

# A Survey on Multi-view Learning

**Chang Xu**

CHANGXU1989@GMAIL.COM

*Centre for Quantum Computation and Intelligent Systems  
Faculty of Engineering and Information Technology  
University of Technology, Sydney  
Sydney, NSW 2007, Australia*

**Dacheng Tao**

DACHENG.TAO@UTS.EDU.AU

*Centre for Quantum Computation and Intelligent Systems  
Faculty of Engineering and Information Technology  
University of Technology, Sydney  
Sydney, NSW 2007, Australia*

**Chao Xu**

XUCHAO@CIS.PKU.EDU.CN

*Key Laboratory of Machine Perception (Ministry of Education)  
School of Electronics Engineering and Computer Science  
Peking University  
Beijing 100871, China*

## Abstract

In recent years, a great many methods of learning from multi-view data by considering the diversity of different views have been proposed. These views may be obtained from multiple sources or different feature subsets. For example, a person can be identified by face, fingerprint, signature or iris with information obtained from multiple sources, while an image can be represented by its color or texture features, which can be seen as different feature subsets of the image. In trying to organize and highlight similarities and differences between the variety of multi-view learning approaches, we review a number of representative multi-view learning algorithms in different areas and classify them into three groups: 1) co-training, 2) multiple kernel learning, and 3) subspace learning. Notably, co-training style algorithms train alternately to maximize the mutual agreement on two distinct views of the data; multiple kernel learning algorithms exploit kernels that naturally correspond to different views and combine kernels either linearly or non-linearly to improve learning performance; and subspace learning algorithms aim to obtain a latent subspace shared by multiple views by assuming that the input views are generated from this latent subspace. Though there is significant variance in the approaches to integrating multiple views to improve learning performance, they mainly exploit either the consensus principle or the complementary principle to ensure the success of multi-view learning. Since accessing multiple views is the fundament of multi-view learning, with the exception of study on learning a model from multiple views, it is also valuable to study how to construct multiple views and how to evaluate these views. Overall, by exploring the consistency and complementary properties of different views, multi-view learning is rendered more effective, more promising, and has better generalization ability than single-view learning.

**Keywords:** Multi-view Learning, Survey, Machine Learning

## 1. Introduction

In most scientific data analytics problems in video surveillance, social computing, and environmental sciences, data are collected from diverse domains or obtained from various feature extractors and exhibit heterogeneous properties, because variables of each data example can be naturally partitioned into groups. Each variable group is referred to as a particular view, and the multiple views for a particular problem can take different forms, e.g. a) colour descriptor, local binary patterns, local shape descriptor, slow features and spatial temporal context captured by multiple cameras for person re-identification and global activity understanding in sparse camera network, and b) words in documents, information describing documents (e.g. title, author and journal) and the co-citation network graph for scientific document management (see Figure 1).

Conventional machine learning algorithms, such as support vector machines, discriminant analysis, kernel machines, and spectral clustering, concatenate all multiple views into one single view to adapt to the learning setting. However, this concatenation causes overfitting in the case of a small size training sample and is not physically meaningful because each view has a specific statistical property. In contrast to single view learning, multi-view learning as a new paradigm introduces one function to model a particular view and jointly optimizes all the functions to exploit the redundant views of the same input data and improve the learning performance. Therefore, multi-view learning has been receiving increased attention and existing algorithms can be classified into three groups: 1) co-training, 2) multiple kernel learning, and 3) subspace learning.

Co-training (Blum and Mitchell, 1998) is one of the earliest schemes for multi-view learning. It trains alternately to maximize the mutual agreement on two distinct views of the unlabeled data. Many variants have since been developed. Nigam and Ghani (2000) generalized expectation-maximization (EM) by assigning changeable probabilistic labels to unlabeled data. Muslea et al. (2002a, 2003, 2006) combined active learning with co-training and proposed robust semi-supervised learning algorithms. Yu et al. (2007, 2011) developed a Bayesian undirected graphical model for co-training and a novel co-training kernel for Gaussian process classifiers. Wang and Zhou (2010) treated co-training as the combinative label propagation over two views and unified the graph- and disagreement-based semi-supervised learning into one framework. Sindhwani et al. (2005) constructed a data-dependent “co-regularization” norm. The resultant reproducing kernel associated with a single RKHS simplified the theoretical analysis and extended the algorithmic scope of co-regularization. Bickel and Scheffer (2004) and Kumar et al. (2010, 2011) advanced co-training for data clustering and designed effective algorithms for multi-view data. The success of co-training algorithms mainly relies on three assumptions: (a) sufficiency - each view is sufficient for classification on its own, (b) compatibility- the target function of both views predict the same labels for co-occurring features with a high probability, and (c) conditional independence- views are conditionally independent given the label. The conditional independence assumption is critical, but it is usually too strong to satisfy in practice and thus several weaker alternatives (Abney, 2002; Balcan et al., 2004; Wang and Zhou, 2007) have been considered.

Multiple kernel learning (MKL) was originally developed to control the search space capacity of possible kernel matrices to achieve good generalization but has been widely applied

to problems involving multi-view data. This is because kernels in MKL naturally correspond to different views and combining kernels either linearly or non-linearly improves learning performance. Lanckriet et al. (2002, 2004) formulated MKL as a semi-definite programming problem. Bach et al. (2004) treated MKL as a second order cone program problem and developed an SMO algorithm to efficiently obtain the optimal solution. Sonnenburg et al. (2006a,b) developed an efficient semi-infinite linear program and made MKL applicable to large scale problems. Rakotomamonjy et al. (2007, 2008) proposed simple MKL by exploring an adaptive 2-norm regularization formulation. Szafranski et al. (2008, 2010); Xu et al. (2010) and Subrahmanya and Shin (2010) constructed the connection between MKL and group-LASSO to model group structure. Many generalization bounds have been obtained to theoretically guarantee the performance of MKL. Lanckriet et al. (2004) showed that given  $k$  base kernels, the estimation error is bounded by  $O(\sqrt{\frac{k/\gamma^2}{n}})$ , where  $\gamma$  is the margin of the learned classifier. Ying and Campbell (2009) used the metric entropy integrals and pseudo-dimension of a set of candidate kernels to estimate the empirical Rademacher chaos complexity. The generalization bounds have a logarithmic dependency on  $k$  for the family of convex combinations of  $k$  base kernels with the  $l_1$  constraint. Assuming different views to be uncorrelated, Kloft and Blanchard (2011) derived a tighter upper bound by the local Rademacher complexities for the  $l_p$ -norm MKL. The cited survey (Gönen and Alpaydm, 2011) is believed to contain all the related references omitted from the proposal.

Subspace learning-based approaches aim to obtain a latent subspace shared by multiple views by assuming that the input views are generated from this latent subspace. The dimensionality of the latent subspace is lower than that of any input view, so subspace learning is effective in reducing the “curse of dimensionality”. Given this subspace, it is straightforward to conduct the subsequent tasks, such as classification and clustering. Canonical correlation analysis (CCA) (Hotelling, 1936) and kernel canonical correlation analysis (KCCA) (Akaho, 2006) explore basis vectors for two sets of variables by mutually maximizing the correlations between the projections onto these basis vectors, so it is straightforward to apply them to two-view data to select the shared latent subspace. They have been further developed to conduct multi-view clustering (Chaudhuri et al., 2009) and regression (Kakade and Foster, 2007). Diethe et al. (2008) generalized Fisher’s discriminant analysis to explore the latent subspace spanned by multi-view data. In contrast to CCA, this generalization considers the class label information. Quadrianto and Lampert (2011) and Zhai et al. (2012) studied multi-view metric learning by constructing embedding projections from multi-view data to a shared subspace, where the Euclidean distance is meaningful across different views. The latent subspace is valuable for inferring another view from the observation view. Shon et al. (2006) exploited Gaussian process, Sigal et al. (2009) maximized the mutual information, and Chen et al. (2010) used Markov network to construct the connections between the two views through latent subspaces. Salzman et al. (2010) and Jia et al. (2010) proposed to find a latent subspace in which the information is correctly factorized into shared and private parts across different views. Consistency and finite sample analysis (Fukumizu et al., 2007; Hardoon and Shawe-Taylor, 2009; Cai and Sun, 2011) have been studied for KCCA.

In reviewing the literature on multi-view learning, we find it is tightly connected with other topics in machine learning, such as active learning, ensemble learning and domain adaptation. Active learning (Settles, 2009; Seung et al., 1992), sometimes called query

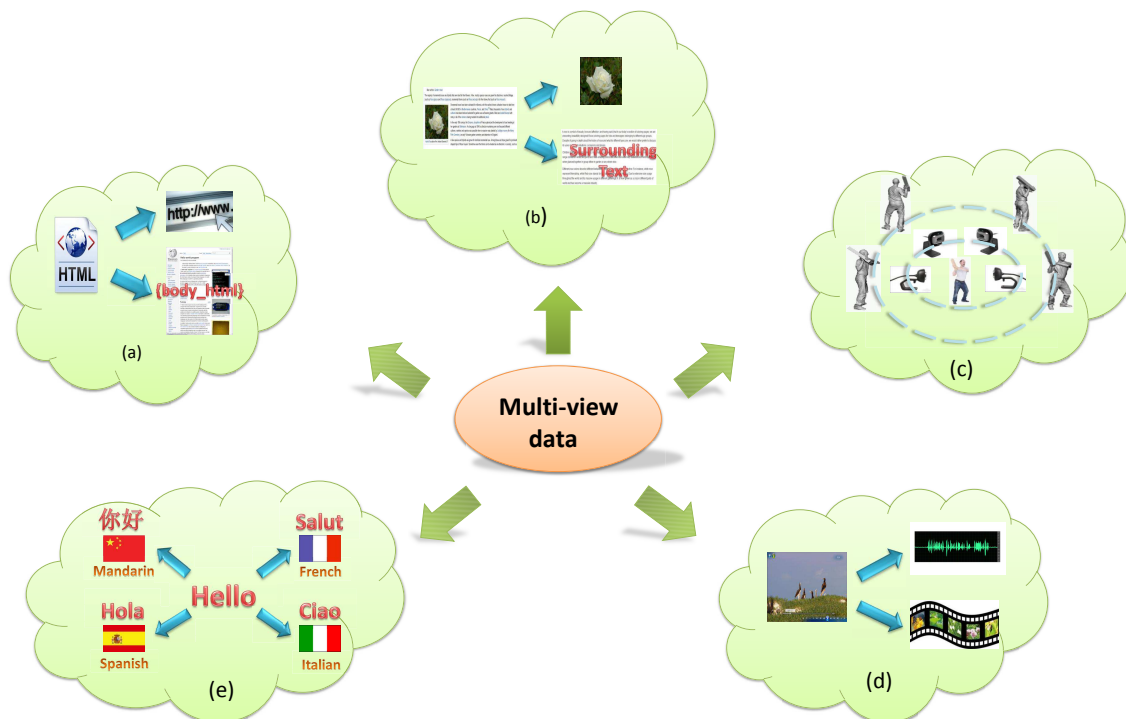


Figure 1: Multi-view data: a) a web document can be represented by its url and words on the page, b) a web image can be depicted by its surrounding text separate to the visual information, c) images of a 3D object taken from different viewpoints, d) video clips are combinations of audio signals and visual frames, e) multilingual documents have one view in each language.

