Neural Comput & Applic (2013) 23:2031–2038
DOI 10.1007/s00521-013-1362-6

INVITED REVIEW

A survey of multi-view machine learning

Shiliang Sun

Received: 4 February 2013 / Accepted: 6 February 2013 / Published online: 17 February 2013

Abstract Multi-view learning or learning with multiple distinct feature sets is a rapidly growing direction in machine learning with well theoretical underpinnings and great practical success. This paper reviews theories developed to understand the properties and behaviors of multi-view learning and gives a taxonomy of approaches according to the machine learning mechanisms involved and the fashions in which multiple views are exploited. This survey aims to provide an insightful organization of current developments in the field of multi-view learning, identify their limitations, and give suggestions for further research. One feature of this survey is that we attempt to point out specific open problems which can hopefully be useful to promote the research of multi-view machine learning.

Keywords Multi-view learning · Statistical learning theory · Canonical correlation analysis · Co-training · Co-regularization

1 Introduction

Multi-view learning is concerned with the problem of machine learning from data represented by multiple distinct feature sets. The recent emergence of this learning mechanism is largely motivated by the property of data from real applications where examples are described by different feature sets or different "views". For instance, in multimedia-content understanding, multimedia segments can be simultaneously described by their video and audio signals. In web page classification, a web page can be described by the document text itself and at the same time by the anchor text attached to hyperlinks pointing to this page. As another example, in content-based web-image retrieval, an object is simultaneously described by visual features from the image and the text surrounding the image. Moreover, a noteworthy fact for multi-view learning is that when a natural feature split does not exist, performance improvements can still be observed using manufactured splits. Therefore, multi-view learning is a very promising topic with widespread applicability.

Canonical correlation analysis (CCA) [21] and co-training [8] are two representative techniques in early studies of multi-view learning. Some theories and methods were later devised to investigate their theoretical properties, explain their success, and extend their applications to other machine learning problems. In 2005, a workshop on learning with multiple views was held in conjunction with the 22nd international conference on machine learning to attract attentions and promote research in this area. So far, the idea of multi-view learning has penetrated multiple existing machine learning branches and a large number of multi-view learning algorithms have been presented. For example, the applications of multi-view learning range from dimensionality reduction [10, 20, 50] and semi-supervised learning [35, 36, 38, 39, 42, 54, 56] to supervised learning [11, 16], active learning [28, 41], ensemble learning [45, 51, 55], transfer learning [12, 52, 53], and clustering [7, 15, 23, 24].

The goal of this survey is to review key advancements in the area of multi-view learning, in particular, on theories

S. Sun (✉)
Department of Computer Science and Technology, East China Normal University, 500 Dongchuan Road, Shanghai 200241, China
e-mail: slsun@cs.ecnu.edu.cn; shiliangsun@gmail.com

and methodologies, and provide useful suggestions for further research. Through this survey, we would like to deliver a whole picture of what is going on and what can be done in the future to make multi-view learning more successful.

The remainder of this paper proceeds as follows. In Sect. 2, we introduce existing theories on multi-view learning, especially on CCA, effectiveness of co-training, and generalization error analysis for co-training and other multi-view learning approaches. Section 3 surveys representative multi-view approaches according to the machine learning mechanisms involved and also provides another taxonomy in terms of the specific manners in which multiple views are exploited. Then, in Sect. 4, we list some open problems which may be helpful for promoting further research of multi-view learning. Finally, we provide concluding remarks in Sect. 5.

2 Theories on multi-view learning

We classify current theories on multi-view learning into four categories which are CCA, effectiveness of co-training, generalization error analysis for co-training, and generalization error analysis for other multi-view learning approaches. These theories can partially answer at least the following three questions: why multi-view learning is useful, what are the underlying assumptions, and how we should perform multi-view learning.

2.1 CCA

Canonical correlation analysis, first proposed by Hotelling [21], works on a paired data set (e.g., data represented by two views) to find two linear transformations each for one view such that the correlations between the transformed variables are maximized. It was later generalized to data with more than two representations in several ways [3, 22]. Here, we only consider the case of two views.

