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/© 2017 Elsevier B.V. All rights reserved.

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 $c(\mathbf{x}) = \mathrm{sign}(\mathbf{u}^\top\phi(\mathbf{x}))$ 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

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

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]*

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

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.

Lemma 2.

$$-\ln\left|\mathbf{I}+\frac{\mathbb{E}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top)}{\sigma^2}\right| \le -d\ln\mathbb{E}\left[\left|\mathbf{I}+\frac{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top}{\sigma^2}\right|^{1/d}\right], \tag{4}$$

$$-\ln\left|\mathbf{I}+\frac{\mathbb{E}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top)}{\sigma^2}\right| \le -\mathbb{E}\ln\left|\mathbf{I}+\frac{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top}{\sigma^2}\right|. \tag{5}$$

Denote $R=\sup_{\tilde{\mathbf{x}}}\|\tilde{\mathbf{x}}\|$. From inequality (4), a new multi-view PAC-Bayes bound is derived as follows.

Theorem 1 (Multi-view PAC-Bayes bound 1)**.** *Consider a classifier prior given in (1) and a classifier posterior given in (2). For any data distribution $\mathcal{D}$, for any $\delta \in (0, 1]$, with probability at least $1-\delta$ over $S\sim\mathcal{D}^m$, the following inequality holds*

$$\forall \mathbf{w},\mu : KL_+(\hat{E}_{Q,S}||E_{Q,\mathcal{D}}) \le \frac{-\frac{d}{2}\ln\left[f_m-(\sqrt[d]{(R/\sigma)^2+1}-1)\sqrt{\frac{1}{2m}\ln\frac{3}{\delta}}\,\right]_+ + \frac{H_m}{2\sigma^2}+\frac{(1+\mu^2)R^2}{2\sigma^2}\sqrt{\frac{1}{2m}\ln\frac{3}{\delta}}+\frac{\mu^2}{2}+\ln\left(\frac{m+1}{\delta/3}\right)}{m},$$

where

$$f_m=\frac{1}{m}\sum_{i=1}^m\left|\mathbf{I}+\frac{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top}{\sigma^2}\right|^{1/d},$$

$$H_m=\frac{1}{m}\sum_{i=1}^m[\tilde{\mathbf{x}}_i^\top\tilde{\mathbf{x}}_i+\mu^2(\mathbf{w}^\top\tilde{\mathbf{x}}_i)^2],$$

and $\|\mathbf{w}\|=1$.

From the bound formulation, we see that if $(\mathbf{w}^\top\tilde{\mathbf{x}}_i)^2$ is small, that is, if the outputs of the two views tend to agree, the bound will be tight. Note that, although the formulation of f_m involves the outer product of feature vectors, it can actually be represented by the inner product, which is obvious through the following determinant equality [42],

$$\left|\mathbf{I}+\frac{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top}{\sigma^2}\right| = \frac{\tilde{\mathbf{x}}_i^\top\tilde{\mathbf{x}}_i}{\sigma^2}+1. \tag{6}$$

The matrix $\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top$ has rank 1 and has only one nonzero eigenvalue.

Then inequality (5) instead of (4) was used to derive a d-independent bound (see Theorem 2 below), which is independent of the dimensionality of the feature representation space.

Theorem 2 (Multi-view PAC-Bayes bound 2)**.** *Consider a classifier prior given in (1) and a classifier posterior given in (2). For any data distribution $\mathcal{D}$, for any $\delta \in (0, 1]$, with probability at least $1-\delta$ over $S\sim\mathcal{D}^m$, the following inequality holds*

$$\forall \mathbf{w},\mu : KL_+(\hat{E}_{Q,S}||E_{Q,\mathcal{D}}) \le \frac{\tilde{f}/2+\frac{1}{2}\left(\frac{(1+\mu^2)R^2}{\sigma^2}+\ln(1+\frac{R^2}{\sigma^2})\right)\sqrt{\frac{1}{2m}\ln\frac{2}{\delta}}+\frac{\mu^2}{2}+\ln\left(\frac{m+1}{\delta/2}\right)}{m},$$

where

$$\tilde{f}=\frac{1}{m}\sum_{i=1}^m\left(\frac{1}{\sigma^2}[\tilde{\mathbf{x}}_i^\top\tilde{\mathbf{x}}_i+\mu^2(\mathbf{w}^\top\tilde{\mathbf{x}}_i)^2]-\ln\left|\mathbf{I}+\frac{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top}{\sigma^2}\right|\right),$$

and $\|\mathbf{w}\|=1$.

Since this bound is independent of d and the term $\left|\mathbf{I}+\frac{\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top}{\sigma^2}\right|$ involving the outer product can be represented by the inner product through (6), this bound can be employed when the dimension of the kernelized feature space goes to infinity.

By changing the prior distribution of $\mathbf{u}$, and applying the two inequality (4) and (5) respectively, another four PAC-Bayes bounds can be deduced [42].

2.1.2. Semi-supervised multi-view PAC-Bayes bounds

PAC-Bayes analysis was considered for semi-supervised multi-view learning, where besides the m labeled examples, u unlabeled examples $U=\{\tilde{X}\}_{j=m+1}^{m+u}$ are further provided. In this case, the weight function $V(\mathbf{u}_1,\mathbf{u}_2)$ was replaced with $\tilde{V}(\mathbf{u}_1,\mathbf{u}_2)$, which has the form $\tilde{V}(\mathbf{u}_1,\mathbf{u}_2)=\exp\left\{-\frac{1}{2\sigma^2}\mathbf{u}^\top\mathbb{E}_U(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\top)\mathbf{u}\right\}$, where $\mathbb{E}_U$ means the empirical average over the unlabeled set U. There are two kinds of semi-supervised multi-view PAC-Bayes bounds by using the noninformative prior and informative prior. If the prior distribution of $\mathbf{u}$ is assumed as $P(\mathbf{u})\propto\mathcal{N}(\mathbf{0},\mathbf{I})\times V(\mathbf{u}_1,\mathbf{u}_2)$, the semi-supervised multi-view PAC-Bayes bound is derived as Theorem 11 in Sun et al. [42]. If the prior distribution of $\mathbf{u}$ is assumed as $P(\mathbf{u})\propto\mathcal{N}(\eta W_p,\mathbf{I})\times V(\mathbf{u}_1,\mathbf{u}_2)$, where $W_p=\mathbb{E}_{(\mathbf{x},y)\sim D}[y\mathbf{x}]$ with $\mathbf{x}=[\mathbf{x}_1^\top,\mathbf{x}_2^\top]^\top$, the semi-supervised multi-view PAC-Bayes bound is derived as Theorem 12 in Sun et al. [42].

2.2. Rademacher complexity based generalization error bounds

Another attempt to analyze the generalization of two-view learning was made using Rademacher complexity [43]. Farquhar et al. [44] analyzed the generalization error bounds of SVM-2K, which rely on the empirical estimate for Rademacher complexity. Szedmak and Shawe-Taylor [45] characterized the generalization performance of its extended version for semi-supervised learning. Rosenberg and Bartlett [46] gave the empirical Rademacher complexity of co-regularized least squares. Then Sindhwani and Rosenberg [47] recovered the generalization bound [47]. Sun and Shawe-Taylor [6] proposed a sparse semi-supervised learning framework using Fenchel–Legendre conjugates and instantiated an algorithm named sparse multi-view SVMs. They gave the generalization error bound of the sparse multi-view SVMs. Sun [8] also presented multi-view Laplacian SVMs whose generalization error analysis and empirical Rademacher complexity were provided as well [8].

Recently, Xu et al. [48] proposed a multi-view intact space learning algorithm, which integrates the encoded complementary information in multiple views to discover a latent intact representation of data. Simultaneously, they proposed a new definition of multi-view stability and derived the generalization error bound based on the multi-view stability and Rademacher complexity, which shows that the complementarity between multiple views is beneficial for the stability and generalization.

3. Multi-view learning methods

From the perspectives of strategies of utilizing multi-view data information, multi-view learning methods can be divided into three major categories: co-training style algorithms, co-regularization style algorithms and margin-consistency style algorithms.

Co-training style algorithms are a kind of mechanisms of multi-view learning which override on single-view learning algorithms. They are mostly used for solving semi-supervised problems. Co-training style algorithms make use of multiple views of data to iteratively learning multiple classifiers that can provide predicted labels for the unlabeled data for each other. Co-regularization style algorithms often add regularization terms of discriminant or regression function onto the objective function. The regularization terms constrain that the prediction results from multiple

Table 1
The outline of multi-view learning methods.

Category	Representatives	Applications
co-training	co-training [2,5] co-EM [3] co-testing [4] co-clustering [15]	multi-view semi-supervised learning multi-view transfer learning,
co-regularization	CCA [49] DCCA [50] MvDA [51] MULDA [30] SVM-2K [44] MvTSVM [7]	multi-view dimension reduction multi-view clustering multi-view supervised learning multi-view semi-supervised learning
margin consistency	MVMED [10] SMVMED [12] MED-2C [13]	multi-view classification

views should be close. Margin-consistency style algorithms model the margin variables from multiple views to be consistent in the framework of MED [14]. Instead of making restrictions directly on the discriminant or regression function, margin-consistency style algorithms constrain that the product of every output variable and discriminant function should be greater than every margin variable. In margin-consistency style algorithms, the values of multiple views' discriminant functions may have large difference.

We first show the outline of multi-view learning methods in Table 1 in terms of representative algorithms and applications to machine learning problems corresponding to each category. Then we will introduce the three categories of multi-view learning methods and the progress on multi-view machine learning applications in detail in the following subsections. We hope that the descriptions of the multi-view learning methods under different categories and applications under different machine learning problems could provide some inspirations for multi-view researchers.