learning, aims to minimize the amount of labeled data required for learning a concept of interest. Muslea et al. (2000) introduced co-testing, which is a novel approach to conducting active learning with multiple views. They combined co-testing with co-EM and derived a novel method co-EMT (Muslea et al., 2002a), which uses co-EM to generate accurate classifiers and chooses the most informative unlabeled examples for co-Testing to label. Furthermore, considering strong and weak views, Muslea et al. (2003, 2006) advanced co-Testing by assuming that the concentrated examples whose labels from strong classifiers are different and inconsistent with the prediction of the weak classifier provide more information for labeling. The idea of ensemble learning (Dietterichl, 2002; Lappalainen and Miskin, 2000) is to employ multiple learners and combine their predictions. The bagging algorithm (Breiman, 1996) uses different training datasets to construct each member of the ensemble and predicts through uniform averaging or voting over class labels. In contrast to co-training, which ensures the diversities of the learned models by training on distinct views, bagging requires different training datasets for generating models with different judgments. AdaBoost (Freund and Schapire, 1996) is another well-known ensemble learning algorithm, in which the principal idea is to train a new model to compensate for the errors made by earlier models. In each round, the misclassified examples are identified and their emphasis will

be increased in a new training set for the next training process. Both co-training and AdaBoost rely on a growing ensemble of classifiers trained on resamples of the data; however, AdaBoost tries to find the error-labeled examples, whereas co-training attempts to exploit the agreement of the learners. Co-training is confidence-driven whereas AdaBoost is error-driven. Domain adaptation refers to the problem of adapting a prediction model trained on data from a source domain to a different target domain, where the data distributions in the two domains are different. Many domain adaptation techniques (Wei and Pal, 2010; Wan et al., 2011) have been proposed to solve the cross-language text classification problem, where the source domain includes the documents translated from the source language and the target domain includes the original documents in the target language. Moreover, these documents in different languages can be seen as different views of the original document; thus, methods like co-training (Wan, 2009), multi-view majority voting (Amini et al., 2010) and multi-view co-classification (Amini and Goutte, 2010) have been designed and successfully applied for this problem.

In this survey paper, we provide a comprehensive overview of multi-view learning. The rest of this paper is organized as follows: we first illustrate the principles underlying multi-view learning algorithms in Section 2. In Section 3, different approaches to the construction of multiple views and methods to evaluate these views are introduced. We present different ways to combine multiple views in Section 4 and illustrate different kinds of multi-view learning algorithm in detail in Sections 5, 6 and 7. The applications of multi-view learning are introduced in Section 8 and experimental results reflecting the performance of multi-view learning are shown in Section 9. Finally, we conclude the paper in Section 10.

## 2. Principles for Multi-view Learning

The demand for redundant views of the same input data is a major difference between multi-view and single-view learning algorithms. Thanks to these multiple views, the learning task can be conducted with abundant information. However if the learning method is unable to cope appropriately with multiple views, these views may even degrade the performance of multi-view learning. Through fully considering the relationships between multiple views, several successful multi-view learning techniques have been proposed. We analyze these various algorithms and observe that there are two significant principles ensuring their success: *consensus* and *complementary* principles.

### 2.1 Consensus Principle

Consensus principle aims to maximize the agreement on multiple distinct views. Suppose the available data set  $X$  has two views  $X^1$  and  $X^2$ . An example  $(x_i, y_i)$  is therefore viewed as  $(x_i^1, x_i^2, y_i)$ , where  $y_i$  is the label associated with the example. Dasgupta et al. (2002) demonstrated the connection between the consensus of two hypotheses on two views respectively and their error rates. Under some mild assumptions they gave the inequality

$$P(f^1 \neq f^2) \geq \max\{P_{err}(f^1), P_{err}(f^2)\}.$$

From the inequality, we conclude that the probability of a disagreement of two independent hypotheses upper bounds the error rate of either hypothesis. Thus by minimizing the disagreement rate of the two hypotheses, the error rate of each hypothesis will be minimized.

In recent years, a great number of methods appear to have utilized this consensus principle in one way or another, even though in many cases the contributors are not aware of the relationship between their methods and this common underlying principle. For example, the co-training algorithm trains alternately to maximize the mutual agreement on two distinct views of the unlabeled data. By minimizing the error on labeled examples and maximizing the agreement on unlabeled examples, the co-training algorithm finally achieves one accurate classifier on each view. In the co-regularization algorithm, the consensus principle can be formulated by regularization terms as

$$\min \sum_{i \in U} [f^1(x_i) - f^2(x_i)]^2 + \sum_{i \in L} V(y_i, f(x_i)), \quad (1)$$

where the first term enforces the agreement on two distinct views on unlabeled examples, and the second term evaluates the empirical loss on the labeled examples with respect to a loss function  $V(\cdot, \cdot)$ . By additionally considering the complexity of the hypotheses, we will achieve the complete objective function, and solving it will result in learning two optimal hypotheses. Observing that applying the kernel canonical correlation analysis (KCCA) to the two feature spaces can improve the performance of the classifier, Farquhar et al. (2005) proposed a supervised learning algorithm called SVM-2K, which combines the idea of KCCA with SVM. An SVM can be thought of as projecting the feature to a 1-dimensional space followed by thresholding, after which SVM-2K forces the constraint of consensus of two views on this 1-dimensional space. Formally this constraint can be written as

$$\|f^1(x_i^1) - f^2(x_i^2)\| \leq \eta_i + \epsilon$$

where  $\eta_i$  is a variable that imposes consensus between the two views, and  $\epsilon$  is a slack variable. In multi-view embedding, we conduct the embedding for multiple features simultaneously while considering the consistency and complement of different views. For example, the multi-view spectral embedding (Xia et al., 2010) first builds a patch for a sample on each view, in which the arbitrary point and its  $k$  nearest neighbors are forced to have similar outputs in the low-dimensional embedding space. Following this local consensus optimization, all the patches from different views are unified as a whole by global coordinate alignment. This can be seen as a global consensus optimization.

## 2.2 Complementary Principle

The complementary principle states that in a multi-view setting, each view of the data may contain some knowledge that other views do not have; therefore, multiple views can be employed to comprehensively and accurately describe the data. In machine learning problems involving multi-view data, the complementary information underlying multiple views can be exploited to improve the learning performance by utilizing the complementary principle.

Nigam and Ghani (2000) used the classifier learned on one view to label the unlabeled data, and then prepared these newly labeled examples for the next iteration of classifier training on another view. On the unlabeled data set, the models on two views therefore shared the complementary information with each other, which led to an improvement in the learning performance. Wang and Zhou (2007) studied why co-training style algorithms can

succeed when there are no redundant views. They used different configurations of the same base learner, which can be seen as another kind of view, to describe the data in different approaches, and showed that when the diversity between the two learners is greater than the amount of errors, the performance of the learners can be improved by co-training style algorithms. The two classifiers which have different biases will label some examples with different labels. If the examples labeled by the classifier  $h_1$  on one view are to be useful for the classifier  $h_2$  on the other view,  $h_1$  should contain some information that  $h_2$  does not know. The two classifiers will thus exchange complementary information with each other and learn from each other under the complementary principle. As the co-training process proceeds, the two classifiers will become increasingly similar, until the performance cannot be further improved.

In multiple kernel learning, different kernels may correspond to different notations of similarity. Since different ways of measuring the similarity of the data have specific advantages, we resort to a learning method that makes an appropriate combination under the complementary principle rather than trying to establish which kernel works best. Thus all kinds of notations of similarity will work together to achieve an accurate evaluation of the data. In addition, different kernels can also use inputs from a variety views, possibly from alternative sources or modalities. Thus by considering the complementary information underlying various views of the data and combining multiple kernels from these distinct views, a comprehensive measurement of the similarity can be obtained.

One traditional solution for the multi-view problem is to concatenate vectors from different views into a new vector and then apply single-view learning algorithms straightforwardly on the concatenated vector. However, this concatenation causes over-fitting on a small training sample, and the specific statistical property of each view is ignored. For many applications with long feature vectors on more than one view as input, it is therefore reasonable to construct a shared low-dimensional representation for these views. In human pose inference, the image features and 3D poses can be seen as two complementary views that describe human poses. Several methods (Shon et al., 2006; Sigal et al., 2009) have been designed to tackle this problem by constructing a latent subspace shared by multiple views, in which distinct views are connected with one another in this subspace, integrating the complementary information underlying different views. At inference, given a new observation on one view, it is possible to find the corresponding latent embedding, which is also connected with the point on the other view. Xia et al. (2010) developed a new spectral embedding algorithm, namely, multi-view spectral embedding (MSE), which encodes multi-view features to achieve a physical meaningful embedding. Yu et al. (2012b) proposed a semi-supervised multi-view distance metric learning (SSM-DML) for cartoon character retrieval. Since various low-level features can be extracted to represent the image, each feature space will give one measurement of similarity of the data, so it is difficult to decide which measurement is the most suitable. By considering the complementary information underlying distinct views, advantage can be taken of metric learning to construct a shared latent subspace to precisely measure the dissimilarity between different examples.

Both complementary and consensus principles play important roles in multi-view learning. For example, in co-training style algorithms, Dasgupta et al. (2002) have shown that by minimizing the disagreement rate of the two hypotheses on two views respectively, the error rate of each hypothesis can be minimized. On the other hand, Wang and Zhou (2007)

established that the reason for the success of co-training style algorithms is the extent of the diversity between the two learners; in other words, it is the complementary information in distinct views that influences the performance of co-training style algorithms. In addressing the problem of multi-view learning, both the consensus and complementary principles should be kept in mind to take full advantage of multiple views.

### 3. View Generation

The priority for multi-view learning is the acquisition of redundant views, which is also a major difference from single-view learning. Multiple view generation not only aims to obtain the views of different attributes, but also involves the problem of ensuring that the views sufficiently represent the data and satisfy the assumptions required for learning. In this section, we will illustrate how to construct multiple views and how to evaluate these views.

#### 3.1 View Construction

In practice, objects can frequently be described from different points of view. One classic multi-view example is the web classification problem. Usually, a web document can be described by either the words occurring on the page or the words contained in the anchor text of links pointing to this page. In many cases, however, no natural multiple views are available because of certain limitations, so that only one view may be provided to represent the data. Since it is difficult to straightforwardly conduct multi-view learning on this single view, the preliminary work of multi-view learning concerns the construction of multiple views from this single view.

Generating different views corresponds to feature set partitioning, which generalizes the task of feature selection. Instead of providing a single representative set of features, feature set partitioning decomposes the original set of features into multiple disjoint subsets to construct each view. A simple way to convert from a single view to multiple views is to split the original feature set into different views at random, and there indeed a number of experiments in multi-view learning employing this trick (Brefeld et al., 2005; Bickel and Scheffer, 2004; Brefeld and Scheffer, 2004). However, there is no guarantee that a satisfactory result will be obtained using this approach. Therefore, subsetting the feature set in a way that adheres to the multi-view learning paradigm is not a trivial task, and is dependent on both the chosen learner and the data domain.