Suppose we have a two-view data set $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_m, \mathbf{y}_m)\}$, and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_m]$. CCA attempts to seek two projection directions $\mathbf{w}_x$ and $\mathbf{w}_y$ to maximize the following linear correlation coefficient

$$\frac{\mathrm{cov}(\mathbf{w}_x^\top \mathbf{X}, \mathbf{w}_y^\top \mathbf{Y})}{\sqrt{\mathrm{var}(\mathbf{w}_x^\top \mathbf{X})\mathrm{var}(\mathbf{w}_y^\top \mathbf{Y})}} = \frac{\mathbf{w}_x^\top \mathbf{C}_{xy} \mathbf{w}_y}{\sqrt{(\mathbf{w}_x^\top \mathbf{C}_{xx} \mathbf{w}_x)(\mathbf{w}_y^\top \mathbf{C}_{yy} \mathbf{w}_y)}}, \quad (1)$$

where covariance matrix $\mathbf{C}_{xy}$ is defined as

$$\mathbf{C}_{xy} = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i - \mathbf{m}_x)(\mathbf{y}_i - \mathbf{m}_y)^\top \quad (2)$$

with $\mathbf{m}_x$ and $\mathbf{m}_y$ being the means from the two views, respectively,

$$\mathbf{m}_x = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i, \quad \mathbf{m}_y = \frac{1}{m} \sum_{i=1}^{m} \mathbf{y}_i, \quad (3)$$

and $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ can be defined analogously.

Since the scales of $\mathbf{w}_x$ and $\mathbf{w}_y$ have no effects on the value of (1), each of the two factors in the denominator can be constrained to have value 1. This results in another widely used objective for CCA

$$\begin{aligned} \max_{\mathbf{w}_x, \mathbf{w}_y} \quad & \mathbf{w}_x^\top \mathbf{C}_{xy} \mathbf{w}_y \\ \text{s.t.} \quad & \mathbf{w}_x^\top \mathbf{C}_{xx} \mathbf{w}_x = 1, \ \mathbf{w}_y^\top \mathbf{C}_{yy} \mathbf{w}_y = 1. \end{aligned} \quad (4)$$

The corresponding Lagrangian function is

$$\begin{aligned} L(\mathbf{w}_x, \mathbf{w}_y, \lambda_x, \lambda_y) = \mathbf{w}_x^\top \mathbf{C}_{xy} \mathbf{w}_y - \frac{\lambda_x}{2} (\mathbf{w}_x^\top \mathbf{C}_{xx} \mathbf{w}_x - 1) \\ - \frac{\lambda_y}{2} (\mathbf{w}_y^\top \mathbf{C}_{yy} \mathbf{w}_y - 1). \end{aligned} \quad (5)$$

Taking its derivatives with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$ to be zero, we have

$$\mathbf{C}_{xy} \mathbf{w}_y - \lambda_x \mathbf{C}_{xx} \mathbf{w}_x = 0 \quad (6)$$

$$\mathbf{C}_{yx} \mathbf{w}_x - \lambda_y \mathbf{C}_{yy} \mathbf{w}_y = 0. \quad (7)$$

Subtracting $\mathbf{w}_y^\top \times (7)$ from $\mathbf{w}_x^\top \times (6)$, we get

$$\lambda_y \mathbf{w}_y^\top \mathbf{C}_{yy} \mathbf{w}_y - \lambda_x \mathbf{w}_x^\top \mathbf{C}_{xx} \mathbf{w}_x = \lambda_y - \lambda_x = 0. \quad (8)$$

Therefore, $\lambda_x = \lambda_y$. Suppose $\lambda_x = \lambda_y = \lambda$. Given that $\mathbf{C}_{yy}$ is invertible, $\mathbf{w}_y$ can be obtained from (7) as

$$\mathbf{w}_y = \frac{1}{\lambda} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \mathbf{w}_x. \quad (9)$$

Substituting (9) into (6) results in the following generalized eigenvalue decomposition problem [39]

$$\mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \mathbf{w}_x = \lambda^2 \mathbf{C}_{xx} \mathbf{w}_x. \quad (10)$$

Now $\mathbf{w}_x$ can be solved, which should then be normalized according to (4). The corresponding $\mathbf{w}_y$ is obtained from (9) which should also be normalized according to (4).