3.1. Co-training style algorithms

Co-training was originally proposed for the problem of semi-supervised learning, in which there is access to labeled as well as unlabeled data. It considers a setting in which each example can be partitioned into two distinct views, and makes two main assumptions for success: sufficiency and conditional independence. In order to deal with more kinds of multi-view learning tasks, the idea of co-training was employed and some extended co-training style algorithms are developed such as co-EM [3], co-testing [4] and co-clustering [15]. In addition, some interesting and valuable analysis for co-training style algorithms was made, which promotes the developments of co-training.

Wang and Zhou [52] showed that co-training can work without two views when the two learners have large difference, and co-training could not improve the performance further after many learning rounds. A series of deep analysis revealed some interesting properties of co-training, for example on the large-diversity of classifiers [52], label propagation over two views [53] and co-training with insufficient views [54]. They further provided a sufficient and necessary condition for co-training to succeed with proper assumptions. Nigam and Ghani [3] showed that co-training on multiple views manually generated by random splits of features can result in performance improvements even when no natural multiple views are available. They also proposed co-EM algorithm which extends the original bootstrap method of the co-training algorithm to operate simultaneously on all unlabeled samples in an iterative batch mode [3]. Brefeld and Scheffer [55] successfully developed a co-EM version of support vector machines. Muslea et al. [56] introduced co-testing, which is a novel approach to combine active learning with multiple views. Then they combined co-testing with co-EM, and derived a novel method called co-EMT [57].

The original co-training algorithm cannot examine the reliability of labels obtained by the classifiers from each view. Even very few inaccurately labeled examples can deteriorate the performance of learned classifiers to a large extent. To overcome this drawback, Sun and Jin [58] proposed robust co-training, which integrates CCA to inspect the predictions of co-training on the unlabeled training data. Yu et al. [59] proposed an improved version of co-training called Bayesian co-training with the Bayesian undirected graphical model. The model can query <example, view> pairs to improve the learning performance. Zhao et al. [60] presented an algorithm that combines the simplicity of k-means clustering and linear discriminant analysis within a co-training scheme, which exploits labels learned automatically in one view to learn discriminative subspaces in another.

Multi-view Transfer Learning Based on Co-training Transfer learning is an emerging and active topic which learns a new task through the transfer of knowledge from a related task that has already been learned. Chen et al. [61] presented a variant of co-training for domain adaptation which connects source and target domains whose distributions can differ substantially. Xu and Sun [62,63] proposed an algorithm involving a variant of EMV-Adaboost for multi-view transfer learning and further extended it to multiple source case. Zhang et al. [64] proposed multi-view transfer learning with a large margin approach. On one hand, labeled data from the source domain are effectively utilized to construct a large margin classifier. On the other hand, data from both domains are employed to impose consistencies among multiple views. Yang and Gao [15] proposed an information-theoretical multi-view adaptation model that combines the paradigms of multi-view learning and domain adaptation based on a co-clustering framework, and aims to transfer knowledge across domains in multiple subspaces of features complementarily. They incorporated multiple views of data in a perceptive transfer learning framework and proposed a multi-view discriminant transfer learning approach for domain adaptation [16]. Tan et al. [17] proposed a novel algorithm to leverage knowledge from different views and sources collaboratively, by assuring different views from different sources to complement each other through a co-training style framework.

3.2. Co-regularization style algorithms

Co-regularization style algorithms usually add regularization terms to the objective function to make sure that data from multiple views are consistent. The regularization styles can be summarized as three different ways. (1) One is to construct linear or nonlinear transformations from the original space in different views to a new space, and constrain that the multiple transformed feature sets should be as close as possible. Typical methods for this kind of regularization are CCA based algorithms. (2) Another one is to apply the label information to the space transformation based algorithms, and add constrains for intra-class and inter-class characteristics. These kinds of algorithms are mostly based on discriminative CCA and multi-view linear discriminate analysis (LDA). (3) The third one is to combine the data and label information by use of classifiers or regressors, and regularize that the outcomes got by classifiers or regressors from multiple views should be as consistent as possible. Multi-view SVMs and multi-view twin SVMs are recently proposed and representative algorithms for this kind of regularization.

3.2.1. CCA based algorithms

One representative co-regularization style algorithm is canonical correlation analysis (CCA) [49,65,66]. CCA is an approach to correlating linear relationships between two-view feature sets. It seeks linear transformations each for one view such that the correlation between these transformed feature sets is maximized in the common subspace while regularizing the self covariance of each transformed feature sets to be small enough. The aim of CCA is to find two projection directions w_x and w_y corresponding to each view, and maximize the following linear correlation coefficient

$$\frac{cov\left(w_x^TX, w_y^TY\right)}{\sqrt{var\left(w_x^TX\right)var\left(w_y^TY\right)}} = \frac{w_x^TC_{xy}w_y}{\sqrt{\left(w_x^TC_{xx}w_x\right)\left(w_y^TC_{yy}w_y\right)}}, \tag{7}$$

where the covariance matrices C_{xy}, C_{xx} and C_{yy} are calculated as $C_{xy} = \frac{1}{n}XY^T$, $C_{xx} = \frac{1}{n}XX^T$, $C_{yy} = \frac{1}{n}YY^T$. The constant $\frac{1}{n}$ can be canceled out when calculating the correlation coefficient. Since w_x, w_y are scale-independent, the objective expressed by (7) is equivalent to the following optimization problem

$$\begin{aligned} &\max_{w_x,w_y} && w_x^TC_{xy}w_y \\ &\text{s.t.} && w_x^TC_{xx}w_x = 1,\ w_y^TC_{yy}w_y = 1. \end{aligned} \tag{8}$$

The optimal solution for the projection directions w_x and w_y can be obtained through solving a generalized eigenvalue problem as

$$\begin{bmatrix} \mathbf{0} & C_{xy} \\ C_{yx} & \mathbf{0} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix} = \lambda \begin{bmatrix} C_{xx} & \mathbf{0} \\ \mathbf{0} & C_{yy} \end{bmatrix} \begin{bmatrix} w_x \\ w_y \end{bmatrix}. \tag{9}$$

Here, **0** represents the zero vector with an appropriate number of zero elements.

It's worth mentioning that maximizing correlation as in (8) corresponds to minimizing the empirical expectation of the pattern function $g_{w_x,w_y}(X,Y) = ||w_xX - w_yY||^2$ subject to the same conditions [67]. Seen from this pattern function which constrains that the value of two views' projection functions should be as identical as possible, CCA is actually a kind of co-regularization style algorithms.

Kernel canonical correlation analysis (KCCA) [68–71] is an kernel extension of CCA for pursuing maximally correlated nonlinear projections. The desired projection vectors w_x^ϕ and w_y^ϕ can be expressed as a linear combination of all training examples in the feature space, and there exist coefficient vectors $a = [a^1, \dots, a^n]^\top$ and $b = [b^1, \dots, b^n]^\top$, such that

$$w_x^\phi = \sum_{i=1}^{n} a^i\phi_x(x_i) = \phi(X)a,\ w_y^\phi = \sum_{i=1}^{n} b^i\phi_y(y_i) = \phi(Y)b. \tag{10}$$

Substituting (10) into (8) and using the definition of the kernel matrix, one can formulate the optimization problem of KCCA as

$$\begin{aligned} &\max_{a,b} && a^TK_xK_yb \\ &\text{s.t.} && a^TK_xK_xa = 1,\ b^TK_yK_yb = 1, \end{aligned} \tag{11}$$

which can be solved in a similar way like CCA.

Bayesian CCA, Deep CCA and Tensor CCA CCA has attracted a lot of researchers in past years [72,73]. CCA has been extended to sparse CCA [74,75] and has been widely used for multi-view classification [76], clustering [77], regression [78], etc. Bach and Jordan [79] gave a probabilistic interpretation of CCA, such that the maximum likelihood estimates of the model parameters can be derived from CCA. Given this probabilistic interpretation, CCA has been extended to Bayesian CCA in fully Bayesian treatment recently [80]. It can avoid overfitting by adding regularization [81]. In addition, some extensions of probabilistic CCA models have been provided as non-Gaussian CCA, discrete CCA and mixed CCA which were adapted to applications where one or both of the data-views are either counts [82]. Deep canonical correlation analysis (DCCA) [83] is a kind of method to learn complex nonlinear transformations of two views of data such that the resulting representations are highly linearly correlated. Unlike KCCA, DCCA does not require an inner product, and has the advantages of a parametric method: training time scales well with the data size and the training data need not be referenced when computing the representations of unseen instances. CCA can be extended to multi-view CCA [84] by maximizing the sum of pairwise correlations between different views. However, the main drawback of this strategy is that only correlation information between pairs of features is explored, while high-order statistics are ignored. Luo et al. [85] develop tensor CCA (TCCA) to generalize CCA to handle any number of views in a direct and yet natural way. In particular, TCCA can directly maximize the correlation between the canonical variables of all views, and this is achieved by analyzing the high-order covariance tensor over the data from all views [86].

Multi-view Dimension Reduction As an important branch of multi-view unsupervised learning, multi-view dimension reduction seeks a low-dimensional common subspace to represent multi-view data [87]. For example, CCA is a typical multi-view dimensionality reduction method. Some new multi-view dimension reduction methods were developed by involving CCA or other space transformation methods. Chen et al. [88] proposed a multi-view latent subspace Markov network to accomplish multi-view dimension reduction. This network fulfills a weak conditional independence assumption that multi-view observations and response variables are conditionally independent given a set of latent variables. Ding and Fu [18] proposed a low-rank common subspace for multi-view data analysis, which seeks a common low-rank linear projection to mitigate the semantic gap among different views. The low-rank common projection is able to capture compatible intrinsic information across different views and also well-aligns the within-class samples from different views. As a result, it offers effective methods for robust subspace learning. White et al. [89] and Guo [19] provided a convex formulation of multi-view subspace learning. The new formulation of multi-view subspace learning allows a global solution, and can be derived with efficient optimization algorithms. More recently, a Bayesian multi-view dimensionality reduction method was proposed, where data points from different views are projected into a unified subspace without the restriction of matching data examples from these views [20]. Regularization for projection functions from different views was also employed to achieve multi-view denoising [90].