The random subspace method (RSM) (Ho, 1998), as an example of a random sampling algorithm, incorporates the benefits of bootstrapping and aggregation. Unlike bagging the bootstraps training samples, RSM performs the bootstrapping in the feature space. This method relies on an autonomous, pseudo random procedure to select a small number of dimensions from a given feature space. This selection is made and a subspace is fixed by giving all points a constant value (zero) in the unselected dimensions, in each pass. For a given feature space of  $n$  dimensions, there are  $2^n$  such selections that can be constructed. All the subspaces can then be regarded as different views of the data. While most other methods suffer from the curse of dimensionality, this method takes advantage of the high dimensionality. Tao et al. (2006) employed the random subspace method to sample several small sets of features to reduce the discrepancy between the training data size and the



feature vector length. Based on the sampled subspaces, multiple SVMs can be constructed and then be combined to obtain a more powerful classifier to solve the over-fitting problem.

Di and Crawford (2012) conducted a thorough investigation of view generation for hyperspectral image data. Considering the key issues: diversity, compatibility and accuracy, several strategies have been proposed to construct multiple views for hyperspectral data, as follows. 1) Clustering: these methods involve feature aggregation based on similarity metrics, with the goal of promoting diversity between views. 2) Random selection: in conjunction with feature space bagging, random selection can result in greater information exploration from the spectral feature space and can eliminate the impact of generating uninformative or corrupted views. 3) Uniform band slicing: uniform division of the data across the full spectral range creates views that contain bands separated by equal intervals, thus guaranteeing view sufficiency. The authors also proposed that increasing the number of views to increase diversity, or increasing randomness from the feature space to avoid insufficient or noisy views, further improves performance.

With respect to learning problems involving textual documents, Matsubara et al. (2005) proposed a pre-processing approach to easily construct different views required by multi-view learning algorithms. By identifying terms as bag-of-words and using different numbers of words to constitute each term, different representations of one document for different views can be obtained. This is a simple yet effective approach to the construction of multiple views for textual documents, although it is difficult to apply to other domains. Wang et al. (2011) developed a novel technique to reshape the original vector representation of a single view into multiple matrix representations. For instance, a vector  $x = [a, b, c, d, e, f]^T$  can be reshaped into two different matrices:

$$\begin{pmatrix} a & c & e \\ b & d & f \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}^T.$$

Different ways of reshaping the vector induce multiple matrix patterns with a variety of dimensional sizes of rows and columns. These matrixes can be regarded as multiple independent or weaker correlated views of the input data. Utilizing the matrix representation, the required memory can be saved, new implicit information is introduced through the new constraint in the structure, and then the performance of the classifiers learned will be improved, compared to the vector representation.

Chen et al. (2011) suggested a novel feature decomposition algorithm called Pseudo Multi-view Co-training (PMC) to automatically divide the features of a single view dataset into two mutually exclusive subsets. Considering the linear classifier,  $f(x) = \mathbf{w}x + b$  given the weight vector  $\mathbf{w}$ , the optimization can be written as

$$\min_{\mathbf{w}_1, \mathbf{w}_2} \log(e^{\mathcal{L}(\mathbf{w}_1; L)} + e^{\mathcal{L}(\mathbf{w}_2; L)}), \quad (2)$$

where  $\mathbf{w}_1$  and  $\mathbf{w}_2$  are weight vectors for two classifiers respectively, and  $\mathcal{L}(\mathbf{w}; L)$  is the log-loss over the dataset  $L$ . To make sure that the two classifiers are trained on different views of the dataset, for each feature  $i$ , at least one of the two classifiers must have a zero weight in the  $i$ -th dimension. This constraint can be written as

$$\forall i, 1 \leq i \leq d, \quad \mathbf{w}_1^i \mathbf{w}_2^i = 0. \quad (3)$$

In each iteration, solving the above optimization problem will automatically find an optimal split of the features.

To obtain the feature subsets automatically, Sun et al. (2011) turned to genetic algorithms (GAs) for help. Each bit in the binary bit strings in GAs is associated with one feature. If the  $i$ -th feature is selected, the  $i$ -th bit is 1, otherwise this bit is 0. Suppose the size of the population is  $n$ , then in each iteration, the best  $n$  individuals will be selected as the next generation. Each individual in the final genetic population corresponds to a candidate feature subset, which can be regarded as one view of the data.

The literature shows that several kernel functions have been successfully used, such as the linear kernel, the polynomial kernel, and the Gaussian kernel. Since different kinds of kernel function correspond to different notations of similarity, it is reasonable instead of selecting one specific kernel function to describe the data to obtain an optimal combination of these kernel functions. These different kinds of kernel function can be seen as distinct views of the data, and the problem of how to learn the kernel combination can therefore be cast as multiple kernel learning.

The above view construction methods can be analyzed and categorized into three classes. The first class includes techniques that construct multiple views from meta data through random approaches. The second class consists of algorithms that reshape or decompose the original single-view feature into multiple views, such as the above matrix representations or different kernel functions. The third class is composed of methods that perform feature set partitioning automatically, such as PMC (Chen et al., 2011). This last type of algorithm bears some connections with the mature feature selection algorithms (Jain and Zongker, 1997; Guyon and Elisseeff, 2003); however, there are significant differences between multi-view feature selection and single-view feature selection. In multi-view feature selection, the relationships between multiple views should additionally be considered, besides the information within each view.

### 3.2 View Evaluation

Constructing multiple views is just one task of view generation; another significant aspect is to evaluate these views and ensure their effectiveness for multi-view learning algorithms. Several approaches have been proposed in the multi-view learning literature that analyze the relationships between multiple views or cope with the problems resulting from the violation of view assumptions or the noise in the views.

Muslea et al. (2002b) first introduced a view validation algorithm which predicts whether or not the views are sufficiently compatible for solving multi-view learning tasks. This algorithm tries to learn a decision tree in a supervised way to discriminate between learning tasks according to whether or not the views are sufficiently compatible for multi-view learning. A set of features is designed to indicate how incompatible the views are, and the label of each instance is generated automatically by comparing the accuracy of single- and multi-view algorithms on a test set.

The assumption of view sufficiency does not generally hold in practice. For example, in the task of video concept detection, one frame contains an airplane and the other contains an eagle, but both frames may have the same color histogram feature. Therefore, it is difficult for the low-level visual features alone to sufficiently represent the concepts.

Yan and Naphade (2005) proposed semi-supervised cross-feature learning (SCFL) to alleviate the problems of co-training when some views are inadequate for learning concepts by themselves. When view sufficiency assumption fails, the main concern in applying co-training is that the additional training data associated with classification noise are likely to corrupt the initial classifiers. After labeling unlabeled data using the initial classifiers of two views, two separate classifiers from each view, based solely on the unlabeled data, are trained to eliminate this problem. With the assistance of validation data  $V$ , all four classifiers can be weighted combined to detect how much benefit can be achieved from the unlabeled data without hurting the performance of the initial classifiers. If the predictions from the unlabeled data are too noisy to use, the combined weights of the two classifiers newly learned on the unlabeled data can simply be zeroed, and we back off to the initial classifiers trained on the labeled data.

The performance of multi-view learning algorithms may be influenced by the noises in the views. Christoudias et al. (2008) defined a view disagreement problem, stating that the samples from each view do not always belong to the same class but sometimes belong to an additional background class as a result of noise. To detect and filter view disagreement, a conditional view entropy  $H(x^i|x^j)$  was defined as a measure of the uncertainty in view  $x^i$  given the observed view  $x^j$ . The conditional view entropy is expected to be larger when conditioning takes place on the background rather than the foreground. By thresholding the conditional view entropy, the samples whose views display disagreement can be discarded in each iteration of the co-training algorithm, and then the performance of the classifiers is improved.

Yu et al. (2011) proposed a probabilistic approach to co-training, called Bayesian co-training, which copes with per-view noise. This algorithm employs a latent variable  $f_j$  for each view and a consensus latent variable,  $f_c$  to model the agreement on different views. Finally  $\psi(f_j, f_c)$  is defined to denote the compatibility between the  $j$ -th view and the consensus function and can be written as,  $\psi(f_j, f_c) = \exp(-\frac{\|f_j - f_c\|^2}{2\sigma_j^2})$ . The parameters  $\{\sigma_j\}$  act as reliability indicators and control the strength of interaction between the  $j$ -th view and the consensus latent variable. A small value of  $\sigma_j$  has a strong influence on the view in the final output, whereas a large value allows the model to discount observations from that view. Thus the Bayesian co-training model can handle per-view noise, where each sample of a given view is assumed to be corrupted by the same amount of noise. Christoudias et al. (2009a) extended Bayesian co-training to the heteroscedastic case, in which each observation can be corrupted by a different noise level. Assume that the latent functions can be corrupted with arbitrary Gaussian noise such that

$$\psi(f_j, f_c) = \mathcal{N}(f_j, \mathbf{A}_j),$$

where  $\mathbf{A}_j$  is the noise covariance matrix. When assuming i.i.d. noise, the noise matrix can be written as

$$\mathbf{A}_j = \text{diag}(\sigma_{1,j}^2, \dots, \sigma_{N,j}^2),$$

where  $\sigma_{i,j}^2$  is the estimation of the noise corrupting sample  $i$  in view  $j$ . Thus the heteroscedastic Bayesian co-training model can incorporate sample-dependent noise modeled by the per-view noise covariance matrices  $\mathbf{A}_j$ .

In multiple kernel learning, different kernels may use inputs coming from various representations, possibly from a range of modalities or sources. These representations may have contrasting measures of similarity corresponding to different kernels, and can be regarded as different views of the data. In this case, combining kernels is one possible way to combine multiple information sources; however, in the real world, the sources may be corrupted by disparate noises, so when some of the kernels are noisy or irrelevant, it is necessary to optimize the kernel weights in the learning process. Lewis et al. (2006) compared the performances of unweighted and weighted sums of kernels on a gene functional classification task. They considered a case in which additional, noisy kernels are added to the system. As more noise is added to the system, the performance of the unweighted average deteriorates, but the weighted kernel approach learns to down-weight the noise kernels and hence continues to work well. Most multiple kernel learning algorithms are global techniques under the assumption of a per-view kernel weighting, and these methods therefore cannot cope with the presence of complex noise processes, such as heteroscedastic noise, or missing data. Christoudias et al. (2009b) presented a Bayesian localized approach for combining different feature representations with Gaussian processes that learns a local weighting over each view. Let  $\bar{\mathbf{X}} = [\mathbf{X}^1, \dots, \mathbf{X}^V]$  be the set of all observations with  $V$  views, let  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]^T$  be the set of labels, and let  $\mathbf{f} = [\mathbf{f}_1, \dots, \mathbf{f}_N]^T$  be a set of latent functions. The Gaussian Process (GP) prior over the latent functions can be written as  $p(\mathbf{f}|\bar{\mathbf{X}}) = \mathcal{N}(0, \bar{\mathbf{K}})$ . If a Gaussian noise model is used, then  $p(\mathbf{Y}|\mathbf{f}) = \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})$  is obtained. The covariance function can be obtained by combining the covariances of feature representations in a non-linear manner; thus, classification is performed using the standard GP approach with common covariance function.