To make the relationship between the eigenvalue λ^2 in (10) and the correlation coefficient clear, we rewrite the objective function as

$$\begin{aligned} \mathbf{w}_x^\top \mathbf{C}_{xy} \mathbf{w}_y &= \frac{1}{\lambda} \mathbf{w}_x^\top \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \mathbf{w}_x \\ &= \frac{1}{\lambda} \mathbf{w}_x^\top \lambda^2 \mathbf{C}_{xx} \mathbf{w}_x = \lambda \mathbf{w}_x^\top \mathbf{C}_{xx} \mathbf{w}_x = \lambda. \end{aligned} \quad (11)$$

Thus, λ reflects the degree of correlation between projections, which must lie in the interval $[-1, +1]$. Interestingly, if $\begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix}$, λ is a solution pair, then $\begin{pmatrix} \mathbf{w}_x \\ -\mathbf{w}_y \end{pmatrix}$, $-\lambda$ would give an equal but negative correlation. However, these two kinds of solutions are equivalent in the sense

that we are only seeking projection directions. Therefore, we just need to consider the positive correlation, as reflected by the objective function in (4). To maximize the correlation between different views, the eigenvector corresponding to the largest eigenvalue in (10) should be retained. For real applications, there are often a lot of projection vector pairs ($\mathbf{w}_x$, $\mathbf{w}_y$) required to reflect different correlations. If CCA retains q pairs of correlated projections, an example (**x, y**) will be transformed to q projection pairs.

It was shown that overfitting with perfect correlations but failing to distinguish spurious from useful features can appear using CCA [3, 33]. Therefore, regularization is needed to detect meaningful patterns. The objective function of the regularized CCA is to maximize

$$\frac{\mathbf{w}_x^\top \mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\left((1-\tau_x)\mathbf{w}_x^\top \mathbf{C}_{xx}\mathbf{w}_x + \tau_x\|\mathbf{w}_x\|^2\right)\left((1-\tau_y)\mathbf{w}_y^\top \mathbf{C}_{yy}\mathbf{w}_y + \tau_y\|\mathbf{w}_y\|^2\right)}}, \tag{12}$$

where regularization parameters τ_x and τ_y vary in the interval [0, 1]. Recent statistical analysis, based on a close relationship between maximizing the correlation and minimizing the discrepancy of the two views in terms of the squared loss, has justified that controlling the norms of the projection directions is a principled way for regularization [19].

Canonical correlation analysis was extended to kernel CCA [3, 17] by means of the kernel trick [34], which corresponds to performing CCA in a kernel-induced feature space. The formulation of the regularized kernel CCA can be found in [19, 34]. Lately, sparse CCA was also presented [10, 20].

2.2 Effectiveness of co-training

The original co-training algorithm was introduced by Blum and Mitchell [8] for semi-supervised classification that combines both labeled and unlabeled data under a two-view setting. From a limited labeled data set, it first trains two weakly-useful classifiers from the two views separately. Then, the two classifiers find their confident predictions from a pool of unlabeled data to enlarge the labeled data set for further training. The process repeats until a termination condition is satisfied. Finally, the two classifiers are used separately or jointly to make predictions on a new example. Later on, the applicability of co-training was further broadened, e.g., Nigam and Ghani [29] showed experimentally that when there are no natural multiple views available, co-training on multiple views manually generated by random splits of features can still improve performance.

The probably approximately correct (PAC) learning framework can provide a theoretical characterization of the capabilities of machine learning algorithms and the difficulty of some machine learning problems. Loosely speaking, a concept class C is PAC-learnable by a learner L using a hypothesis space H if, for any target concept in C, L will with probability at least $(1-\delta)$ output a hypothesis whose error is less than or equal to ϵ, after training with a reasonable number of examples and performing a reasonable amount of computation [27].

To justify the effectiveness of co-training, Blum and Mitchell [8] gave a PAC-style analysis. They showed that under assumptions that (1) each view in itself is sufficient for correct classification (i.e., target functions from the two views and the combined view have label consistency on every example) and (2) the two views of any example are conditionally independent given the class label, PAC learnability on semi-supervised learning holds with an initial weakly useful predictor trained from the labeled data. For a special case of co-training, Balcan and Blum [4] proved that there is a polynomial-time algorithm to learn a linear separator under proper assumptions, using a single-labeled example and polynomially many unlabeled examples.

It was shown that the second assumption of co-training can be relaxed to a weaker expansion assumption on the underlying data distribution for iterative co-training to succeed, given appropriately strong PAC-learning algorithms on each view, and the expansion assumption is to some extent necessary as well [5].