Multi-view Clustering Multi-view clustering, which aims to obtain a partition of the data in multiple views that often provide complementary information to each other, has received considerable attention in the past years. Most work was designed based on space transformation methods [77,91–95]. Recently, Liu et al. [96] presented a novel tensor-based framework for integrating heterogeneous multi-view data in the context of spectral clustering. Zhang et al. [21] proposed low-rank tensor constrained multi-view subspace clustering which regards the subspace representation matrices of different views as a tensor and the tensor is equipped with a low-rank constraint. The multi-linear relationship among multi-view data is taken into account through their tensor-based strategy. In order to deal with large-scale data clustering problems, a new robust large-scale multi-view clustering method [22] was proposed to integrate multiple representations of large scale data. Li et al. [23] presented partial multi-view clustering in the case that every view suffers from the missing of some data and results in many partial examples. Wang et al. [97] proposed a multi-view learning model to integrate all features and learn the weight for every feature with respect to each cluster individually via new joint structured sparsity-inducing norms. Some multi-view clustering algorithms based on the nonnegative matrix factorization were proposed [24,25,98]. Xia et al. [26] proposed a novel Markov chain method for robust multi-view spectral clustering, which combines the transition probability matrices constructed from each view into a shared transition probability matrix via low-rank and sparse decomposition. Based on max-product belief propagation, Zhang et al. [27] proposed a novel multi-view clustering algorithm termed multi-view affinity propagation. Diversity induced multi-view subspace clustering [28] was proposed to explore the complementary information. Some researchers have proposed multi-view clustering ensemble learning that combines different ensemble techniques for multi-view clustering [99–101]. Chikhi [102] proposed a multi-view normalized cut approach with spectral partitioning and local refinement.

3.2.2. Discriminative CCA and multi-view LDA based algorithms

Although CCA can obtain a common space for multiple views, it does not take label information into account. To learn a discriminant common space for two views, correlation discriminant analysis and discriminative canonical correlation analysis (DCCA) [50,103,104] were proposed to extend CCA by maximizing the difference of within-class and between-class variations across two views. Moreover, as extensions from LDA, multi-view Fisher discriminant analysis for binary classification problem [105,106] and generalized multi-view linear discriminant analysis (GMvDA) for multi-class classification from multiple views [107] were proposed. While GMvDA requires setting hyper-parameters for regularization, multi-view discriminant analysis (MvDA) [51] provides more direct derivation from LDA for multiple view projection matrices without any hyper-parameter. In addition, MvDA simultaneous obtains a concatenation of projection matrices from multiple views by solving a single generalized eigenvalue problem. Makihara et al. [29] described a multi-view discriminant analysis with tensor representation and applied it to cross-view gait recognition. The large-margin idea was also integrated into the Gaussian processes to discover the latent subspace shared by multiple views [108,109].

MULDA Uncorrelated LDA (ULDA) is an extension of LDA by adding some constraints into the optimization objective of LDA, so that the feature vectors extracted by ULDA could contain minimum redundancy. Multi-view uncorrelated linear discriminant analysis (MULDA) [30] was recently proposed by imposing two more constraints in each view, which extracts uncorrelated features in each view and computes transformations of each view to project data into a common subspace. Here we briefly introduce MULDA. Let (w_{x1}, w_{y1}) represent the vector pair solved by the existing multi-view LDA method which corresponds to the maximum eigenvalue. Suppose the vector pairs (w_{xj}, w_{yj}) with $j = 1, 2, \dots, r-1$ of the two-view data are obtained. MULDA aims to find the rth discriminant vector pair (w_{xr}, w_{yr}) of datasets X and Y with the following conjugated orthogonality constraints

$$w_{xr}^T S_{t_x} w_{xj} = w_{yr}^T S_{t_y} w_{yj} = 0 \quad (j = 1, 2, \dots, r-1), \tag{12}$$

where S_{t_x} and S_{t_y} represent the total scatter matrix for two views. With S_{b_x} and S_{b_y} denoting the between-class scatter matrix, the optimization problem of MULDA can be formulated as

$$\begin{aligned} \max_{w_{xr}, w_{yr}} \quad & w_{xr}^T S_{b_x} w_{xr} + w_{yr}^T S_{b_y} w_{yr} + 2\gamma w_{xr}^T C_{xy} w_{yr} \\ \text{s.t.} \quad & w_{xr}^T S_{t_x} w_{xr} + \sigma w_{yr}^T S_{t_y} w_{yr} = 1 \\ & w_{xr}^T S_{t_x} w_{xj} = w_{yr}^T S_{t_y} w_{yj} = 0 \\ & (j = 1, 2, \dots, r-1), \end{aligned} \tag{13}$$

where w_{xr} and w_{yr} represent the rth discriminant vectors of datasets X and Y, respectively. Through optimizing (13), we obtain d feature vectors for each view: $z_{xi} = w_{xi}^T X$, $z_{yi} = w_{yi}^T Y$, $i = 1, \dots, d$.

3.2.3. Multi-view SVMs and multi-view twin SVMs

Many multi-view supervised learning methods build upon SVMs. SVM-2K [44] is a representative multi-view algorithm, which combines the two views by introducing the constraint of similarity between two one-dimensional projections identifying two distinct SVMs from the two feature spaces. Assuming that the two view data are expressed through two feature projections, i.e., ϕ_1 with corresponding kernel κ_1 and ϕ_2 with corresponding kernel κ_2, the constraint is expressed as an ϵ-insensitive 1-norm regularization using slack variables η_i,

$$| < \mathbf{w}_1, \phi_1(\mathbf{x}_i) > +b_1 - < \mathbf{w}_2, \phi_2(\mathbf{x}_i) > -b_2| \leq \eta_i + \epsilon, \tag{14}$$

where $\mathbf{w}_1$, b_1 , $(\mathbf{w}_2, b_2)$ are the weight and threshold of the first (second) view's SVM.

Recently, a new method called multi-view twin support vector machines (MvTSVMs) was proposed [7]. On one view, positive examples are represented by $A_1^{'}$ and negative examples are represented by $B_1^{'}$. On the other view, positive examples are represented by $A_2^{'}$ and negative examples are represented by $B_2^{'}$. For simplicity, suppose that all e are vectors of ones of appropriate dimensions and

$$\begin{aligned} & A_1 = (A_1^{'}, e), B_1 = (B_1^{'}, e), A_2 = (A_2^{'}, e), B_2 = (B_2^{'}, e), \\ & v_1 = \begin{pmatrix} w_1 \\ b_1 \end{pmatrix}, \ v_2 = \begin{pmatrix} w_2 \\ b_2 \end{pmatrix}, u_1 = \begin{pmatrix} w_3 \\ b_3 \end{pmatrix}, \ u_2 = \begin{pmatrix} w_4 \\ b_4 \end{pmatrix}, \end{aligned} \tag{15}$$

where (w_1, b_1) and (w_2, b_2) are classifier parameters of +1 class, and (w_3, b_3) and (w_4, b_4) are classifier parameters of −1 class. The optimization problems for MvTSVMs are written as

$$\begin{aligned} \min_{v_1, v_2, q_1, q_2, \eta} \quad & \frac{1}{2}\|A_1 v_1\|^2 + \frac{1}{2}\|A_2 v_2\|^2 + c_1 e_2^\top q_1 + c_2 e_2^\top q_2 + D e_1^\top \eta \\ \text{s.t.} \quad & |A_1 v_1 - A_2 v_2| \preceq \eta, \\ & - B_1 v_1 + q_1 \succeq e_2, \\ & - B_2 v_2 + q_2 \succeq e_2, \\ & q_1 \succeq 0, \ q_2 \succeq 0, \\ & \eta \succeq 0, \end{aligned} \tag{16}$$

$$\begin{aligned} \min_{u_1, u_2, k_1, k_2, \zeta} \quad & \frac{1}{2}\|B_1 u_1\|^2 + \frac{1}{2}\|B_2 u_2\|^2 + d_1 e_1^\top k_1 + d_2 e_1^\top k_2 + H e_2^\top \zeta \\ \text{s.t.} \quad & |B_1 u_1 - B_2 u_2| \preceq \zeta, \\ & - A_1 u_1 + k_1 \succeq e_1, \\ & - A_2 v_2 + k_2 \succeq e_1, \\ & k_1 \succeq 0, \ k_2 \succeq 0, \\ & \zeta \succeq 0, \end{aligned} \tag{17}$$

where e_1 and e_2 are vectors of ones of appropriate dimensions, v_1, v_2, u_1, u_2 are classifier parameters, c_1, c_2, d_1, d_2, D, H are nonnegative parameters, and q_1, q_2, η, ζ, k_1, k_2 are slack vectors of appropriate dimensions.