Liu and Yuen (2011) introduced two new confidence measures, namely, inter-view confidence and intra-view confidence, to describe the view sufficiency and view dependency issues in multi-view learning. Considering the sample  $X$  associated with  $M$  views, the observed data are represented as  $X^1, \dots, X^M$  respectively; based on the mutual information definition, the inter-view confidence of  $X$  is defined as

$$C_{inter}(X) = \sum_{i=1}^M \sum_{j=i}^M \frac{1}{I(X^i, X^j)},$$

where  $I(X^i, X^j)$  measures the mutual information between  $X^i$  and  $X^j$ . By maximizing the inter-view confidence, the total data dependency is minimized. In addition, the authors proposed the calculation and minimization of the total inconsistency of labeled and unlabeled data iteratively in a semi-supervised manner. Thus the view sufficiency can be defined as

$$C_{intra}(X) = \sum_{i=1}^M \frac{1}{F(X_i^L, X_i^U, S_i)},$$

where  $X_i^L$  and  $X_i^U$  are the labeled and unlabeled dataset respectively,  $S_i$  is the similarity matrix for view  $i$ , and  $F$  measures the data consistency between  $X_i^L$  and  $X_i^U$ .

Correlation between views is an important consideration in subspace-based approaches for multi-view learning. Hotelling (1936) introduced canonical correlation analysis (CCA) to describe the linear relation between two views which aims to compute a low-dimensional

shared embedding of both views of variables such that the correlations among the variables between the two views is maximized in the embedded space. Since the new subspace is simply a linear system of the original space, CCA can only be used to describe linear relation. Under Gaussian assumption, the CCA can also be used to test stochastic independence between two views of variables. Akaho (2006) studied a hybrid approach of CCA with a kernel machine, called kernel canonical correlation analysis (KCCA), to identify non-linearly correlated projections between the two views. Formally, for two views  $X \in \mathbb{R}^{d \times n}$  and  $Y \in \mathbb{R}^{k \times n}$ , CCA computes two projection vectors,  $w_x \in \mathbb{R}^d$  and  $w_y \in \mathbb{R}^k$ , such that the following correlation coefficient:

$$\rho = \frac{w_x^T X Y^T w_y}{\sqrt{(w_x^T X X^T w_x)(w_y^T Y Y^T w_y)}}$$

is maximized. Similarly in KCCA, we express the projection direction as  $w_x = X\alpha$  and  $w_y = Y\beta$ , where  $\alpha$  and  $\beta$  are vectors of size  $N$ . Irrespective of whether CCA or KCCA is used, a sequence of correlation coefficients  $\{\rho_1, \rho_2, \dots\}$  arranged in descending order can be obtained. Several measures of association in the literature are constructed as functions of the correlation coefficients, of which the two most common association measures are as follows. One is the maximal correlation

$$r(X, Y) = \rho_1,$$

and the other is

$$r(X, Y) = -\sum_{i=1} \log(1 - \rho_i^2).$$

#### 4. View Combination

One traditional way to combine multiple views is to concatenate all multiple views into a single view to adapt to the single-view learning setting. However, this concatenation causes over-fitting on a small training sample and is not physically meaningful because each view has a specific statistical property. Thus we resort to advanced methods of combining multiple views to achieve the improvement in learning performance compared to single-view learning algorithms.

Co-training style algorithms usually train separate but correlated learners on each view, and the outputs of learners are forced to be similar on the same validation points, as shown in Figure 2, Under the consensus principle, the goal of each iteration is to maximize the consistency of two learners on the validation set. Certainly there may be some disagreement between the predictions from the two learners on the validation set; however, this disagreement is propagated back to the training set to help to train more accurate learners, thus minimizing the disagreement on the validation set in the next iteration.

Co-training is a classical algorithm in semi-supervised learning. In co-training, a classifier is trained on per-view, which only uses the features from that view. By maximizing the agreement on the predictions of two classifiers on the labeled dataset, as well as minimizing the disagreement on the predictions of two classifiers on the unlabeled dataset, the classifiers learn from each other and reach an optimal solution. Here, the unlabeled set is considered

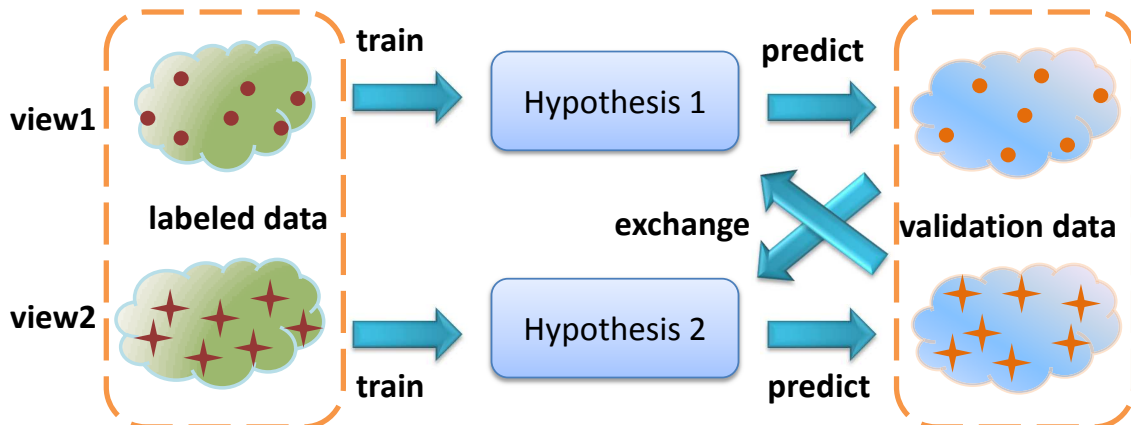


Figure 2: The process of co-training style algorithms.

to be the validation set. In each iteration, the learner on one view labels unlabeled data which are then added to the training pool of the other learner; therefore, the information underlying two views can be exchanged in this scheme. Co-regularization can be regarded as a regularized version of the co-training algorithm. Unlike co-training, the co-regularization algorithm formally measures the agreement on two distinct views using Eq. 1. By solving the corresponding objective problem, two optimal classifiers can be obtained.

If a validation set is not provided, for example in an unsupervised learning setting, it is necessary to train the classifier on each view as well as validate the combination of views on the same training set. Kumar and Daumé III (2011) applied the idea of co-training to the unsupervised learning setting and proposed a spectral clustering algorithm for multi-view data. Under the assumption that the true underlying clustering would assign corresponding points in each view to the same cluster, this algorithm solves spectral clustering on individual graphs to obtain the discriminative eigenvectors  $\mathbf{U}_1(\mathbf{U}_2)$  in each view, then clusters points using  $\mathbf{U}_1(\mathbf{U}_2)$  and uses this clustering to modify the graph structure in views 2(1) respectively. This process is repeated for a number of iterations. Similar to many other multi-view clustering algorithms (Kumar et al., 2010, 2011), multiple views in this setting are usually combined on the training set considering the consensus principle. In multi-view supervised learning problems, an implicit validation set is also employed to combine multiple views. For example, in the Bayesian co-training proposed by Yu et al. (2011), a Bayesian undirected graphical model for co-training through gauss process is constructed. A latent function  $f_c$  is introduced to ensure the conditional independence between the output  $y$  of each example and latent functions  $f_j$  for each view. Thus  $\{f_c\}$  can be seen as an implicit validation set which connects multiple views in a latent space.

Instead of choosing a single kernel function for multiple kernel learning, it is better to use a set and allow an algorithm to choose suitable kernels and the kernel combination. Since different kernels may correspond to various notions of similarity or inputs coming from different representations, possibly from a number of sources or modalities, combining kernels is one possible way to integrate multiple information sources and find a better solution, as shown in Figure 3. There are several ways in which the combination can be

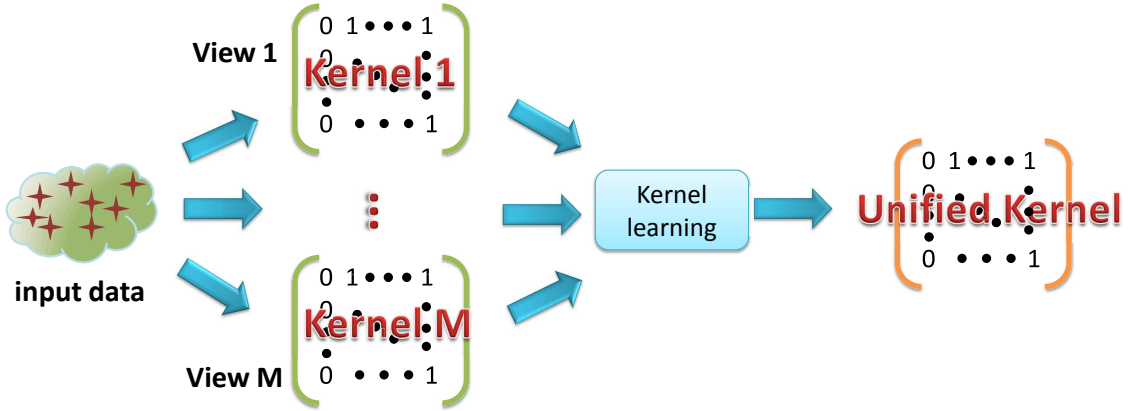


Figure 3: Sketch map of multiple kernel learning.

made, and each has its own combination parameter characteristics. These methods can be grouped into two categories:

1. *Linear combination methods*

There are several linear ways to combine multiple kernels. These methods have two basic categories:

- Direct summation kernel

$$K(x_i, x_j) = \sum_{k=1}^M K_k(x_i, x_j), \quad (4)$$

- Weighted summation kernel

$$K(x_i, x_j) = \sum_{k=1}^M d_k K_k(x_i, x_j). \quad (5)$$

Using an unweighted sum gives equal preference to all kernels, which may not be ideal; a weighted sum may be a better choice. In the latter case, versions of this approach differ in the way they place restrictions on the kernel weights  $\{d_k\}_{k=1}^M$ . Lanckriet et al. (2002, 2004) used a direct approach to optimize the unrestricted kernel combination weights. The combined kernel matrix is selected from the following set:

$$\mathcal{K} = \{K : K = \sum_{k=1}^M d_k K_k, K \geq 0, tr(K) \leq c\}.$$

Lanckriet et al. (2004) restricted the combination weights to non-negative values by selecting the combined kernel matrix from the set:

$$\mathcal{K} = \{K : K = \sum_{k=1}^M d_k K_k, d_k \geq 0, K \geq 0, tr(K) \leq c\}$$

Thorsten Joachims and Shawe-Taylor (2001) followed the constraints  $d_k \geq 0$ ,  $\sum_{k=1}^M d_k = 1$ , and considered the convex combination of kernel weights. If only binary  $d_k$  for kernel selection is allowed, the kernels whose  $d_k = 0$  can be discarded and only the kernels whose  $d_k = 1$  are used. Xu et al. (2009b) used this definition to perform feature selection. Usually the same weight is assigned to a kernel over the whole input space, which ignores the data distribution of each local region. Gönen and Alpaydin (2008) proposed to assign different weights to kernel functions according to data distribution, and defined the locally combined kernel matrix as

$$K(x_i, x_j) = \sum_{k=1}^M d_k(x_i) K_k(x_i, x_j) d_k(x_j), \quad (6)$$

where  $d_k(x)$  is the gating function which chooses feature space as a function of input  $x$ .