Wang and Zhou [48] proved that the co-training process can succeed even without two views, given that the labeled data set is sufficient to learn good classifiers and the two classifiers have a large diversity. Under the setting that the learner in each view is viewed as label propagation and thus the co-training process is viewed as the combinative label propagation over the two views, they further provided a sufficient and necessary condition for co-training to succeed with appropriate assumptions [49].

In practice, the original co-training algorithm may be problematic in the sense that it does not examine the reliability of labels provided by the classifiers from each view. Actually, even very few inaccurately labeled examples can greatly deteriorate the performance of subsequent classifiers. To overcome this drawback, Sun and Jin [39] proposed robust co-training, which integrates CCA to inspect the predictions of co-training on the unlabeled training data. Based on the low-dimensional representations recovered by CCA, it calculates the similarities between an unlabeled example and the original labeled examples. Only those examples whose predicted labels are consistent with the outcome of CCA label inspection are eligible to enlarge the labeled set.

2.3 Generalization error analysis for co-training

Early theoretical work on co-training such as [8] was only loosely related to its empirical success. In particular, it does not provide a generalization error bound as a function of empirically measurable quantities, and there is no very direct and apparent relationship between the PAC-learnability analysis and the iterative co-training algorithm, as stated in [14].

Based on the conditional independence assumption of views, Dasgupta et al. [14] gave a PAC generalization bound for co-training, which 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. This justifies the kind of empirical work that encourages agreements between classifiers from different views over the unlabeled data [13].

The assumption that views are conditionally independent is rather strong and hardly holds in practice. Abney [1] generalized the error bound in [14] with weaker assumptions that are classifiers from different views are weakly dependent and nontrivial.

2.4 Generalization error analysis for other multi-view learning approaches

In order to gain insights into the roles played by the multi-view regularization and even unlabeled data in the generalization performance, researchers have provided generalization error analysis for some other multi-view learning approaches. This kind of generalization analysis is built upon the Rademacher complexity theory which we briefly introduce below through a definition and theorem.

Definition 1 (**Rademacher complexity** [6, 33]) For a sample $S = \{x_1, \ldots, x_\ell\}$ generated by a distribution $\mathcal{D}_x$ on a set X and a real-valued function class $\mathcal{F}$ with domain X, the empirical Rademacher complexity of $\mathcal{F}$ is the random variable

$$\hat{R}_\ell(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i f(x_i)\right| | x_1, \ldots, x_\ell\right], \tag{13}$$

where $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_\ell\}$ are independent uniform $\{\pm 1\}$-valued (Rademacher) random variables. The Rademacher complexity of $\mathcal{F}$ is

$$R_\ell(\mathcal{F}) = \mathbb{E}_S[\hat{R}_\ell(\mathcal{F})] = \mathbb{E}_{S\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i f(x_i)\right|\right]. \tag{14}$$

Theorem 1 ([33]) *Fix $\delta \in (0,1)$ and let $\mathcal{F}$ be a class of functions mapping from an input space Z (for supervised learning having the form $Z = X \times Y$) to [0, 1]. Let $\{z_i\}_{i=1}^{\ell}$ be drawn independently according to a probability distribution $\mathcal{D}$. Then with probability at least $1-\delta$ over random draws of samples of size ℓ, every $f \in \mathcal{F}$ satisfies*

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[f(z)] &\le \hat{\mathbb{E}}[f(z)] + R_\ell(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2\ell}} \\ &\le \hat{\mathbb{E}}[f(z)] + \hat{R}_\ell(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}},\end{aligned} \tag{15}$$

where $\hat{\mathbb{E}}[f(z)]$ is the empirical error averaged on the ℓ examples.

Making use of the Rademacher complexity theory, Farquhar et al. [16] analyzed the generalization error bound of the supervised SVM-2K algorithm, and Szedmak and Shawe-Taylor [46] characterized the generalization performance of its extended version for semi-supervised learning.

Rosenberg and Bartlett [31] derived the empirical Rademacher complexity for the function class of co-regularized least squares and gave the generalization bound which was later recovered by Sindhwani and Rosenberg [36] but with a much simpler derivation. Potentially tighter bounds were also reported in terms of the localized Rademacher complexity [36]. This kind of work was further extended to a more general setting, e.g., with more than two views [32].

Recently, Sun and Shawe-Taylor [42] 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 where the empirical Rademacher complexity has two different forms depending on whether the used iterative procedure iterates only once or multiple steps. Taking manifold regularization into account, Sun [38] presented multi-view Laplacian SVMs whose generalization error analysis and empirical Rademacher complexity were also provided.

