} = [\mathbf{u}_1^\top, \mathbf{u}_2^\top]^\top$ is the concatenated weight vector from two views, and $\phi(\mathbf{x})$ can be the concatenated $\mathbf{x} = [\mathbf{x}_1^\top, \mathbf{x}_2^\top]^\top$ itself or a concatenation of maps of $\mathbf{x}$ to kernel-induced feature spaces. Note that $\mathbf{x}_1$ and $\mathbf{x}_2$ indicate features of one example from the two views, respectively. For simplicity, they use the original features to derive their results, though kernel maps can be implicitly employed as well.

According to the setting, the classifier prior is fixed to be

$$P(\mathbf{u}) \propto \mathcal{N}(\mathbf{0}, \mathbf{I}) \times V(\mathbf{u}_1, \mathbf{u}_2), \tag{1}$$

where function $V(\mathbf{u}_1, \mathbf{u}_2)$ makes the prior place a large probability mass on parameters with which the classifiers from two views agree well on all examples averagely. The posterior is chosen to be of the form

$$Q(\mathbf{u}) = \mathcal{N}(\mu\mathbf{w}, \mathbf{I}), \tag{2}$$

where $\|\mathbf{w}\| = 1$. Define $\tilde{\mathbf{x}} = [\mathbf{x}_1^\top, -\mathbf{x}_2^\top]^\top$. The following is obtained

$$P(\mathbf{u}) \propto \mathcal{N}(\mathbf{0}, \mathbf{I}) \times V(\mathbf{u}_1, \mathbf{u}_2) \propto \exp\left\{-\frac{1}{2}\mathbf{u}^\top\left(\mathbf{I} + \frac{\mathbb{E}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top)}{\sigma^2}\right)\mathbf{u}\right\}.$$

That is, $P(\mathbf{u}) = \mathcal{N}(\mathbf{0}, \Sigma)$ with $\Sigma = \left(\mathbf{I} + \frac{\mathbb{E}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top)}{\sigma^2}\right)^{-1}$.

Suppose $\dim(\mathbf{u}) = d$. Given the above prior and posterior, their divergence is characterized by the following lemma.

Lemma 1. *[42]*

$$KL(Q(\mathbf{u})\|P(\mathbf{u})) = \frac{1}{2}\left(-\ln\left(\left|\mathbf{I} + \frac{\mathbb{E}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top)}{\sigma^2}\right|\right) + \frac{1}{\sigma^2}\mathbb{E}[\tilde{\mathbf{x}}^\top\tilde{\mathbf{x}} + \mu^2(\mathbf{w}^\top\tilde{\mathbf{x}})^2] + \mu^2\right). \tag{3}$$

In addition, they provided and proved two inequalities on the involved logarithmic determinant function, which are very important for the subsequent multi-view PAC-Bayes bounds.