## 2. Nonlinear combination methods

Linear combinations of base kernels are limited, thus far richer representation can be achieved by combining kernels in other fashions. Varma and Babu (2009) tried to use the products of base kernels and other combinations which yield positive definite kernels to perform multiple kernel learning; for example, the exponentiation and power way of combining kernels:

$$K(x_i, x_j) = \exp\left(-\sum_{k=1}^M d_k x_i^T \mathbf{A}_k x_j\right),$$

or

$$K(x_i, x_j) = \left(d_0 + \sum_{k=1}^M d_k x_i^T \mathbf{A}_k x_j\right)^n.$$

Another work by Cortes et al. (2009) is a non-linear kernel combination method based on kernel regression and the polynomial combination of kernels. They proposed to combine kernels as follows:

$$K = \sum_{0 \leq k_1 + \dots + k_M \leq d, k_m \geq 0} \mu_{k_1 \dots k_M} \prod_{m=1}^M K_m(x_i, x_j)^{k_m}$$

with

$$\mu_{k_1 \dots k_M} \geq 0.$$

A special case is considered:

$$K = \sum_{k_1 + \dots + k_M = d, k_m \geq 0} \prod_{m=1}^M \mu_m^{k_m} K_m(x_i, x_j)^{k_m},$$

with

$$\mu_{k_1 \dots k_M} \geq 0.$$



Consequently, the objective of the algorithm is to find the vector  $\mu = (\mu_1, \dots, \mu_M)^T$ . However, the empirical results do not show consistent performance improvement, bringing into question whether the non-linear combination of kernel functions is necessary or efficient.

Subspace learning-based approaches aim to obtain a latent subspace shared by multiple views by assuming that the input views are generated from this latent subspace, as illustrated in Figure 4. In the literature on single-view learning, principal component analysis (PCA) is the time-honoured and simplest technique to exploit the subspace for single-view data. Canonical correlation analysis (CCA) can be viewed as the multi-view version of PCA, and it has become a general tool for performing subspace learning for multi-view data. Through maximizing the correlation between the two views in the subspace, CCA outputs one optimal projection on each view; however, since the subspace constructed by CCA is linear, it is impossible to straightforwardly apply this to many real world datasets exhibiting non-linearities. Thus the kernel variant of CCA, namely KCCA, can be thought of in terms of first mapping each data point to a higher space in which linear CCA operates. Both CCA and KCCA exploit the subspace in an unsupervised way, so that the label information is ignored. Motivated by the generation of CCA from PCA, multi-view Fisher discriminant analysis is developed to find informative projections with label information. Lawrence (2004) cast the Gaussian process as a tool to construct a latent variable model which could accomplish the task of non-linear dimensional reduction. Chen et al. (2010) developed a statistical framework that learns a predictive subspace shared by multiple views based on a generic multi-view latent space Markov network. Quadrianto and Lampert (2011) studied the metric learning problem in cross-media retrieval tasks. The goal of metric learning for multi-view data is to learn metrics with which the original multi-view higher dimensional features can be projected into a shared feature space, so that the Euclidean distance in this space is meaningful not only within a single view, but also among different views. Since the subspace constructed through different methods usually has lower dimensionality than that of any input view, the “curse of dimensionality” problem is effectively eliminated, and given the subspaces, it is straightforward to conduct subsequent tasks such as classification and clustering.

After analyzing the various approaches above to combine multiple views, we sum up their similarities and differences as follows. (a) Co-training style algorithms usually train separate learners on distinct views, which are then forced to be consistent across views. Thus this kind of approach can be regarded as a late combination of multiple views because the views are considered independently while training base learners. (b) Multiple kernel learning algorithms calculate separate kernels on each view which are combined with a kernel-based method. This kind of approach can be thought of as an intermediate combination of multiple views because kernels (views) are combined just before or during the training of the learner. (c) Subspace learning-based approaches aim to obtain an appropriate subspace by assuming that input views are generated from a latent view. This kind of approach can be seen as the prior combination of multiple views because multiple views are straightforwardly considered together to exploit the shared subspace.

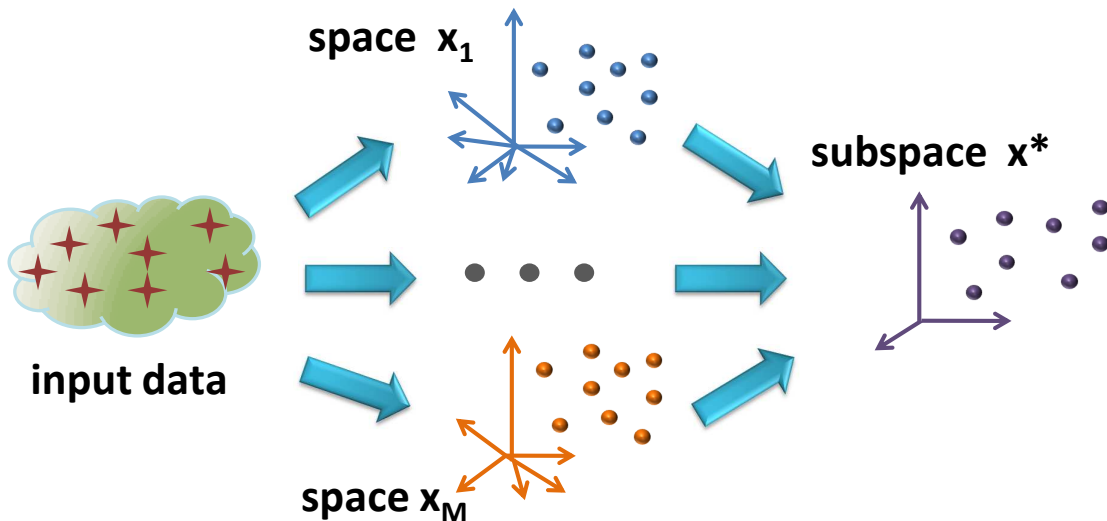


Figure 4: Sketch map of subspace learning for multi-view data.

## 5. Co-training Style Algorithms

Co-training (Blum and Mitchell, 1998) was one of the earliest schemes for multi-view learning. Since then, many variants have been developed. Besides the research on designing various algorithms, there are also a number of works on assumptions for co-training, which ensure the success of algorithms.

### 5.1 Assumptions for Co-training

Co-training considers a setting in which each example can be partitioned into two distinct views, and makes three main assumptions: (a) *Sufficiency*: each view is sufficient for classification on its own, (b) *Compatibility*: the target functions in both views predict the same labels for co-occurring features with high probability, and (c) *Conditional independence*: the views are conditionally independent given the class label. The conditional independence assumption plays a critical role, but it is usually too strong to be satisfied in practice and several weaker alternatives have thus been considered.

#### 5.1.1 CONDITIONAL INDEPENDENCE ASSUMPTION

Blum and Mitchell (1998) proved that when two sufficient views are conditionally independent given the class label, co-training can be successful. They gave a theorem that if the concept classes  $C_{1,2}$  on view  $X_{1,2}$  are learnable in the PAC model in spite of the classification noise, and if the conditional independence assumption is satisfied, then  $(C_1, C_2)$  is learnable in the co-training model from unlabeled data only, given an initial weakly-useful predictor  $h(x_1)$ . Specifically, let classification noise  $(\alpha, \beta)$  be a setting in which true positive examples are incorrectly labeled (independently) with probability  $\alpha$ , and the true negative examples are incorrectly labeled (independently) with probability  $\beta$ . Again define  $f(x)$  as the target concept and  $p = Pr_D(f(x) = 1)$  as the probability that a random example from

$D$  is positive. The sum of the two noise rates satisfies

$$\alpha + \beta < 1 - \varepsilon^2/(p(1 - p)), \quad (7)$$

with the probability at most  $1 - 4\varepsilon^2$ .

The above inequation gives some degree of justification for the co-training restriction on rule-based bootstrapping. However, it does not provide a bound on generalization error as a function of empirically measurable quantities; hence, based on the same conditional independence assumption, Dasgupta et al. (2002) gave the PAC style bounds for co-training. Let  $S$  be an i.i.d sample consisting of individual samples  $s_1, \dots, s_m$ . A partial rule  $h$  on a dataset  $X$  is a mapping from  $X$  to the label set  $\{1, \dots, k, \perp\}$ , where  $k$  is the number of class labels and  $\perp$  denotes the partial rule  $h$  gives no opinion. Then with probability at least  $1 - \delta$  over the choice of  $S$  we have the following for all pairs of rules  $h_1$  and  $h_2$ : if  $\gamma_i(h_1, h_2, \delta/2) > 0$  for  $1 \leq i \leq k$  then  $f$  is a permutation and for all  $1 \leq i \leq k$ ,

$$P(h_1 = i | f(y) = i, h_1 \neq \perp) \leq \frac{1}{\gamma_i(h_1, h_2, \delta)} (\epsilon_i(h_1, h_2, \delta) + \hat{P}(h_1 \neq i | h_2 = i, h_1 \neq \perp)),$$

where

$$\epsilon_i(h_1, h_2, \delta) = \sqrt{\frac{\ln 2(|h_1| + |h_2|) + \ln 2/\delta}{2|S(h_2 = i, h_1 \neq \perp)|}}$$

$$\gamma_i(h_1, h_2, \delta) = \hat{P}(h_1 = i | h_2 = i, h_1 \neq \perp) - \hat{P}(h_1 \neq i | h_2 = i, h_1 \neq \perp) - 2\epsilon_i(h_1, h_2, \delta). \quad (8)$$

Note that if the sample size is sufficiently large (relative to  $|h_1|$  and  $|h_2|$ ) then  $\epsilon_i(h_1, h_2, \delta)$  is near to zero. Also note that if  $h_1$  and  $h_2$  have near perfect agreement when neither is  $\perp$  then  $\gamma_i(h_1, h_2, \delta)$  is near one. The agreement between  $h_1$  and  $h_2$  upper bounds the error of  $h_1$ . The co-training algorithm therefore needs to maximize the agreement on unlabeled data between classifiers based on different views under the conditional dependence assumption to improve the accuracy of each hypothesis.

### 5.1.2 WEAK DEPENDENCE ASSUMPTION

The above-mentioned conditional independence assumption is overly strong to be satisfied for the two views in real applications. Abney (2002) relaxed the assumptions and found that weak dependence alone can lead to successful co-training. Given the mapping function  $Y = y$ , the conditional dependence of opposing-view rules  $h_1$  and  $h_2$  is defined as

$$d_y = \frac{1}{2} \sum_{u,v} |Pr[h_1 = v | Y = y, h_2 = u] - Pr[h_1 = v | Y = y]|.$$

If  $h_1$  and  $h_2$  are conditionally independent, then  $d_y = 0$ . The  $h_1$  and  $h_2$  satisfy weak rule dependence just in case:

$$d_y \leq p_2 \frac{q_1 - p_1}{2p_1 q_1},$$

where  $p_1 = \min_u Pr[h_2 = u | Y = y]$ ,  $p_2 = \min_u Pr[h_1 = u | Y = y]$ , and  $q_1 = 1 - p_1$ .