3 Multi-view learning methods

We proceed to review representative multi-view learning methods according to the machine learning mechanisms that multi-view learning is applied to or combined with. Then, we give a high-level taxonomy of multi-view learning methods in terms of how multiple views are exploited.

3.1 Multi-view dimensionality reduction

As an important branch of unsupervised learning, dimensionality reduction aims to express high-dimensional data with low-dimensional representations to reveal significant latent information. It can be used to compress, visualize or

re-organize data, and as a preprocessing step for other machine learning tasks.

Canonical correlation analysis is an early and classical method for multi-view dimensionality reduction by learning subspaces jointly from different views [21]. It was further extended to nonlinear subspace learning [3, 17] and sparse formulations [2, 10, 20]. Recently, White et al. [50] adapted new advances of single-view subspace learning to the multi-view case and provided a convex formulation for multi-view subspace learning. This work permits an arbitrary loss function that is convex in the first argument and replaces the usual rank constraint with a rank-reducing regularizer.

3.2 Multi-view semi-supervised learning

Semi-supervised learning or learning from both labeled and unlabeled data has attracted much attention during the last decade. For many practical applications, label information is expensive or time-consuming to obtain but unlabeled examples are very easy to collect. In this scenario, it is helpful to combine the limited labeled data together with the unlabeled data for effective function learning. Semi-supervised learning can address this problem by learning with few labeled data and a large number of unlabeled data jointly, where the unlabeled data can play the role of induction preference toward functions with some properties.

Multi-view semi-supervised learning has an additional approach for induction preference, namely view agreements. By requiring that functions from different views have similar outputs, it can reduce the size of the hypothesis space and thus a better generalization performance is possible. Representative multi-view semi-supervised learning methods include co-training [8], co-EM [29], multi-view sequential learning [9], Bayesian co-training [54], multi-view point cloud regularization [32], sparse multi-view SVMs [42], and robust co-training [39]. The recent multi-view Laplacian SVMs [38] integrate the multi-view regularization with manifold regularization and bring further improvements.

3.3 Multi-view supervised learning

Unlike semi-supervised learning, supervised learning only uses labeled data for function learning. However, research on multi-view supervised learning is comparatively less than multi-view semi-supervised learning. One reason may be that multi-view semi-supervised learning can often be regarded as a more difficult and general problem than multi-view supervised learning. Multi-view supervised learning is almost direct to adapt if one already has a multi-view semi-supervised learning method. But we should note that these two problems are intrinsically distinct. For example, effective model selection is more difficult for semi-supervised learning than for supervised learning.

For multi-view supervised learning, Chen and Sun [11] proposed the multi-view Fisher discriminant analysis which is applicable for both binary and multi-class classification. Farquhar et al. [16] introduced supervised SVM-2K that was later extended to multi-view semi-supervised learning [46].

3.4 Multi-view active learning

Active learning is concerned with the scenario, where a learning algorithm can actively query the user for labels. Due to this interactive nature, the number of examples needed to learn a function can often be much lower than the corresponding supervised learning case. In other words, the aim of active learning is to alleviate the burden of labeling abundant examples by discovering and asking the user to label only the most informative ones.

Muslea et al. [28] gave a multi-view active learning method co-testing which is a two-step iterative process. First, it uses a few labeled examples to learn a classifier in each view. Then, it queries an unlabeled example (a contention point) for which the views predict different labels. After adding the queried example to the labeled training set, the entire procedure is repeated for a number of iterations. Yu et al. [54] introduced an active sensing framework with Bayesian co-training, in which the $\langle \text{example}, \text{view} \rangle$ pairs are actively queried to improve learning performance.

However, for some applications, there are very limited labeled examples available. For instance, in the extreme case, each category can have a single-labeled example where most existing active learning methods cannot be directly applied. Sun and Hardoon [41] proposed an approach for multi-view active learning with extremely sparse labeled examples, which adopts a similarity rule defined with CCA [56].

3.5 Multi-view ensemble learning

The goal of ensemble learning is to use multiple models (e.g., classifiers or regressors) to obtain a better predictive performance than could be obtained from any of the constituent models. It is widely acknowledged that an effective ensemble learning system should consist of individuals that are not only accurate, but are diverse as well, that is, a good balance should hold between diversity and individual performance [37, 43, 44].