Multi-view Semi-supervised Learning Semi-supervised learning, which learns from few labeled examples and a large number of unlabeled examples, is an active research direction. Its prevalence is mainly motivated by the need to reduce the expensive or time-consuming label acquisition process. Szedmak and Shawe-Taylor [45] exploited unlabeled data via multi-view learning effectively [45]. Representative multi-view semi-supervised learning methods include co-training [39], co-EM [3], multi-view sequential learning [110], Bayesian co-training [59], multi-view point cloud regularization [111], sparse multi-view SVMs [6], two-view transductive support vector machines [112], robust co-training [5], multi-view vector-valued manifold regularization [113]. The recent multi-view Laplacian SVMs [8] and multi-view Laplacian TSVMs [9] integrate the multi-view regularization with manifold regularization and bring inspiring results. Here we introduce the multi-view Laplacian TSVMs. On view one, positive examples are represented by A_1' and negative examples are represented by B_1'. On view two, positive examples are represented by A_2' and negative examples are represented by B_2'. The optimization problems of multi-view Laplacian TSVMs can be written as

$$\begin{aligned}
\min_{w_1,w_2,b_1,b_2,q_1,q_2,\eta}\ & \frac{1}{2}\|A_1'w_1+e_1b_1\|^2+\frac{1}{2}\|A_2'w_2+e_1b_2\|^2\\
& +c_1e_2^\top q_1+c_2e_2^\top q_2\\
& +\frac{1}{2}c_3(\|w_1\|^2+b_1^2+\|w_2\|^2+b_2^2)\\
& +\frac{1}{2}c_4[(w_1^\top M_1'^\top+e^\top b_1)L_1(M_1'w_1+eb_1)\\
& +(w_2^\top M_2'^\top+e^\top b_2)L_2(M_2'w_2+eb_2)]+De_1^\top\eta \qquad (18)\\
\text{s.t.}\quad & |A_1'w_1+e_1b_1-A_2'w_2-e_1b_2|\preceq\eta,\\
& -B_1'w_1-e_2b_1+q_1\succeq e_2,\\
& -B_2'w_2-e_2b_2+q_2\succeq e_2,\\
& q_1\succeq 0,\ q_2\succeq 0\\
& \eta\succeq 0,
\end{aligned}$$

$$\begin{aligned}
\min_{w_3,w_4,b_3,b_4,q_3,q_4,\zeta}\ & \frac{1}{2}\|B_1'w_3+e_2b_3\|^2+\frac{1}{2}\|B_2'w_4+e_2b_4\|^2\\
& +c_1e_1^\top q_3+c_2e_1^\top q_4\\
& +\frac{1}{2}c_3(\|w_3\|^2+b_3^2+\|w_4\|^2+b_4^2)\\
& +\frac{1}{2}c_4[(w_3^\top M_1'^\top+e^\top b_3)L_1(M_1'w_3+eb_3)\\
& +(w_4^\top M_2'^\top+e^\top b_4)L_2(M_2'w_4+eb_4)]+He_2^\top\zeta\\
\text{s.t.}\quad & |B_1'w_3+e_2b_3-B_2'w_4-e_2b_4|\preceq\zeta,\\
& -A_1'w_3-e_1b_3+q_3\succeq e_1,\\
& -A_2'w_4-e_1b_4+q_4\succeq e_1,\\
& q_3\succeq 0,\ q_4\succeq 0\\
& \zeta\succeq 0.
\end{aligned} \qquad (19)$$

M_1' includes all of labeled data and unlabeled data from view 1. M_2' includes all of labeled data and unlabeled data from view 2. L_1 is the graph Laplacian of view 1 and L_2 is the graph Laplacian of view 2. e_1, e_2 and e are vectors of ones of appropriate dimensions. w_1, b_1, w_2, b_2, w_3, b_3, w_4, b_4 are classifier parameters. c_1, c_2, c_3 and c_4 are nonnegative parameters. q_1, q_2, q_3, q_4, η and ζ are slack vectors of appropriate dimensions.

3.3. Margin-consistency style algorithms

Margin-consistency style algorithms were proposed under the consideration of the characteristics of classification in multi-view cases. Especially for large-margin classifiers, the margins between the samples and the hyperplanes well depict the relationship between the models and the data. It is a valid method to utilize consistency of multi-view data to regularize that the margins from two views are the same or have the same posteriors. The strategy of using margin consistency was firstly proposed in the framework of multi-view maximum entropy discrimination (MVMED) [10]. Some variants such as soft margin consistency based multi-view MED (SMVMED) [11] and consensus and complementarity based MED (MED-2C) [13] were also developed and obtained promising performance.

3.3.1. MVMED

Multi-view maximum entropy discrimination (MVMED) [10] was proposed as an extension of MED to the multi-view learning setting. It considers a joint distribution $p(\boldsymbol{\Theta_1}, \boldsymbol{\Theta_2})$ over the view one classifier parameter $\boldsymbol{\Theta}_1$ and view two classifier parameter $\boldsymbol{\Theta}_2$. $\boldsymbol{\gamma}$ is the shared margin variable by two views. Using the augmented joint distribution $p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma})$ and the joint prior distribution $p_0(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\gamma})$, MVMED can be formulated as follows

$$\begin{aligned}
&\min_{p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})} KL(p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})\,||\,p_0(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma}))\\
&\text{s.t.}\begin{cases}\int p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})[y_tL_1(X_t^1|\boldsymbol{\Theta}_1)-\gamma_t]d\boldsymbol{\Theta}_1d\boldsymbol{\Theta}_2d\boldsymbol{\gamma}\geq 0\\ \int p(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,\boldsymbol{\gamma})[y_tL_2(X_t^2|\boldsymbol{\Theta}_2)-\gamma_t]d\boldsymbol{\Theta}_1d\boldsymbol{\Theta}_2d\boldsymbol{\gamma}\geq 0\\ 1\leq t\leq N,\end{cases}
\end{aligned} \qquad (20)$$

where $L_1(X_t^1|\boldsymbol{\Theta}_1)$ and $L_2(X_t^2|\boldsymbol{\Theta}_2)$ are discriminant functions from two views, respectively.

Chao and Sun [11] also proposed a more flexible MVMED framework called alternative MVMED (AMVMED) [11], which considers two separate distributions $p_1(\boldsymbol{\Theta}_1)$ over $\boldsymbol{\Theta}_1$ and $p_2(\boldsymbol{\Theta}_2)$ over $\boldsymbol{\Theta}_2$, and balances KL divergences of their augmented distributions with respect to the corresponding prior distributions $p_0(\,\cdot\,)$. AMVMED is formulated as

$$\begin{aligned}
&\min_{p_1(\boldsymbol{\Theta}_1,\boldsymbol{\gamma}),\,p_2(\boldsymbol{\Theta}_2,\boldsymbol{\gamma})} \rho KL(p_1(\boldsymbol{\Theta}_1,\boldsymbol{\gamma})\,||\,p_0(\boldsymbol{\Theta}_1,\boldsymbol{\gamma}))\\
&\quad+(1-\rho)KL(p_2(\boldsymbol{\Theta}_2,\boldsymbol{\gamma})\,||\,p_0(\boldsymbol{\Theta}_2,\boldsymbol{\gamma}))\\
&\text{s.t.}\begin{cases}\int p_1(\boldsymbol{\Theta}_1,\boldsymbol{\gamma})\,[y_tL_1(X_t^1|\boldsymbol{\Theta}_1)-\gamma_t]\,d\boldsymbol{\Theta}_1d\boldsymbol{\gamma}\geq 0\\ \int p_2(\boldsymbol{\Theta}_2,\boldsymbol{\gamma})\,[y_tL_2(X_t^2|\boldsymbol{\Theta}_2)-\gamma_t]\,d\boldsymbol{\Theta}_2d\boldsymbol{\gamma}\geq 0\\ \int p_1(\boldsymbol{\Theta}_1,\boldsymbol{\gamma})d\boldsymbol{\Theta}_1=\int p_2(\boldsymbol{\Theta}_2,\boldsymbol{\gamma})d\boldsymbol{\Theta}_2\\ 1\leq t\leq N.\end{cases}
\end{aligned} \qquad (21)$$

3.3.2. SMVMED

Unlike conventional multi-view learning method, MVMED and AMVMED exploits the multiple views in a style called margin consistency, that is, to enforce margins from the two views to be identical. Although they have provided state-of-art multi-view learning performance, this margin consistency requirement may be too strong to fulfill in some cases and hinder effective model learning. It is thus interesting to explore the possibility of relaxing the requirement. The proposed soft margin consistency based MVMED (SMVMED) has achieved improvements through the relaxing of the requirement on margin consistency [12]. It assumes two different posterior distributions, $p_1(\boldsymbol{\gamma})$ and $p_2(\boldsymbol{\gamma})$, for margin variables and ensures that the KL divergence between the two distributions as small as possible. Similar to MVMED, the objective optimization problem can be expressed by

$$\begin{aligned}
&\min_{p_1(\boldsymbol{\Theta}_1,\boldsymbol{\gamma}),p_2(\boldsymbol{\Theta}_2,\boldsymbol{\gamma})} \mathbb{KL}(p_1(\boldsymbol{\Theta}_1)\,||\,p_1^0(\boldsymbol{\Theta}_1)) + \mathbb{KL}(p_2(\boldsymbol{\Theta}_2)\,||\,p_2^0(\boldsymbol{\Theta}_2))\\
&\quad +(1-\alpha)\mathbb{KL}(p_1(\boldsymbol{\gamma})\,||\,p_1^0(\boldsymbol{\gamma})) + (1-\alpha)\mathbb{KL}(p_2(\boldsymbol{\gamma})\,||\,p_2^0(\boldsymbol{\gamma}))\\
&\quad +\alpha\mathbb{KL}(p_1(\boldsymbol{\gamma})\,||\,p_2(\boldsymbol{\gamma})) + \alpha\mathbb{KL}(p_2(\boldsymbol{\gamma})\,||\,p_1(\boldsymbol{\gamma}))\\
&\text{s.t.} \begin{cases} \int p_1(\boldsymbol{\Theta}_1,\boldsymbol{\gamma})[y_t L_1(X_t^1|\boldsymbol{\Theta}_1)-\gamma_t]d\boldsymbol{\Theta}_1 d\boldsymbol{\gamma} \geq 0\\ \int p_2(\boldsymbol{\Theta}_2,\boldsymbol{\gamma})[y_t L_2(X_t^2|\boldsymbol{\Theta}_2)-\gamma_t]d\boldsymbol{\Theta}_2 d\boldsymbol{\gamma} \geq 0\\ 1 \leq t \leq N. \end{cases}
\end{aligned} \tag{22}$$