### 5.1.3 EXPANSION ASSUMPTION

Balcan et al. (2004) proposed a much weaker “expansion” assumption on the underlying data distribution and proved that it was sufficient for iterative co-training to succeed given appropriately strong PAC-learning algorithms on each feature set. Assume that examples are drawn from a distribution  $D$  over an instance space  $X$ . Let  $X^+$  and  $X^-$  denote the positive and negative regions of  $X$  respectively. For  $S_1 \subseteq X_1$  and  $S_2 \subseteq X_2$ , let  $\mathbf{S}_i (i = 1, 2)$  denote the event that an example  $\langle x_1, x_2 \rangle$  has  $x_i \in S_i$ . If we think of  $S_1$  and  $S_2$  as confident sets in each view, then  $Pr(\mathbf{S}_1 \wedge \mathbf{S}_2)$  denotes the probability mass on examples for which we are confident about both views, and  $Pr(\mathbf{S}_1 \oplus \mathbf{S}_2)$  denotes the probability mass on examples for which we are confident about just one view. Define that  $D^+$  is expanding if for any  $S_1 \subseteq X_1^+$  and  $S_2 \subseteq X_2^+$ ,

$$Pr(\mathbf{S}_1 \oplus \mathbf{S}_2) \geq \epsilon \min[Pr(\mathbf{S}_1 \wedge \mathbf{S}_2), Pr(\bar{\mathbf{S}}_1 \wedge \bar{\mathbf{S}}_2)]. \quad (9)$$

Another slightly stronger kind of expansion called “left-right expansion” can be defined as below.  $D^+$  is right-expanding if for any  $S_1 \subseteq X_1^+$  and  $S_2 \subseteq X_2^+$ , if

$$Pr(\mathbf{S}_1) \leq 1/2, Pr(\mathbf{S}_2 | \mathbf{S}_1) \geq 1 - \epsilon,$$

then

$$Pr(\mathbf{S}_2) \geq (1 + \epsilon)Pr(\mathbf{S}_1).$$

$D^+$  is left-expanding if above holds with indices 1 and 2 reversed.

It can clearly be seen that if  $S_i$  is the confident set in  $X_i^+$  and this set is small ( $Pr(\mathbf{S}_i) \leq 1/2$ ), a classifier, which learns from positive data on the conditional distribution that  $S_i$  induces over  $X_{3-i} (i = 1, 2)$ , is trained until it has error  $\leq \epsilon$  on that distribution. The definition implies that the confident set on  $X_{3-i}$  will have noticeably larger probability than  $S_i$ , so it is clear why this is useful for co-training.

### 5.1.4 LARGE DIVERSITY ASSUMPTION

Goldman and Zhou (2000) used two different supervised learning algorithms, and Zhou and Li (2005b) used two different parameter configurations of the same base learner for co-training style algorithms without redundant views, but neither of them had addressed the reasons of their successes. Afterwards, Wang and Zhou (2007) showed that when the diversity between the two learners is larger than their errors, the performance of the learner can be improved by co-training style algorithms. The difference  $d(h_i, h_j)$  between the two classifiers  $h_i$  and  $h_j$  implies the different biases between them, and the two classifiers will label some instances with different labels. If the examples labeled by the classifier  $h_i$  are to be useful for the classifier  $h_j$ ,  $h_i$  should know some information that  $h_j$  does not know. In other words,  $h_i$  and  $h_j$  should have significant differences. As the co-training process proceeds, the two classifiers will become increasingly similar and the difference between them will become smaller as the two classifiers label more and more unlabeled instances for each other. The co-training process would therefore not improve performance further after a number of learning rounds.

## 5.2 Co-training

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 three main assumptions for its success: sufficiency, compatibility, and conditional independence.

In the original co-training algorithm (Blum and Mitchell, 1998), given a set  $L$  of labeled examples and a set  $U$  of unlabeled examples, the algorithm first creates a smaller pool  $U'$  containing  $u$  unlabeled examples. It then iterates the following procedure. First, use  $L$  to train two naive Bayes classifiers  $h_1$  and  $h_2$  on the view  $x_1$  and  $x_2$  respectively. Second, allow each of these two classifiers to examine the unlabeled set  $U'$  and add the  $p$  examples it most confidently labels as positive, and  $n$  examples it most confidently labels as negative to  $L$ , along with the labels assigned by the corresponding classifier. Finally, the pool  $U'$  is replenished by drawing  $2p + 2n$  examples from  $U$  at random.

## 5.3 Co-EM

The intuition behind the co-training algorithm is that classifier  $h_1$  adds examples to the labeled set that classifier  $h_2$  will then be able to use for learning. If the conditional independence assumption holds, then on average each added example will be as informative as a random example and learning should progress, subject to adding many examples belonging to the wrong class. If the independence assumption is violated, then on average the added examples will be less informative and co-training may not be successful. Instead of committing labels for the unlabeled examples, we thus choose to run EM in each view and give unlabeled examples probabilistic labels that may change from one iteration to another. This is the principal idea of co-EM (Nigam and Ghani, 2000).

Co-EM outperforms co-training for many problems, but it requires the algorithm to process probabilistically labeled training data and the classifier to output class probabilities. Hence, the co-EM algorithm has only been studied with naive Bayes as the underlying learner, even though Support Vector Machine (SVM) is known to better fit the characteristics of many classification problems. By reformulating the SVM in a probabilistic way and estimating the labels of unlabeled data with probabilities, Brefeld and Scheffer (2004) successfully developed a co-EM version of SVM to close this gap.

## 5.4 Co-regularization

Suppose we have two hypothesis spaces,  $H^1$  and  $H^2$ , each of which contains a predictor that well-approximates the target function. In the case of co-training, these two are defined over different representations, or “views”, of the data, and trained alternately to maximize mutual agreement on unlabeled examples. More recently, several papers have formulated those intuitions as joint complexity regularization, or co-regularization (Sindhwani et al., 2005; Sindhwani and Rosenberg, 2008), between  $H^1$  and  $H^2$  which are taken to be Reproducing Kernel Hilbert Spaces (RKHSs) of functions defined on the input space  $X$ . Given a few labeled examples  $(x_i, y_i)_{i \in L}$  and a collection of unlabeled data  $\{x_i\}_{i \in U}$ , co-regularization learns a prediction function,

$$f_*(x) = \frac{1}{2}(f_*^1(x) + f_*^2(x)), \quad (10)$$

where  $f_*^1 \in H^1$  and  $f_*^2 \in H^2$  are obtained by solving the following optimization problem,

$$(f_*^1, f_*^2) = \min_{f^1 \in H^1, f^2 \in H^2} \gamma_1 \|f^1\|_{H^1}^2 + \gamma_2 \|f^2\|_{H^2}^2 + \mu \sum_{i \in U} [f^1(x_i) - f^2(x_i)]^2 + \sum_{i \in L} V(y_i, f(x_i)).$$

In this objective function, the first two terms measure complexity by RKHS norms  $\|\cdot\|_{H_1}^2$  and  $\|\cdot\|_{H_2}^2$  in  $H_1$  and  $H_2$  respectively, the third term enforces agreement among predictors on unlabeled examples, and the final term evaluates the empirical loss of the mean function  $f - (f^1 + f^2)/2$  on the labeled data with respect to a loss function  $V(\cdot, \cdot)$ . The real valued parameters  $\gamma_1$ ,  $\gamma_2$  and  $\mu$  allow different tradeoffs between the regularization terms.  $L$  and  $U$  are index sets over labeled and unlabeled examples respectively.

## 5.5 Co-regression

Most studies on multi-view and semi-supervised learning focus on classification problems, and regression problems can also be solved in a similar way. For instance, Zhou and Li (2005a) developed a co-training style semi-supervised regression algorithm called CoREG. This algorithm employs two k-nearest neighbor (kNN) regressors, each of which labels the unlabeled data for the other during the learning process. For the sake of choosing the appropriate unlabeled examples to label, CoREG estimates the labeling confidence by consulting the influence of the labeling of unlabeled examples on the labeled examples. The final prediction is made by averaging the regression estimates generated by both regressors. Inspired by the co-regularization algorithm, Brefeld et al. (2006) proposed a co-regression algorithm. Formally given  $M$  views, the training instances  $\{X_v\}_{v=1}^M$  with labels  $y(x) \in \mathbb{R}$ , and a finite set of instances  $Z \subseteq X$  for which the labels are unknown, we attempt to find  $f_1 : X \rightarrow \mathbb{R}, \dots, f_M : X \rightarrow \mathbb{R}$  that minimize

$$Q(f) = \sum_{v=1}^M \left[ \sum_{x \in X_v} V(y(x), f_v(x)) + \nu \|f_v(\cdot)\|^2 \right] + \lambda \sum_{u,v=1}^M \sum_{z \in Z} V(f_u(z), f_v(z)),$$

where the norms measuring complexity are in the respective Hilbert spaces,  $V(y(x), f_v(x))$  evaluates the losses between the predictors and target values of labeled examples, and  $V(f_u(z), f_v(z))$  imposes the agreement among predictors on unlabeled examples.

## 5.6 Co-clustering

The co-training algorithm was originally designed for semi-supervised learning, but the idea of co-training can also be applied in unsupervised and supervised learning settings. Under the assumption that the true underlying clustering will assign corresponding points in each view to the same cluster, several clustering techniques have been developed using the multi-view approach. Bickel and Scheffer (2004) studied a multi-view version of the most frequently used clustering approaches such as k-means, k-medoids, and EM. Taking k-means as an example: in each iteration, run k-means in one view, then interchange the partition information to another view and run k-means in the second view again. After termination, compute a consensus mean for each cluster and view, then assign each example to one distinct cluster that is determined through the closed concept vector. Considering that spectral clustering algorithms have good performance on arbitrary shaped clusters and a

well-defined mathematical framework, some methods are designed to utilize the idea of co-training to conduct spectral clustering (Kumar et al., 2010, 2011; Kumar and Daumé III, 2011). For example, Kumar and Daumé III (2011) developed a multi-view spectral clustering algorithm which solves spectral clustering on individual graphs to obtain the discriminative eigenvectors  $\mathbf{U}_1(\mathbf{U}_2)$  in each view, then clusters points using  $\mathbf{U}_1(\mathbf{U}_2)$  and uses this clustering to modify the graph structure in views 2(1) respectively. This process is repeated for a number of iterations.

### 5.7 Graph-based Co-training

Most co-training style algorithms focus on how to minimize the disagreement between two classifiers in order to obtain satisfactory performance of multi-view learners, thus these methods can be seen as disagreement-based approaches. Graph-based methods for co-training also exist; for instance, Yu et al. (2007, 2011) proposed a Bayesian undirected graphical model for co-training through Gaussian process (GP). Suppose we have  $m$  different views of  $n$  data examples  $\{x_i\}$  with outputs  $\{y_i\}$ . Let  $f_j$  denote the latent function for the  $j$ -th view, and let  $f_j \sim GP(0, \kappa)$  be its GP prior in view  $j$ . A latent function  $f_c$  is then introduced to ensure conditional independence between the output  $y$  and the  $m$  latent functions  $f_j$  for the  $m$  views. At the functional level, the output  $y$  depends only on  $f_c$ , and latent functions  $f_j$  depend on each other only via the consensus function  $f_c$ . That is, we have the joint probability:

$$p(y, f_c, f_1, \dots, f_m) = \frac{1}{Z} \psi(y, f_c) \prod_{j=1}^m \psi(f_j, f_c). \quad (11)$$

In the ground network with  $n$  data examples, let  $\mathbf{f}_c = \{f_c(x_i)\}_{i=1}^n$  and  $\mathbf{f}_j = \{f_j(x_i^j)\}_{i=1}^n$ . The graphical model leads to the following factorization:

$$p(\mathbf{y}, \mathbf{f}_c, \mathbf{f}_1, \dots, \mathbf{f}_m) = \frac{1}{Z} \prod_i \psi(y_i, f_c(x_i)) \prod_{j=1}^m \psi(\mathbf{f}_j) \psi(\mathbf{f}_j, \mathbf{f}_c). \quad (12)$$

Here, the within-view potential  $\psi(\mathbf{f}_j)$  specifies the dependency structure within each view  $j$ , and the consensus potential  $\psi(\mathbf{f}_j, \mathbf{f}_c)$  describes how the latent function in each view is related to the consensus function  $f_c$ . Employing a GP prior for each of the views, we can define the following potentials:

$$\psi(\mathbf{f}_j) = \exp\left(-\frac{1}{2} \mathbf{f}_j^T \mathbf{K}_j^{-1} \mathbf{f}_j\right), \quad \psi(\mathbf{f}_j, \mathbf{f}_c) = \exp\left(-\frac{\|\mathbf{f}_j - \mathbf{f}_c\|^2}{2\sigma_j^2}\right). \quad (13)$$

Integrating all the  $m$  latent functions in Eq. (12), we get the co-training kernel for multi-view learning as

$$\mathbf{K}_c = \left[ \sum_j (\mathbf{K}_j + \sigma_j^2 \mathbf{I})^{-1} \right]^{-1}. \quad (14)$$

This co-training kernel reveals a previously unclear insight into how the kernels from different views are combined in multi-view learning and allows us to solve GP classification simply.