Xu and Sun [51] extended the well-known ensemble learning method Adaboost to the multi-view learning scenario and proposed the embedded multi-view Adaboost algorithm (EMV-Adaboost). The key idea of EMV-

Adaboost is that during every iteration an example will contribute to the error rate as long as it is predicted incorrectly by either of the weaker learners from the two views. Sun and Zhang introduced a multi-view ensemble learning framework possessing both multiple views and multiple learners and applied it successfully to semi-supervised learning [45] and active learning [55], respectively.

3.6 Multi-view transfer learning

Transfer learning is one emerging and active topic in current machine learning research. Traditional machine learning algorithms are usually designed for solving a certain single task. The recent developments of transfer learning or multitask learning have shown that it is often advantageous to transfer knowledge learned in one or more source tasks to a related target task to improve learning.

Chen et al. [12] introduced a variant of co-training for domain adaptation which attempts to bridge the gap between source and target domains whose distributions can differ substantially. This variant gradually adds to the training set both the target features and instances that are regarded as the most confident. Specifically, for each iteration of co-training, it simultaneously learns a target predictor, a split of the feature space into views, and a subset of source and target features to include in the predictor. Xu and Sun proposed an algorithm involving a variant of EMV-Adaboost for multi-view transfer learning [52] and further extended it to taking the advantages of learning with multiple sources [53].

3.7 Multi-view clustering

Multi-view learning has also been applied to improve single-view clustering methods. Bickel and Scheffer [7] studied multi-view versions of several clustering algorithms for text data and found that EM-based multi-view algorithms significantly outperform the single-view counterparts, while the agglomerative hierarchical multi-view clustering leads to negative results.

Recently, Tzortzis and Likas [47] proposed a multi-view convex mixture model that extends convex mixture models to the multi-view clustering setting. de Sa et al. [15] developed an algorithm to leverage information from multiple views for clustering by constructing a multi-view affinity matrix. They used this multi-view affinity matrix as the affinity matrix for spectral clustering. Kumar and Daumé [23] presented a co-training approach for multi-view spectral clustering, where the clusterings of different views are bootstrapped using information from one another. In particular, the spectral embedding from one view is adopted to constrain the similarity graph used for the other view. Kumar et al. [24] further proposed two co-regularization based approaches for multi-view spectral clustering by enforcing the clustering hypotheses on different views to agree with each other. They constructed an objective function that consists of the graph Laplacians from all views and made regularizations on the eigenvectors of the Laplacians such that the resulting cluster structures would be consistent.

3.8 A high-level taxonomy

Current multi-view learning methods can be divided into two major categories: co-training style algorithms and co-regularization style algorithms. They are two different approaches for exploiting multiple views.

The co-training style algorithms are inspired by the co-training algorithm [8], which essentially involve an iterative procedure to exploit different views. For example, co-EM [29], co-testing [28], and robust co-training [39] are of this category.

For the co-regularization style algorithms such as sparse multi-view SVMs [42] and multi-view Laplacian SVMs [38], the disagreement between the functions of two views is taken as one part of the objective function to be minimized. Note that, CCA [21] and Bayesian co-training [54] also belong to the co-regularization style category.

4 Open problems

Now we present several important open problems which can be very useful for further developments of multi-view learning.

4.1 PAC-Bayes analysis of multi-view learners

For generalization error analysis of multi-view learners, we have witnessed some results based on the Rademacher complexity bounds. However, the tightest bounds so far for practical applications appear to be the PAC-Bayes bound [25, 26] for which the most recent research outcome is using data-dependent priors [30]. It would be interesting to show whether tighter and more insightful bounds can be obtained for multi-view learners with the theory of PAC-Bayes analysis.

4.2 New approaches to exploiting distinct views

From the survey of existing multi-view methods, especially Sect. 3.8, we know that the two major categories of approaches to exploiting distinct views are co-training style algorithms and co-regularization style algorithms. Different from these approaches, Ganchev et al. [18]

introduced stochastic agreement regularization for multi-view learning over structured outputs, which uses the Bhattacharyya distance between distributions. Therefore, a natural question to ask is: can we go further beyond these approaches?

4.3 Theory and practical methods for view construction

It is shown that multi-view learning often works even with multiple views generated from data with one single view. Typical view construction methods include the random split [29] and principal component analysis [45]. Recently, Sun et al. [40] proposed to use genetic algorithms for view construction. However, the practical problem of effective view construction is still not as highly valued as it should be.