3.3.3. MED-2C

Another margin-consistency style multi-view learning method is called consensus and complementarity based MED (MED-2C) [13]. It is used for multi-view classification, which well utilizes the two principles consensus and complementarity for multi-view learning. MED-2C first transforms data from two views into a common subspace, and makes the transformed data in the new subspace identical to respect the consensus principle. Then it augments the transformed data with their original features to take into account the complementarity principle. Similar to MVMED, the objective optimization problem in MED-2C can be formulated by

$$\begin{aligned}
&\min_{P,Q}\min_{p(\boldsymbol{\Theta},\boldsymbol{\gamma})} \mathbb{KL}(p(\boldsymbol{\Theta},\boldsymbol{\gamma})\,||\,p_0(\boldsymbol{\Theta},\boldsymbol{\gamma})) + \beta||PX_1 - QX_2||_F^2\\
&\text{s.t.} \begin{cases} \int p(\boldsymbol{\Theta},\boldsymbol{\gamma})[y_t L(\tilde{X}_t^1|\boldsymbol{\Theta})-\gamma_t]d\boldsymbol{\Theta} d\boldsymbol{\gamma} \geq 0\\ \int p(\boldsymbol{\Theta},\boldsymbol{\gamma})[y_t L(\tilde{X}_t^2|\boldsymbol{\Theta})-\gamma_t]d\boldsymbol{\Theta} d\boldsymbol{\gamma} \geq 0\\ 1 \leq t \leq N, \end{cases}
\end{aligned} \tag{23}$$

where $\tilde{X}_t^1 = [[PX_t^1]^\top, {X_t^1}^\top, \mathbf{0}^\top]^\top$ and $\tilde{X}_t^2 = [[QX_t^2]^\top, \mathbf{0}^\top, {X_t^2}^\top]^\top$. MED-2C is a successful method of combining margin-consistency style algorithms and co-regularization style algorithms.

3.4. Progress and applications of combining multi-view learning algorithms

Multi-task Multi-view Learning Multi-task multi-view learning (MTMV) can learn multiple related tasks with multi-view data. He and Lawrence [114] proposed a graph-based framework which takes full advantage of information among multiple tasks and multiple views. Zhang and Huan [115] proposed a general inductive learning framework for the challenging MTMV problems using co-regularization and task relationship learning. Yang and He [31] modeled task relatedness using a normal penalty with sparse covariances to couple multiple tasks and view relatedness using matrix Dirichlet process. Two MTMV tracking methods were proposed based on joint sparse representation [32] and based on an approximate least absolute deviation [33] to exploit the related information shared between particles and views in order to obtain improved performance. However, they all tackle the classification problem. Zhang et al. [34] introduced an MTMV clustering framework which integrates within view-task clustering, multi-view relationship learning and multi-task relationship learning. To facilitate information sharing among different tasks on multi-view representation, Jin et al. [35] proposed an efficient inductive convex shared structure learning method for the MTMV problem. In real world, there exist quite a few applications where the tasks with several views correspond to different set of class labels. This new learning scenario is called MTMV learning for heterogeneous tasks [116], for which an MTMV discriminant analysis method was proposed to solve this problem.

Multi-view Ensemble Learning The goal of ensemble learning is to construct strong learners by combining weak learners to make very accurate predictions. Many algorithms have been developed and widely used, such as bagging, boosting, random subspace. Xu and Sun [117] extended Adaboost to the multi-view learning scenario and presented the embedded multi-view Adaboost algorithm (EMV-Adaboost). Sun and Zhang [118] extended a multi-view ensemble learning framework with both multiple views and multiple learners to semi-supervised learning [118] and active learning [119], respectively. Kumar and Minz [120] [121] proposed supervised feature set partitioning method and optimal feature set partitioning for performance enhancement of multi-view ensemble learning. Multi-view ensemble learning has successfully addressed the issue related to high dimensionality of the data and poem data classification using sentiwordnet [122].

4. Multi-view datasets and software packages

In order to provide experimental supports for the research on multi-view learning, we describe some widely used multi-view datasets and representative software packages.

Handwritten Digit Dataset[1] Handwritten digit dataset [7] is marked as multiple feature sets in the UCI repository. It consists of feature sets of handwritten numerals ($0 \sim 9$) extracted from a collection of Dutch utility maps. Each digit (class) digitized in binary images has 200 examples (for a total of 2000 examples) which are represented in six feature sets (views) in this dataset.

Advertisement Dataset[2] Advertisement data [10] consist of 3279 examples including 459 ads images (positive examples) and 2820 non-ads images (negative examples). The first view describes the image itself (words in the image's URL, alt text and caption), while the other view contains all other features (words from the URLs of the pages that contain the image and the image points to).

WebKB Dataset[3] WebKB data [10] consist of 1051 two-view web pages collected from computer science department web sites at four universities: Cornell University, University of Washington, University of Wisconsin, and University of Texas. There are 230 course pages and 821 non-course pages.

Multi-PIE dataset[4] Multi-PIE dataset [123] is employed to evaluate face recognition across poses. It contains more than 750,000 images of 337 people under various view points.

CUFSF Dataset[5] CUHK face sketch FERET (CUFSF) dataset [123] is used to evaluate photo-sketch face recognition. It contains 1194 subjects with lighting variations, where examples in this dataset come from only two views, photo and sketch.

HFB Dataset[6] Heterogeneous face biometrics (HFB) dataset [123] contains images from 100 subjects, which is used to evaluate visual (VIS) image vs. near infrared (NIR) image heterogeneous face recognition, where examples are only from two views, visual image and near infrared image.

Corel Images Dataset[7] Corel images dataset [124] consists of 34 categories, each with 100 images. Attribute vectors represent the images in terms of seven views, three color-related views (color histogram, moment and coherence) and four texture-related views (coarseness and directionality of tamura texture, wavelet and mrsar texture).

Software packages Besides the above valuable datasets, there are also some representative public software packages which can bring convenience to multi-view researchers. For co-training style

[1] Data are available at http://archive.ics.uci.edu/ml/datasets/Multiple+Features.

[2] Data are available at https://archive.ics.uci.edu/ml/datasets/Internet+Advertisements.

[3] Data are available at http://www.cs.cmu.edu/afs/cs/project/theo-11/www/wwkb/.

[4] Data are available at https://datahub.io/dataset/multipie.

[5] Data are available at http://mmlab.ie.cuhk.edu.hk/archive/cufsf/.

[6] Data are available at http://www.cbsr.ia.ac.cn/english/HFB&20Databases.asp.

[7] Data are available at https://archive.ics.uci.edu/ml/datasets/Corel+Image+Features.

algorithms, there is usually no universal software since they depend on specific single-view algorithms. For co-regularization style algorithms, CCA is a very simple algorithm always embedded in popular toolboxes. MvDA[8] [29] and MULDA[8] [30] are two discriminant projection methods. SVM-2K[9] [44] is often regarded as a baseline algorithm. For margin-consistency style algorithms, MVMED[10] [10], SMVMED[11] [12], MED-2C[12] [13] are recently proposed algorithms with public software available.

5. Open problems

With the needs of practical applications and the developments of machine learning methods, multi-view learning has got rapid progress. In this part, we present several open problems that can be important for future research and applications of multi-view learning.

5.1. Large-scale multi-view learning

Nowadays, a tremendous quantity of data are continually generated. It has been witnessed that many real applications involve large-scale multi-view data. For example, hundreds of hours of videos are uploaded to YouTube every minute, which appear in multiple modalities or views, namely visual, audio and text views. A large number of bilingual news are reported every day, with the description in each language as a view. It is noteworthy that most previous multi-view approaches only work on small-size data sets, which makes it difficult to handle large-scale multi-view tasks. Therefore, it is a challenge for previous approaches to deal with the task of learning with large-scale multi-view data.

There are some multi-view stereo algorithms applied in large-scale data sets [125,126]. Zhu et al. [127] concentrated on the large-scale multi-view learning for classification, and proposed the one-pass multi-view framework which goes through the training data only once without storing the entire training examples. The computing in CCA for large data sets can be very slow since it involves implementing QR decomposition or singular value decomposition of large matrices. Lu and Foster [128] introduced large-scale CCA, an iterative algorithm which can compute CCA fast on large sparse data sets [128]. Cai et al. [129] proposed a novel robust large-scale multi-view K-means clustering approach, which can be easily parallelized and performed on multi-core processors for big visual data clustering. Li et al. [22] proposed a novel large-scale multi-view spectral clustering approach based on the bipartite graph. Besides large-scale CCA and large-scale multi-view clustering, it is an urgent need to develop large-scale learning methods for some other multi-view learning algorithms. Large-scale MVMED is a potential method to be studied to handle extensive data.

5.2. Multi-view deep learning

Deep neural networks have recently demonstrated outstanding performance in a variety of tasks such as face recognition, object classification and object detection. They can significantly outperform other methods for the task of large-scale image classification. For multi-view learning, there are also some potential of improving performance through incorporating multi-view learning algorithms and deep learning methods.

So far, multi-view deep representation learning has two main strategies [130]. First, Ngiam et al. [131] proposed to extract shared representations by reconstructing both views from the view that is available at test time which is regarded as a split autoencoder. Second, Andrew et al. [132] proposed a DNN extension of CCA called deep CCA. For practical application, Zhu et al. [133] proposed a multi-view perceptron which is a deep model for learning face identity and view representations. Su et al. [134] presented a novel CNN architecture that combines information from multiple views of a 3D shape into a single and compact shape descriptor. Elhoseiny et al. [135] achieved joint object categorization and pose estimation on multi-view data through employing view-invariant representation within CNNs. Elkahky et al. [136] presented a general recommendation framework that uses deep learning to match rich user features to item features. They also showed how to extend this framework to combine data from different domains to further improve the recommendation quality. Although these methods have realized deep learning in the multi-view learning framework, there is still a lot of room to develop multi-view deep learning in terms of methodologies and applications. For example, multi-view deep Gaussian processes is a kind of interesting and challenging model.