Wang and Zhou (2010) treated the co-training process as a combinative propagation over two views and unified the graph- and disagreement-based semi-supervised learning into one framework. In one view, the labels can be propagated from the initial labeled examples to unlabeled examples, and these newly-labeled examples can be added into the other view. The other view can then propagate the labels of the initial labeled examples and these newly labeled examples to the remaining unlabeled instances. This process can be repeated until the stopping condition is met.

## 5.8 Multi-learner Algorithms

Goldman and Zhou (2000) presented a new “co-training” strategy for using unlabeled data to improve the performance of standard supervised learning algorithms. Without assuming that both of the views are sufficient for perfect classification, the only requirement of this co-training strategy is that its hypothesis partitions the example space into a set of equivalence classes. Assume that  $A$  and  $B$  are two different supervised algorithms,  $U$  are unlabeled data,  $L$  are the original labeled data,  $L_A$  are the data that  $B$  labeled for  $A$ , and  $L_B$  are the data  $A$  labeled for  $B$ . At the start of each iteration, train  $A$  on the labeled examples  $L \cup L_A$  to obtain the hypothesis  $H_A$ . Similarly, train  $B$  on  $L \cup L_B$  to obtain  $H_B$ . Each algorithm considers each of its equivalence classes and decides which to use to label data from  $U$  for the other algorithm. This co-training algorithm repeats until neither  $L_A$  nor  $L_B$  change during an iteration.

Zhou and Li (2005b) proposed another co-training style semi-supervised algorithm called tri-training, which does not require that the instance space be described with sufficient and redundant views, nor does it put any constraints on the supervised algorithm, as do Goldman and Zhou (2000). Tri-training generates three classifiers from the original labeled example set which are then refined using unlabeled examples in the iterations. For each iteration, an unlabeled example is labeled for a classifier if the other two classifiers agree on the labeling, under certain conditions.

The performance of traditional SVM-based relevance feedback approaches is often poor when the number of labeled feedback samples is small, thus Li et al. (2006) developed a new machine learning technique, namely multi-training SVM (MTSVM), to mitigate this problem. MTSVM combines the merits of the co-training technique and a random sampling method in the feature space. However, simply using the co-training algorithm with SVM is not realistic, because the co-training algorithm requires that the initial sub-classifiers have good generalization ability before the co-training procedure commences. Thus the authors employed classifier committee learning to enhance the generalization ability of each sub-classifier. Initially, a series of subsets of feature - in other words, multiple views of the data can be obtained from the original input feature using the random subspace method. Multiple classifiers can then be learned on these generated views and can train one another in a semi-supervised relevance feedback setting. Finally, the majority voting rule is used to generate the optimal classifier.

## 6. Multiple Kernel Learning

Multiple Kernel Learning (MKL) was originally developed to control the search space capacity of possible kernel matrices to achieve good generalization, but it has been widely applied to



problems involving multi-view data. This is because kernels in MKL naturally correspond to different views and combining kernels appropriately may improve learning performance. Gönen and Alpaydın (2011) have reviewed the literature on MKL. Since MKL can be regarded as just one part of multi-view learning, we place more weight on the connections between MKL and those parts; in this section, we illustrate the representative MKL algorithms and theoretical studies to present a complete picture in this survey.

### 6.1 Boosting Methods

Inspired by ensemble and boosting methods (Duffy and Helmbold, 2000; Friedman et al., 2001), Bennett et al. (2002) proposed the Multiple Additive Regression Kernels (MARK) algorithm which considers a large library of kernel matrices formed by different kernel functions and parameters. The decision function is modified as

$$f(x) = \sum_{i=1}^N \sum_{k=1}^M d_i^k K_k(x_i^m, x^m) + b, \quad (15)$$

which is composed of a linear combination of heterogeneous kernel functions  $K_1, \dots, K_M$ , and each kernel can be of any type; for example,  $\{K_k\}$  could be RBF kernels with different parameters. Like ensemble methods, each column of the kernel is treated as a hypothesis and the kernel columns are generated on the fly. Gradient-based ensemble algorithms, such as gradient boosting, can be adapted to this optimization problem.

Column Generation (CG) techniques have been widely used for solving large scale linear programs (LPs). Bi et al. (2004) used the 2-norm regularization approach to extend LP-Boost to a quadratic program (QP), so that many successful formulations, such as classic SVMs, ridge regression, etc, could benefit from CG techniques.

Crammer et al. (2002) used the boosting paradigm to perform the kernel construction process. Since numerous interpretations of AdaBoost and its variants regard the boosting process as a procedure that attempts to minimize classification error, the boosting methodology can be modified to work with kernels by rewriting the loss functions for a pair of examples  $(x_1, y_1)$  and  $(x_2, y_2)$  as

$$\begin{aligned} \text{ExpLoss}(K(x_1, x_2), y_1 y_2) &= \exp(-y_1 y_2 K(x_1, x_2)) \\ \text{LogLoss}(K(x_1, x_2), y_1 y_2) &= \log(1 + \exp(-y_1 y_2 K(x_1, x_2))). \end{aligned}$$

A pair of instances is viewed as a single example and pairs of the same labels are regarded as positively labeled examples, while pairs of opposite labels are seen as negatively labeled examples. Along similar lines to boost algorithms for classification, the combined kernel matrix can be updated iteratively using one of these two loss functions.

### 6.2 Semi-Definite Programming

The general form of Semi-Definite Programming (SDP) is

$$\begin{aligned} \min_x \quad & c^T x \\ \text{s.t.} \quad & F(x) = F_0 + x_1 F_1 + \dots + x_n F_n \geq 0 \\ & Ax = b, \end{aligned} \quad (16)$$

where  $x \in \mathbb{R}^p$  and  $F_i = F_i^T \in \mathbb{R}^{p \times p}$ . Note that the object is linear in the unknown  $x$ , and that both inequality and equality constraints are linear in  $x$ .

Lanckriet et al. (2002, 2004) showed how the kernel matrix can be learned from data via SDP techniques. In particular, if all labels of data are known, the task is to find the kernel matrix  $K$  which is maximally aligned with the set of labels  $y$ , and then this problem is formulated as

$$\begin{aligned} \max_{A, K} \quad & \langle K, yy^T \rangle \\ \text{s. t.} \quad & \text{trace}(A) \leq 1 \\ & \begin{pmatrix} A & K^T \\ K & I \end{pmatrix} \geq 0 \\ & K \geq 0. \end{aligned} \tag{17}$$

Given the labeled training set  $S_{n_{tr}} = \{(x_1, y_1), \dots, (x_{n_{tr}}, y_{n_{tr}})\}$  and the unlabeled test set  $T_{n_t} \{x_{n_{tr}+1}, \dots, x_{n_{tr}+n_t}\}$ , formally, we consider a kernel matrix has the form:

$$K = \begin{pmatrix} K_{tr} & K_{trt} \\ K_{trt}^T & K_t \end{pmatrix}, \tag{18}$$

where  $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$ ,  $i, j = 1, \dots, n_{tr}, n_{tr} + 1, n_{tr} + n_t$ . The goal is then to learn the optimal mixed block  $K_{trt}$  and the optimal ‘‘test data block’’  $K_t$  by optimizing a cost function over the ‘‘training data block’’  $K_{tr}$ . Under the constraint  $K = \sum_{i=1}^M \mu_i K_i$ , where the set  $\mathcal{K} = \{K_1, \dots, K_M\}$  is given and  $\mu_i$  are to be optimized, we can replace  $K$  with  $K_{tr}$  in Eq. (17) and obtain the SDP formulation for learning the kernel matrix.

### 6.3 Quadratically Constrained Quadratic Program (QCQP)

Bach et al. (2004) introduced a novel classification algorithm called support kernel machine (SKM). Given a decomposition of  $\mathbb{R}^k$  as a product of  $m$  blocks:  $\mathbb{R}^k = \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_m}$ , then each data  $x_i$  can be decomposed into  $m$  block components,  $x_i = \{x_{1i}, \dots, x_{mi}\}$ . The aim is to find a linear classifier,  $y = \text{sign}(w^T x + b)$ , where  $w = w_1, \dots, w_m$ . To obtain the sparsity of the vector  $w$  and make most of the components in  $w$  zero, the 1-norm and 2-norm are used to penalize  $w$ . Thus the primal problem can be formulated as follow:

$$\begin{aligned} \min \quad & \frac{1}{2} \left( \sum_{j=1}^m d_j \|w_j\|_2 \right)^2 + C \sum_{i=1}^n \xi_i \\ \text{w. r. t.} \quad & w \in \mathbb{R}^k = \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_m}, \quad \xi \in \mathbb{R}_+^n, \quad b \in \mathbb{R} \\ \text{s. t.} \quad & y_i \left( \sum_j w_j^T x_{ji} + b \right) \geq 1 - \xi_i, \quad \forall i \in \{1, \dots, n\}. \end{aligned} \tag{19}$$

This optimization problem can be seen as a second order cone program (SOCP) problem, and then the dual problem is given by:

$$\begin{aligned}
 \min \quad & \frac{1}{2}\gamma^2 - \alpha^T e & (20) \\
 \text{w.r.t.} \quad & \gamma \in \mathbb{R}, \alpha \in \mathbb{R}^n \\
 \text{s.t.} \quad & 0 \leq \alpha \leq C, \alpha^T y = 0 \\
 & \left\| \sum_i \alpha_i y_i x_{ji} \right\|_2 \leq d_j \gamma, \forall j \in \{1, \dots, m\},
 \end{aligned}$$

which is exactly equivalent to the QCQP formulation of Lanckriet et al. (2004). However, the advantage of this SOCP formulation is that Bach et al. (2004) developed an SMO algorithm for the SKM with Moreau-Yosida regularization, and transformed the primal problem as:

$$\begin{aligned}
 \min \quad & \frac{1}{2} \left( \sum_{j=1}^m d_j \|w_j\|_2 \right)^2 + \frac{1}{2} \sum_j a_j^2 \|w_j\|_2^2 + C \sum_{i=1}^n \xi_i & (21) \\
 \text{w.r.t.} \quad & w \in \mathbb{R}^k = \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_m}, \xi \in \mathbb{R}_+^n, b \in \mathbb{R} \\
 \text{s.t.} \quad & y_i \left( \sum_j w_j^T x_{ji} + b \right) \geq 1 - \xi_i, \forall i \in \{1, \dots, n\},
 \end{aligned}$$

where  $\{a_j\}$  are the MY-regularization parameters.