Meanwhile, it remains a problem when we should generate multiple views from a whole single view and apply multi-view learning methods rather than single-view learning methods. Research on this topic is very few. Especially, theoretical insights are in urgent need.

5 Conclusion

We have surveyed recent developments on theories and methodologies of multi-view machine learning where when applicable we tried to provide a neat categorization and organization. Several open problems were also listed, which we think are important for the development of multi-view learning. This paper can be useful for readers to further promote the research of multi-view learning or apply the idea of multi-view learning to other machine learning problems.

Acknowledgments This work is supported by the National Natural Science Foundation of China under Project 61075005, the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and Shanghai Knowledge Service Platform for Trustworthy Internet of Things (No. ZF1213).

References

1. Abney S (2002) Bootstrapping. Proceedings of the 40th annual meeting of the association for computational linguistics, pp 360–367
2. Archambeau C, Bach F (2009) Sparse probabilistic projections. Adv Neural Inform Process Syst 21:17–24
3. Bach F, Jordan M (2002) Kernel independent component analysis. J Mach Learn Res 3:1–48
4. Balcan MF, Blum A (2005) A PAC-style model for learning from labeled and unlabeled data. Proceedings of the 18th annual conference on computational learning theory, pp 111–126
5. Balcan MF, Blum A, Yang K (2005) Co-training and expansion: towards bridging theory and practice. Adv Neural Inform Process Syst 17:89–96
6. Bartlett P, Mendelson S (2002) Rademacher and Gaussian complexities: risk bounds and structural results. J Mach Learn Res 3:463–482
7. Bickel S, Scheffer T (2004) Multi-view clustering. Proceedings of the 4th IEEE international conference on data mining, pp 19–26
8. Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. Proceedings of the 11th annual conference on computational learning theory, pp 92–100
9. Brefeld U, Büscher C, Scheffer T (2005) Multi-view discriminative sequential learning. Lect Notes Aritif Intell 3720:60–71
10. Chen X, Liu H, Carbonell J (2012) Structured sparse canonical correlation analysis. Proceedings of the 15th international conference on artificial intelligence and statistics, pp 199–207
11. Chen Q, Sun S (2009) Hierarchical multi-view Fisher discriminant analysis. Lect Notes Comput Sci 5864:289–298
12. Chen M, Weinberger K, Blitzer J (2011) Co-training for domain adaptation. Adv Neural Inform Process Syst 24: 2456–2464
13. Collins M, Singer Y (1999) Unsupervised models for named entity classification. Proceedings of the Joint SIGDAT conference on empirical methods in natural language processing and very large corpora, pp 100–110
14. Dasgupta S, Littman M, McAllester D (2002) PAC generalization bounds for co-training. Adv Neural Inform Process Syst 14:375–382
15. de Sa V, Gallagher P, Lewis J, Malave V (2010) Multi-view kernel construction. Mach Learn 79:47–71
16. Farquhar J, Hardoon D, Meng H, Shawe-Taylor J, Szedmak S (2006) Two view learning: SVM-2K, theory and practice. Adv Neural Inform Process Syst 18:355–362
17. Fyfe C, Lai P (2000) ICA using kernel canonical correlation analysis. Proceedings of the international workshop on independent component analysis and blind singal separation, pp 279–284
18. Ganchev K, Graça J, Blitzer J, Taskar B (2008) Multi-view learning over structured and non-identical outputs. Proceedings of the 24th conference on uncertainty in artificial intelligence, pp 204–211
19. Hardoon D, Shawe-Taylor J (2009) Convergence analysis of kernel canonical correlation analysis: theory and practice. Mach Learn 74:23–38
20. Hardoon D, Shawe-Taylor J (2011) Sparse canonical correlation analysis. Mach Learn 83:331–353
21. Hotelling H (1936) Relations between two sets of variates. Biometrika 28:321–377
22. Kettenring J (1971) Canonical analysis of several sets of variables. Biometrika 58:433–451
23. Kumar A, Daumé H (2011) A co-training approach for multi-view spectral clustering. Proceedings of the 28th international conference on machine learning, pp 393–400
24. Kumar A, Rai P, Daumé H (2011) Co-regularized multi-view spectral clustering. Adv Neural Inform Process Syst 24:1413–1421
25. Langford J (2005) Tutorial on practical prediction theory for classification. J Mach Learn Res 6:273–306
26. McAllester D (1999) PAC-Bayesian model averaging. Proceedings of the 12th annual conference on computational learning theory, pp 164–170
27. Mitchell T (1997) Machine learning. McGraw Hill, New York
28. Muslea I, Minton S, Knoblock C (2006) Active learning with multiple views. J Artif Intell Res 27:203–233
29. Nigam K, Ghani R (2000) Analyzing the effectiveness and applicability of co-training. Proceedings of the 9th international conference on information and knowledge management, pp 86–93
30. Parrado-Hernández E, Ambroladze A, Shawe-Taylor J, Sun S (2012) PAC-Bayes bounds with data dependent priors. J Mach Learn Res 13:3507–3531