5.3. Model design for more than two views

Many multi-view learning algorithms were proposed based on two views. Actually, multiple view data are very common in practical applications. Some existing methods for handling multiple views are variants of two-view methods. They combine all the pairwise correlations through addition operation in the objective function. However, the main drawback of this strategy is that only correlation information between pairs of features is explored, and high-order statistics are ignored.

Besides the above simple ways of combining two-view learning algorithms, some new strategies of handling multi-view cases were proposed. Among them, tensor product is an effective technique used for learning multi-view data. The proposed TCCA [85] is a valid instance of using tensor product to CCA. How to develop more richer multi-view learning algorithms with tensor product is a problem worth studying. Further, considering the variety of multi-view learning methods, it would be interesting to design algorithms for more than two views under specific settings.

5.4. Multi-view learning with incomplete views

The existing multi-view learning algorithms have shown promising performance in different applications. These algorithms usually work under the full-view assumption where data from all views are required to be observed. In some practical applications, this full-view setting is likely to be violated. For example, data from some certain views may be lost because of sensor faults or man-made errors. As a result, we can only access multi-view data with incomplete views, which brings difficulties for multi-view learning. How to well perform multi-view leaning algorithms in the case of incomplete views or propose new multi-view learning algorithms with the ability to handle the incomplete-view case is an interesting research direction.

By the driving of practical applications where views are incomplete, some work on incomplete view learning was developed. Mostly it was designed to handle specific tasks such as multi-view clustering [137,138]. The main idea of multi-view clustering with incomplete view is to reconstruct the data in the missing views using space transformation and then perform full multi-view learning methods. Since Bayesian methods can deal with incomplete data by involving and integrating out latent variables, it is a feasible method to solve missing view issues in the future.

[8] Available at http://www.cst.ecnu.edu.cn/~slsun/software/MUDAcode.rar.

[9] Available at http://www.davidroihardoon.com/code.html.

[10] Available at http://www.cst.ecnu.edu.cn/~slsun/software/MVMEDcode.zip.

[11] Available at http://www.cst.ecnu.edu.cn/~slsun/software/SMVMEDcode.rar.

[12] Available at http://www.cst.ecnu.edu.cn/~slsun/software/MED-2C.rar.

5.5. Multi-view active learning based on Gaussian processes

In supervised learning, data with labels are very important for model training. Especially for multi-view learning, data from different views need to be labeled to construct training sets. However, more labeling data will cost more. Therefore, it is significant to reduce the number of labeled data without influencing the multi-view learning performance. Active learning is an effective method of selecting less and valuable data points to be labeled. It designs the classifier with data acquisition by ranking the unlabeled data to provide suggestions for the next query which has the highest training utility. Thus, it explores the maximum potential of the learner on both the labeled and unlabeled data, and the training set can be maintained as small as possible. This potentially leads to exploiting the most informative data, while significantly reducing the cost of data labeling. Combing multi-view learning with active learning will promote each other. On one hand, active learning provides a valid approach to select more valuable data from different views, which may improve the effectiveness and efficiency of multi-view learning algorithms. On the other hand, multi-view learning algorithms can help active learning to make better strategy of selecting data points.

There are already some work on multi-view active learning. For example, Muslea et al. [139] proposed a multi-view active learning method called co-testing, which firstly takes advantage of a few labeled examples to learn a hypothesis in each view, and then applies the learned hypothesis to all unlabeled examples and detects the set of contention points. Sun and Hardoon [140] presented an approach for multi-view active learning with extremely sparse labeled examples, which employs a similarity rule defined with CCA. These methods apply active learning to some certain multi-view learning algorithms and work well. They can inspire people to develop more effective multi-view active learning methods. Since Gaussian process active learning [141,142] has been proposed and experimentally proved valid, and multi-view Gaussian processes can be an elegant Bayesian learning method, it is worthwhile to study multi-view active learning algorithms based on Gaussian processes.

5.6. Multi-view sequential models under the Bayesian framework

When considering the type of data expression, sequential data are very common in the daily life. Sequential data also have multi-view information. For example, a sequence of human activities can be expressed as body sensor data or video data. A voice sequence can be expressed as audio data or throat sensor data. For these multi-view sequential data, existing multi-view learning methods will not work. Therefore, developing effective models with the ability to handle sequential data and utilizing multi-view information is an open problem.

Most existing methods for modeling sequential data are based on the Bayesian framework, such as hidden Markov models (HMMs) and Gaussian process dynamical systems (GPDSs) [143–145]. Among these, GPDSs are a kind of valid models with stronger modeling ability for sequential data. Thus, it is a significant research direction to propose multi-view learning methods for processing sequential data based on GPDSs.

6. Conclusions

We have made an overview of the developments of multi-view machine learning methods in terms of theories and methodologies. From perspectives of theories, we introduced the recent PAC-Bayesian bounds and Rademacher complexity based generalization error bounds. From perspectives of methodologies, we tried to provide a neat categorization and organization where the multi-view learning approaches are divided into three major categories. For each category, we described the representative algorithms and introduced the latest developments. In addition, some popular data sets were listed to provide convenience for future researchers. Several interesting and significant open problems were discussed in detail, which we think are worth studying. This paper can be useful for readers to further promote theoretical and methodological research and practical applications of multi-view learning.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Projects 61673179 and 61370175, Shanghai Knowledge Service Platform Project (No. ZF1213), and the Fundamental Research Funds for the Central Universities.

References

[1] S. Sun, A survey of multi-view machine learning, Neural Comput. Appl. 23 (2013) 2031–2038.
[2] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceeding of the 11th Annual Conference on Computational Learning Theory, 1998, pp. 92–100.
[3] K. Nigam, R. Ghani, Analyzing the effectiveness and applicability of co-training, in: Proceedings of the 9th International Conference on Information and Knowledge Management, 2000, pp. 86–93.
[4] I. Muslea, S. Minton, C. Knoblock, Active learning with multiple views, J. Artif. Intell. Res. 27 (2006) 203–233.
[5] S. Sun, F. Jin, Robust co-training, Int. J. Pattern Recognit. Artif. Intell. 25 (2011) 1113–1126.
[6] S. Sun, J. Shawe-Taylor, Sparse semi-supervised learning using conjugate functions, J. Mach. Learn. Res. 11 (2010) 2423–2455.
[7] X. Xie, S. Sun, Multi-view twin support vector machines, Intell. Data Anal. 19 (2015) 701–712.
[8] S. Sun, Multi-view Laplacian support vector machines, Lect. Notes Artif. Intell. 7121 (2011) 209–222.
[9] X. Xie, S. Sun, Multi-view Laplacian twin support vector machines, Appl. Intell. 41 (2014) 1059–1068.
[10] S. Sun, G. Chao, Multi-view maximum entropy discrimination, in: Proceedings of the 23rd International Joint Conference on Artificial Intelligence, 2013, pp. 1706–1712.
[11] G. Chao, S. Sun, Alternative multi-view maximum entropy discrimination, IEEE Trans. Neural Netw. Learn. Syst. 27 (2016) 1445–1456.
[12] L. Mao, S. Sun, Soft margin consistency based scalable multi-view maximum entropy discrimination, in: Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016, pp. 1839–1845.
[13] G. Chao, S. Sun, Consensus and complementarity based maximun entropy discrimination for multi-view classification, Inf. Sci. 367 (2016) 296–310.
[14] T. Jaakkola, M. Meila, T. Jebara, Maximum entropy discrimination, Adv. Neural Inf. Process. Syst. 12 (1999) 470–476.
[15] P. Yang, W. Gao, Information-theoretic multi-view domain adaptation: a theoretical and empirical study, J. Artif. Intell. Res. 49 (2014) 501–525.
[16] P. Yang, W. Gao, Multi-view discriminant transfer learning, in: Proceedings of the 23rd International Joint Conference on Artificial Intelligence, 2013, pp. 1848–1854.
[17] B. Tan, E.H. Zhong, E.W. Xiang, Q. Yang, Multi-transfer: transfer learning with multiple views and multiple sources, Stat. Anal. Data Min. 7 (2014) 282–293.
[18] Z. Ding, Y. Fu, Low-rank common subspace for multi-view learning, in: Proceedings of the 14th International Conference on Data Mining, 2014, pp. 110–119.
[19] Y. Guo, Convex subspace representation learning from multi-view data, in: Proceedings of the 27th AAAI Conference on Artificial Intelligence, 2013, pp. 387–393.
[20] M. Gönen, G.B. Gönen, F. Gürgen, Bayesian multiview dimensionality reduction for learning predictive subspaces, in: Proceedings of the 21st European Conference on Artificial Intelligence, 2014, pp. 387–392.
[21] C. Zhang, H. Fu, S. Liu, G. Liu, X. Cao, Low-rank tensor constrained multi-view subspace clustering, in: Proceedings of IEEE International Conference on Computer Vision, 2015, pp. 1582–1590.
[22] Y. Li, F. Nie, H. Huang, J. Huang, Large-scale multi-view spectral clustering via bipartite graph, in: Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015, pp. 2750–2756.
[23] S. Li, Y. Jiang, Z. Zhou, Partial multi-view clustering, in: Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014, pp. 1968–1974.
[24] X. Zhang, L. Zong, X. Liu, H. Yu., Constrained NMF-based multi-view clustering on unmapped data, in: Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015, pp. 3174–3180.
[25] X. Zhang, L. Zhao, L. Zong, X. Liu, H. Yu, Multi-view clustering via multi-manifold regularized nonnegative matrix factorization, in: Proceedings of the IEEE International Conference on Data Mining, 2014, pp. 1103–1108.

[26] R. Xia, Y. Pan, L. Du, J. Yin, Robust multi-view spectral clustering via low-rank and sparse decomposition, in: Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014, pp. 2149–2155.

[27] C. Zhang, J. Lai, P. Yu, Multi-view clustering based on belief propagation, IEEE Trans. Knowl. Data Eng. 28 (2015) 1007–1021.

[28] X. Cao, C. Zhang, H. Fu, S. Liu, H. Zhang, Diversity-induced multiview subspace clustering, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2015, pp. 586–594.

[29] Y. Makihara, A. Mansur, D. Muramatsu, Z. Uddin, Y. Yagi, Multi-view discriminant analysis with tensor representation and its application to cross-view gait recognition, in: Proceedings of 11th IEEE International Conference on Automatic Face and Gesture Recognition, 2015, pp. 1–8.

[30] S. Sun, X. Xie, M. Yang, Multiview uncorrelated discriminant analysis, IEEE Trans. Cybern. 46 (2015) 3272–3284.

[31] H. Yang, J. He, NOTAM^2: Nonparametric Bayes multi-task multi-view learning, in: Proceedings of World Statistics Conference, 2013, pp. 2351–2356.

[32] X. Mei, Z. Hong, D.V. Prokhorov, D. Tao, Robust multitask multiview tracking in videos, IEEE Trans. Neural Netw. Learn. Syst. 26 (2015) 2874–2890.

[33] Z. Hong, X. Mei, D.V. Prokhorov, D. Tao, Tracking via robust multi-task multi-view joint sparse representation, in: Proceedings of IEEE International Conference on Computer Vision, 2013, pp. 649–656.

[34] X. Zhang, X. Zhang, H. Liu, Multi-task multi-view clustering for non-negative data, in: Proceedings of the 24th International Joint Conference on Artificial Intelligence, 2015, pp. 4055–4061.

[35] X. Jin, F. Zhuang, S. Wang, Q. He, Z. Shi, Shared structure learning for multiple tasks with multiple views, in: Proceedings of ECML PKDD, 2013, pp. 353–368.

[36] S. Dasgupta, M. Littman, D. McAllester, PAC generalization bounds for co–training, Adv. Neural Inf. Process. Syst. 14 (2002) 375–382.

[37] K. Sridharan, S.M. Kakade, An information theoretic framework for multi-view learning, in: Proceedings of the Conference on Learning Theory, 2008, pp. 403–414.

[38] E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, S. Sun, PAC-Bayes bounds with data dependent priors, J. Mach. Learn. Res. 13 (2012) 3507–3531.

[39] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of the 11th Annual Conference on Computational Learning Theory, 1998, pp. 92–100.

[40] S. Abney, Bootstrapping, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 360–367.

[41] M.F. Balcan, A. Blum, K. Yang, Co-training and expansion: towards bridging theory and practice, Adv. Neural Inf. Process. Syst. 17 (2005) 89–96.

[42] S. Sun, J. Shawe-Taylor, L. Mao, PAC-Bayes analysis of multi-view learning, Inf. Fusion 35 (2017) 117–131.

[43] P. Bartlett, S. Mendelson, Rademacher and Gaussian complexities: risk bounds and structural results, J. Mach. Learn. Res. 3 (2002) 463–482.

[44] J. Farquhar, D. Hardoon, H. Meng, J. Shawe-Taylor, S. Szedmak, Two view learning: SVM-2K, theory and practice, Adv. Neural Inf. Process. Syst. 18 (2006) 355–362.

[45] S. Szedmak, J. Shawe-Taylor, Synthesis of maximum margin and multiview learning using unlabeled data, Neurocomputing 70 (2007) 1254–1264.

[46] D. Rosenberg, P. Bartlett, The Rademacher complexity of co-regularized kernel classes, J. Mach. Learn. Res. Workshop Conf. Proc. 2 (2007) 396–403.

[47] V. Sindhwani, D. Rosenberg, An RKHS for multi-view learning and manifold co-regularization, in: Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 976–983.

[48] C. Xu, D. Tao, C. Xu, Multi-view intact space learning, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 2531–2544.

[49] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, Neural Comput. 16 (2004) 2639–2664.

[50] T. Sun, S. Chen, J. Yang, P. Shi, A novel method of combined feature extraction for recognition, in: Proceedings of the International Conference on Data Mining, 2008, pp. 1043–1048.

[51] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 808–821.

[52] W. Wang, Z. Zhou, Analyzing co-training style algorithms, Lect. Notes Artif. Intell. 4701 (2007) 454–465.

[53] W. Wang, Z. Zhou, A new analysis of co-training, in: Proceedings of the 27th International Conference on Machine Learning, 2010, pp. 1135–1142.

[54] W. Wang, Z. Zhou, Co-training with insufficient views, in: Proceedings of the 5th Asian Conference on Machine Learning, 2013, pp. 467–482.

[55] U. Brefeld, T. Scheffer, Co-EM support vector learning, in: Proceedings of the 21st International Conference on Machine Learning, 2004, pp. 99–106.

[56] I. Muslea, S. Minton, C.A. Knoblock, Active learning with multiple views, J. Artif. Intell. Res. 27 (2006) 203–233.

[57] I. Muslea, S. Minton, C.A. Knoblock, Active+semi-supervised learning=robust multiview learning, in: Proceedings of the 19th International Conference on Machine Learning, 2002, pp. 435–442.

[58] S. Sun, F. Jin, Robust co-training, Int. J. Pattern Recognit. Artif. Intell. 25 (2011) 1113–1126.

[59] S. Yu, B. Krishnapuram, R. Rosales, R.B. Rao, Bayesian co-training, J. Mach. Learn. Res. 12 (2011) 2649–2680.

[60] X. Zhao, N. Evans, J.L. Dugelay, A subspace co-training framework for multi-view clustering, Pattern Recognit. Lett. 41 (2014) 73–82.

[61] M. Chen, K.Q. Weinberger, J. Blitzer, Co-training for domain adaptation, Adv. Neural Inf. Process. Syst. 24 (2011) 2456–2464.

[62] Z. Xu, S. Sun, Multi-view transfer learning with adaboost, in: Proceedings of the 23rd IEEE International Conference on Tools with Artificial Intelligence, 2011, pp. 399–402.

[63] Z. Xu, S. Sun, Multi-source transfer learning with multi-view adaboost, Lect. Notes Comput. Sci. 7665 (2012) 332–339.

[64] D. Zhang, J. He, Y. Liu, L. Si, R.D. Lawrence, Multi-view transfer learning with a large margin approach, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011, pp. 1208–1216.

[65] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 321–377.

[66] J.R. Kettenring, Canonical analysis of several sets of variables, Biometrika 58 (1971) 433–451.

[67] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.

[68] P. Lai, C. Fyfe, Kernel and nonlinear canonical correlation analysis, Int. J. Neural Syst. 10 (2000) 365–378.

[69] D. Hardoon, J. Shawe-Taylor, Convergence analysis of kernel canonical correlation analysis: theory and practice, Mach. Learn. 74 (2009) 23–38.

[70] F.R. Bach, M.I. Jordan, Kernel independent component analysis, J. Mach. Learn. Res. 3 (2002) 1–48.

[71] S. Akaho, A kernel method for canonical correlation analysis, in: Proceedings of the International Meeting of the Psychometric Society, 2001, pp. 1–7.

[72] J. Rupnik, J. Shawe-Taylor, Multi-view canonical correlation analysis, in: Proceedings of the Conference on Data Mining and Data Warehouses, 2010, pp. 1–4.

[73] P. Dhillon, D. Foster, L. Ungar, Multi-view learning of word embeddings via CCA, Adv. Neural Inf. Process. Syst. 24 (2011) 199–207.

[74] X. Chen, H. Liu, J. Carbonell, Structured sparse canonical correlation analysis, in: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics, 2012, pp. 199–207.

[75] D. Hardoon, J. Shawe-Taylor, Sparse canonical correlation analysis, Mach. Learn. 83 (2011) 331–353.

[76] L. Sun, S. Ji, J. Ye, Canonical correlation analysis for multilabel classification: a least-squares formulation, extension, and analysis, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 194–200.

[77] K. Chaudhuri, S.M. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 129–136.

[78] S.M. Kakade, D.P. Foster, Multi-view regression via canonical correlation analysis, in: Proceedings of the Annual Conference on Computational Learning Theory, 2007, pp. 82–96.

[79] R.F. Bach, I.M. Jordan, A probabilistic interpretation of canonical correlation analysis, Technical Report, Department of Statistics, University of California, Berkeley, 2005. pp. 1–11.

[80] A. Klami, S. Virtanen, S. Kaski, Bayesian canonical correlation analysis, J. Mach. Learn. Res. 14 (2013) 965–1003.

[81] S. Virtanen, A. Klami, S. Kaski, Bayesian CCA via group sparsity, in: Proceedings of the 28th International Conference on Machine Learning, 2011, pp. 457–464.

[82] A. Podosinnikova, F. Bach, S. Lacoste-Julien, Beyond CCA: moment matching for multi-view models, in: Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 458–467.

[83] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 1247–1255.

[84] J. Vía, I. Santamaría, J. Pérez, A learning algorithm for adaptive canonical correlation analysis of several data sets, Neural Netw. 20 (2007) 139–152.

[85] Y. Luo, D. Tao, Y. Wen, K. Ramamohanarao, C. Xu, Tensor canonical correlation analysis for multi-view dimension reduction, IEEE Trans. Knowl. Data Eng. 27 (2015) 3111–3124.

[86] T.K. Kim, S.F. Wong, R. Cipolla, Tensor canonical correlation analysis for action classification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[87] B. Long, P.S. Yu, Z.M. Zhang, A general model for multiple view unsupervised learning, in: Proceeding of the SIAM International Conference on Data Mining, 2008, pp. 822–833.

[88] N. Chen, J. Zhu, F. Sun, E.P. Xing, Large-margin predictive latent subspace learning for multiview data analysis, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 2365–2378.