31. Rosenberg D, Bartlett P (2007) The Rademacher complexity of co-regularized kernel classes. J Mach Learn Res Workshop Conf Proc 2:396–403
32. Rosenberg D, Sindhwani V, Bartlett P, Niyogi P (2009) Multi-view point cloud kernels for semisupervised learning. IEEE Signal Process Mag 145:145–150
33. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge, UK
34. Shawe-Taylor J, Sun S (2013) Kernel methods and support vector machines. Book Chapter for E-Reference Signal Processing, Elsevier
35. Sindhwani V, Niyogi P, Belkin M (2005) A co-regularization approach to semi-supervised learning with multiple views. Proceedings of the workshop on learning with multiple views, pp 824–831
36. Sindhwani V, Rosenberg D (2008) An RKHS for multi-view learning and manifold co-regularization. Proceedings of the 25th international conference on machine learning, pp 976–983
37. Sun S (2010) Local within-class accuracies for weighting individual outputs in multiple classifier systems. Pattern Recognit Lett 31:119–124
38. Sun S (2011) Multi-view Laplacian support vector machines. Lect Notes Artif Intell 7121:209–222
39. Sun S, Jin F (2011) Robust co-training. Int J Pattern Recognit Artif Intell 25:1113–1126
40. Sun S, Jin F, Tu W (2011). View construction for multi-view semi-supervised learning. Lect Notes Comput Sci 6675:595–601
41. Sun S, Hardoon D (2010) Active learning with extremely sparse labeled examples. Neurocomputing 73:2980–2988
42. Sun S, Shawe-Taylor J (2010) Sparse semi-supervised learning using conjugate functions. J Mach Learn Res 11:2423–2455
43. Sun S, Zhang C (2007) Subspace ensembles for classification. Phys A Stat Mech Appl 385:199–207
44. Sun S, Zhang C, Lu Y (2008) The random electrode selection ensemble for EEG signal classification. Pattern Recognit 41: 1663–1675
45. Sun S, Zhang Q (2011) Multiple-view multiple-learner semi-supervised learning. Neural Process Lett 34:229–240
46. Szedmak S, Shawe-Taylor J (2007) Synthesis of maximum margin and multiview learning using unlabeled data. Neurocomputing 70:1254–1264
47. Tzortzis G, Likas A (2009) Convex mixture models for multi-view clustering. Lect Notes Comput Sci 5769:205–214
48. Wang W, Zhou Z (2007) Analyzing co-training style algorithms. Lect Notes Artif Intell 4701:454–465
49. Wang W, Zhou Z (2010) A new analysis of co-training. Proceedings of the 27th international conference on machine learning, pp 1135–1142
50. White M, Yu Y, Zhang X, Schuurmans D (2012) Convex multi-view subspace learning. Adv Neural Inform Process Syst 25:1–9
51. Xu Z, Sun S (2010) An algorithm on multi-view Adaboost. Lect Notes Comput Sci 6443:355–362
52. Xu Z, Sun S (2011) Multi-view transfer learning with Adaboost. Proceedings of the 23rd IEEE international conference on tools with artificial intelligence, pp 399–402
53. Xu Z, Sun S (2012) Multi-source transfer learning with multi-view Adaboost. Lect Notes Comput Sci 7665:332–339
54. Yu S, Krishnapuram B, Rosales R, Rao R (2011) Bayesian co-training. J Mach Learn Res 12:2649-2680
55. Zhang Q, Sun S (2010) Multiple-view multiple-learner active learning. Pattern Recognit 43:3113–3119
56. Zhou Z, Zhan D, Yang Q (2007) Semi-supervised learning with very few labeled training examples. Proceedings of the 22nd AAAI conference on artificial intelligence, pp 675–680