[89] M. White, X. Zhang, D. Schuurmans, Y. Yu, Convex multi-view subspace learning, Adv. Neural Inf. Process. Syst. 25 (2012) 1–9.

[90] L. Zhang, S. Wang, X. Zhang, Y. Wang, B. Li, D. Shen, S. Ji, Collaborative multi-view denoising, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 2045–2054.

[91] S. Bickel, T. Scheffer, Multi-view clustering, in: Proceedings of the 4th IEEE International Conference on Data Mining, 2004, pp. 19–26.

[92] V.R.D. Sa, Spectral clustering with two views, in: Proceedings of the 22th IEEE International Conference on Machine Learning, 2005, pp. 20–27.

[93] M.B. Blaschko, C.H. Lampert, Correlational spectral clustering, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[94] D. Zhou, C.J.C. Burges, Spectral clustering and transductive learning with multiple views, in: Proceedings of the 12th IEEE International Conference on Data Mining, 2012, pp. 675–684.

[95] E. Bruno, S. Marchand-Maillet, Multiview clustering: a late fusion approach using latent models, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, pp. 736–737.
[96] X. Liu, S. Ji, W. Glänzel, Multiview partitioning via tensor methods, IEEE Trans. Knowl. Data Eng. 25 (2013) 1056–1069.
[97] H. Wang, F. Nie, H. Huang, Multi-view clustering and feature learning via structured sparsity, in: Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 352–360.
[98] J. Liu, C. Wang, J. Gao, J. Han, Multi-view clustering via joint nonnegative matrix factorization, in: Proceedings of the 13th SIAM International Conference on Data Mining, 2013, pp. 252–260.
[99] X. Xie, S. Sun, Multi-view clustering ensembles, in: Proceedings of International Conference on Machine Learning and Cybernetics, 2013, pp. 51–56.
[100] C.A. Méndez, P. Summers, G. Menegaz, Multiview cluster ensembles for multimodal MRI segmentation, Int. J. Imaging Syst. Technol. 25 (2015) 56–67.
[101] S.F. Hussain, M. Mushtaq, Z. Halim, Multi-view document clustering via ensemble method, J. Intell. Inf. Syst. 49 (2014) 81–99.
[102] N.F. Chikhi, Multi-view clustering via spectral partitioning and local refinement, Inf. Process. Manage. 52 (2016) 618–627.
[103] T. Diethe, D.R. Hardoon, J.S. Taylor, Constructing nonlinear discriminants from multiple data views, in: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2010, pp. 328–343.
[104] T. Kim, J. Kittler, R. Cipolla, Learning discriminative canonical correlations for object recognition with image sets, in: Proceedings of the 9th European Conference on Computer Vision, 2006, pp. 251–262.
[105] T. Diethe, D.R. Hardoon, J. Shawe-Taylor, Multiview Fisher discriminant analysis, in: Proceedings of the NIPS Workshop on Learning from Multiple Sources, 2008, pp. 1–8.
[106] Q. Chen, S. Sun, Hierarchical multi-view Fisher discriminant analysis, Lect. Notes Comput. Sci. 5864 (2009) 289–298.
[107] A. Sharma, A. Kumar, H. Daume, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: Proceedings of the Computer Vision and Pattern Recognition, 2012, pp. 2160–2167.
[108] C. Xu, D. Tao, Y. Li, C. Xu, Large-margin multi-view Gaussian process, Multimedia Syst. 21 (2014) 147–157.
[109] S. Eleftheriadis, O. Rudovic, M. Pantic, Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition, IEEE Trans. Image Process. 24 (2015) 189–204.
[110] U. Brefeld, C. Büscher, T. Scheffer, Multi-view discriminative sequential learning, Lect. Notes Artif. Intell. 3720 (2005) 60–71.
[111] D. Rosenberg, V. Sindhwani, P. Bartlett, P. Niyogi, Multi-view point cloud kernels for semisupervised learning, IEEE Signal Process. Mag. 145 (2009) 145–150.
[112] G. Li, S.C.H. Hoi, K. Chang, Multiview semi-supervised learning with consensus, IEEE Trans. Knowl. Data Eng. 24 (2012) 2040–2051.
[113] Y. Luo, D. Tao, C. Xu, C. Xu, H. Liu, Y. Wen, Multiview vector-valued manifold regularization for multilabel image classification, IEEE Trans. Neural Netw. Learn. Syst. 24 (2013) 209–722.
[114] J. He, R. Lawrence, A graph-based framework for multi-task multi-view learning, in: Proceedings of the 28th International Conference on Machine Learning, 2011, pp. 25–32.
[115] J. Zhang, J. Huan, Inductive multi-task learning with multiple view data, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 543–551.
[116] X. Jin, F. Zhuang, H. Xiong, C. Du, P. Luo, Q. He., Multi-task multi-view learning for heterogeneous tasks, in: Proceedings of the ACM International Conference on Information and Knowledge Management, 2014, pp. 441–450.
[117] Z. Xu, S. Sun, An algorithm on multi-view adaboost, Lect. Notes Comput. Sci. 6443 (2010) 355–362.
[118] S. Sun, Q. Zhang, Multiple-view multiple-learner semi-supervised learning, Neural Process. Lett. 34 (2011) 229–240.
[119] Q. Zhang, S. Sun, Multiple-view multiple-learner active learning, Pattern Recognit. 43 (2010) 3113–3119.
[120] V. Kumar, S. Minz, Multi-view ensemble learning: a supervised feature set partitioning for high dimensional data classification, in: Proceedings of the 3rd International Symposium on Women in Computing & Informatics, 2015, pp. 31–37.
[121] V. Kumar, S. Minz, Multi-view ensemble learning: an optimal feature set partitioning for high dimensional data classification, Knowl. Inf. Syst. (2016) 1–59.
[122] V. Kumar, S. Minz, Multi-view ensemble learning for poem data classification using sentiwordnet, in: Proceedings of the 2nd International Conference on Advanced Computing, Networking and Informatics, 2014, pp. 57–66.
[123] R. Gross, I. Matthews, J. Cohn, T. Kanada, S. Baker, The CMU multi-pose, illumination, and expression (multi-pie) face database, Technical Report, Carnegie Mellon University Robotics Institute, 2007. pp. 1–8.
[124] G. Tzortzis, A. Likas, Kernel-based weighted multiview clustering, in: Proceedings of the 12th IEEE International Conference on Data Mining, 2012, pp. 675–684.
[125] E. Tola, C. Strecha, P. Fua, Efficient large-scale multiview stereo for ultra high--resolution image sets, Mach. Vision Appl. 23 (2012) 903–920.
[126] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, H. Aanæs, Large scale multi-view stereopsis evaluation, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 406–413.
[127] Y. Zhu, W. Gao, Z.H. Zhou, One-pass multi-view learning, in: Proceedings of the 7th Asian Conference on Machine Learning, 2015, pp. 407–422.
[128] Y. Lu, D.P. Foster, Large scale canonical correlation analysis with iterative least squares, Adv. Neural Inf. Process. Syst. 27 (2014) 91–99.
[129] X. Cai, F. Nie, H. Huang, Multi-view k-means clustering on big data, in: Proceedings of the 23rd International Joint Conference on Artificial Intelligence, 2013, pp. 2598–2604.
[130] W. Wang, R. Arora, K. Livescu, J. Bilmes, On deep multi-view representation learning, in: Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 1083–1092.
[131] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, Multimodal deep learning, in: Proceedings of the 28th International Conference on Machine Learning, 2011, pp. 689–696.
[132] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 1247–1255.
[133] Z. Zhu, P. Luo, X. Wang, X. Tang, Multi-view perceptron: a deep model for learning face identity and view representations, Adv. Neural Inf. Process. Syst. 27 (2014) 217–225.
[134] H. Su, S. Maji, E. Kalogerakis, E. Learned-Miller, Multi-view convolutional neural networks for 3D shape recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 945–953.
[135] M. Elhoseiny, T. El-Gaaly, A. Bakry, A. Elgammal, A comparative analysis and study of multiview CNN models for joint object categorization and pose estimation, in: Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 888–897.
[136] A.M. Elkahky, Y. Song, A.X. He, Multi-view deep learning approach for cross domain user modeling in recommendation systems, in: Proceedings of the 24th International Conference on World Wide Web, 2015, pp. 278–288.
[137] S. Li, Y. Jiang, Z. Zhou, Partial multi-view clustering, in: Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014, pp. 1968–1974.
[138] W. Shao, L. He, P. Yu, Multiple incomplete views clustering via weighted nonnegative matrix factorization with $L_{2,1}$ regularization, in: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2015, pp. 318–334.
[139] I. Muslea, S. Minton, C. Knoblock, Selective sampling with redundant views, in: Proceedings of the 17th AAAI Conference on Artificial Intelligence, 2000, pp. 621–626.
[140] S. Sun, D. Hardoon, Active learning with extremely sparse labeled examples, Neurocomputing 73 (2010) 2980–2988.
[141] J. Zhou, S. Sun, Active learning of Gaussian processes with manifold-preserving graph reduction, Neural Comput. Appl. 25 (2014) 1615–1625.
[142] J. Zhou, S. Sun, Gaussian process versus margin sampling active learning, Neurocomputing 167 (2015) 122–131.
[143] J. Zhao, S. Sun, Revisiting Gaussian process dynamical models, in: Proceedings of the 24th International Joint Conference on Artificial Intelligence, 2015, pp. 1047–1053.
[144] J. Zhao, S. Sun, High-order Gaussian process dynamical models for traffic flow prediction, IEEE Trans. Intell. Transp. Syst. 17 (2016a) 2014–2019.
[145] J. Zhao, S. Sun, Variational dependent multi-output Gaussian process dynamical systems, J. Mach. Learn. Res. 17 (2016b) 1–36.