#### 6.4 Semi-infinite Linear Program (SILP)

Sonnenburg et al. (2006a,b) followed a different direction and formulated the problem as a semi-infinite linear program (SILP). Beginning with Eq. (20), the equivalent multiple kernel learning dual is modified as:

$$\begin{aligned}
 \min \quad & \gamma & (22) \\
 \text{w.r.t.} \quad & \gamma \in \mathbb{R}, \alpha \in \mathbb{R}^n \\
 \text{s.t.} \quad & 0 \leq \alpha \leq C, \alpha^T y = 0 \\
 \forall k \quad & \underbrace{\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K_k(x_i^k, x_j^k) - \sum_{i=1}^N \alpha_i}_{S_k(\alpha)} \leq \gamma,
 \end{aligned}$$

which may be solved by

$$L = \gamma + \sum_{k=1}^M \beta_k (S_k(\alpha) - \gamma) \quad (23)$$

minimized w.r.t  $\alpha$  and maximized w.r.t  $\beta$ . Setting the derivative w.r.t. to  $\gamma$  to zero, the constraint  $\sum_k \beta_k = 1$  is obtained. Eq. (23) can then be simplified to a min-max problem

$$\begin{aligned} \max_{\beta} \min_{\alpha} \quad & \sum_{k=1}^M \beta_k S_k(\alpha) \\ \text{s. t.} \quad & \sum_{i=1} \alpha_i y_i = 0 \quad \sum_{k=1}^M \beta_k = 1. \end{aligned} \tag{24}$$

Assume that  $\alpha^*$  is the optimal solution, and given the definition of  $\theta = L = S(\alpha^*, \beta)$ , Eq. (23) is equivalent to the following SILP problem:

$$\begin{aligned} \max \quad & \theta \\ \text{s. t.} \quad & \sum_k \beta_k = 1, \quad \sum_k \beta_k S_k(\alpha) \geq \theta \\ \forall \alpha \quad & 0 \leq \alpha \leq C, \quad \sum_i y_i \alpha_i = 0, \end{aligned} \tag{25}$$

where  $\theta$  and  $\beta$  are only linearly constrained, but there are a large number of constraints due to the possible values of  $\alpha$ .

Compared to the SDP and QCQP, the SILP formulation has a lower computational complexity, and this SILP problem can be efficiently solved using an off-the-shelf LP solver and a standard SVM implementation. Thus it allows us to efficiently handle more than a hundred thousand examples or several hundred kernels.

## 6.5 Simple MKL

Rakotomamonjy et al. (2007, 2008) departed from the framework proposed by Bach et al. (2004) and presented a different primal problem for multiple kernel learning through an adaptive 2-norm regularization formulation. Inspired by the multiple smoothing splines framework (Wahba, 1990), the proposed primal formulation is

$$\begin{aligned} \min \quad & \sum_k \frac{1}{d_k} \|w_k\|^2 + C \sum_i \xi_i \\ \text{s. t.} \quad & y_i (\sum_k w_k x_i^k + b) \geq 1 - \xi_i \\ & \sum_k d_k = 1 \\ & \xi_i \geq 0, d_k \geq 0 \forall i, \forall k. \end{aligned} \tag{26}$$

Note that the  $d_k$  controls the smoothness of kernel function, and the 1-norm constraint on the vector  $d$  will lead to a sparse decision function with few basis kernels. By Defining the

optimal SVM objective value  $J(d)$  as

$$\begin{aligned} \min \quad & J(d) = \sum_k \frac{1}{d_k} \|w_k\|^2 + C \sum_i \xi_i \\ \text{s. t.} \quad & y_i \left( \sum_k w_k x_i^k + b \right) \geq 1 - \xi_i \\ & \xi_i \geq 0, \end{aligned} \tag{27}$$

the primal optimization problem can then be reformulated as

$$\min_d J(d) \text{ s. t. } \sum_k d_k = 1, d_k \geq 0. \tag{28}$$

The overall procedure to solve this problem consists of two steps: first, solving a canonical SVM optimization problem  $J(d)$  with given  $d$ ; second, updating  $d$  by gradient descent while ensuring that the constraints on  $d$  are satisfied. This novel multiple kernel learning framework is called simple MKL, which has been shown to be more efficient than the SILP problem.

Chapelle and Rakotomamonjy (2008) investigated the use of second order optimization approaches to solve the MKL problem, and propose hessian MKL as an extension of simple MKL. In each iteration, hessian MKL updates the kernel weights using a Newton step found by minimizing a QP problem. The result shows that hessian MKL outperforms simple MKL in terms of computational efficiency.

The SILP approach often suffers from slow convergence because it updates kernel weights based only on the cutting plane model. The simple MKL is efficient; however, it does not use the gradients computed in previous iterations, which can be useful in improving the efficiency of the search. Xu et al. (2009a) extended the level method, and applied it to multiple kernel learning to overcome the drawbacks of SILP (Sonnenburg et al., 2006b) and simple MKL (Rakotomamonjy et al., 2007). Following the SILP method, this algorithm has an extra step to adjust the solution for kernel weights obtained from a cutting plan model, through a projection to a level set. This adjustment ensures the new solution is close to the current solution and reduces the objective function.

## 6.6 Group-LASSO Approaches

It is reasonable to consider the group structure between the combined kernels when the kernels can be partitioned into groups which correspond to subsets of inputs or sources. In the learning process, it is desirable to suppress the kernels or groups that are irrelevant for the classification task, otherwise all the kernels belonging to the same groups which are relevant to the task will be selected. Based on this idea, Szafranski et al. (2008, 2010) developed the Composite Kernel Learning (CKL) approach, which extends the multiple kernel learning problem to take into account the group structure among kernels and constructs the relationship with group-LASSO (Yuan and Lin, 2006). The MKL formulation

of Rakotomamonjy et al. (2008) is modified to obtain the following formulation of CKL:

$$\begin{aligned}
\min \quad & \frac{1}{2} \sum_k \frac{1}{d_k} \|w_k\|^2 + C \sum_i \xi_i & (29) \\
\text{s. t.} \quad & y_i \left( \sum_k w_k x_i^k + b \right) \geq 1 - \xi_i \\
& \xi_i \geq 0 \\
& \sum_G \left( \|G\|^p \left( \sum_{k \in G} d_k^{1/q} \right)^q \right)^{1/(p+q)} \leq 1 \\
& d_k \geq 0,
\end{aligned}$$

where  $p$  and  $q$  are set according to the problem at hand,  $G$  denotes one subset of kernels, and  $\|G\|$  is the size of group  $G$ . Note the third constraint: in particular cases where  $p = 0, q = 1$ , a LASSO type penalty is imposed on the RKHS norms, and when  $p = 1, q = 0$ , a group-LASSO type penalty is imposed on the RKHS norms.

Xu et al. (2010) discussed the connection between multiple kernel learning and the group-LASSO regularizer, and generalized MKL to  $L_p$ -MKL which constrains the  $p$ -norm kernel weights. This proposed algorithm provides a unified solution for the entire family of  $L_p$  models, besides which the kernel weights can be calculated by a closed-form formulation without dependence on other commercial software. Subrahmanya and Shin (2010) proposed an algorithm called Sparse Multiple Kernel Learning (SMKL), which generalizes group-feature selection to kernel selection by introducing a log-based penalty over the groups. This method can automatically select the optimal number of sources from a large candidate list with a sparser solution compared to the existing multiple kernel learning framework.

## 6.7 Bounds for Learning Kernels

The most common family of kernels examined in multiple kernel learning is that of non-negative or convex combination of some fixed kernels constrained by a trace condition, which can be viewed as an  $L_1$  or  $L_2$  regularization, or  $L_p$  regularization with other values of  $p$ .

Lanckriet et al. (2004) showed that when a kernel is chosen from a convex combination of  $k$  base kernels, the estimation error of the learned classifier is bounded by  $O(\sqrt{\frac{k/\gamma^2}{n}})$ , where  $\gamma$  is the margin of the learned classifier under the kernel. This bound converges and can be viewed as the first informative generalization bound for this family of kernels; however, the multiplicative interaction between the margin complexity term  $1/\gamma^2$  and the number of base kernels  $k$  does not encourage the use of too many base kernels. It suggests that learning even a few kernel parameters leads to a multiplicative increase in the required sample size. Srebro and Ben-David (2006) presented a generalization bound for a kernel family with pseudo-dimension of  $d_\phi$ . The pseudo-dimension of most kernel families is similar to our intuitive notion of the dimensionality of the family; in particular, the pseudo-dimension of a family of linear or convex combinations of  $k$  base kernels is at most  $k$ . The estimation error for SVMs with margin  $\gamma$  is bounded by  $\sqrt{O(d_\phi + 1/\gamma^2)/n}$ , which establishes that the

bound on the required sample size,  $O(d_\phi + 1/\gamma^2)$  grows only additive with the dimensionality of the allowed kernel family. Ying and Campbell (2009) showed that the generalization analysis of the regularized kernel learning system reduces to investigation of the suprema of the Rademacher chaos process of order two over candidate kernels, and they used metric entropy integrals and the pseudo-dimension of the set of candidate kernels to estimate the empirical Rademacher chaos complexity. For a pseudo-dimension of  $k$ , as in the case of a convex combination of  $k$  base kernels, their bound is in  $O(\sqrt{k(R^2/\rho^2)(\log(m)/m)})$  and is thus multiplicative in  $k$ . Based on a combinatorial analysis of the Rademacher complexity of the hypothesis set under consideration, Cortes et al. (2010) presented another generalization bound with an  $L_1$  constraint that has only a logarithmic dependency on the kernel number  $k$ . The bound is in  $O(\sqrt{\frac{(\log k)R^2/\rho^2}{m}})$ , thus it is valid for a very large number of kernels, in particular for  $k \gg m$ , and it contains only a  $\sqrt{\log k}$  dependency on the number of kernels, which is tight and considerably more favorable. Assuming the different views corresponding to the different kernels to be uncorrelated, Kloft and Blanchard (2011) derived an upper bound on the local Rademacher complexity of  $L_p$ -norm multiple kernel learning. Given the number of kernels  $M$  and the radius  $D$ , the bound for centered identical independent kernels is of the order  $O(\sqrt{\sum_j^\infty \min(rM, D^2 M^{\frac{2}{p^*}} \lambda_j)})$ . From the upper bound, a tighter excess risk bound than previous approaches is obtained, which achieves a fast convergence rate of the order  $O(n^{-\frac{\alpha}{1+\alpha}})$ , where  $\alpha$  is the minimum eigenvalue decay rate of the individual kernels.

## 7. Subspace Learning-based Approaches

Subspace learning-based approaches aim to obtain a latent subspace shared by multiple views by assuming that the input views are generated from this subspace. Besides the well known canonical correlation analysis (CCA), other more effective methods to construct the subspaces have recently become available.

### 7.1 Algorithms based on CCA

Canonical correlation analysis (CCA) is a technique for modeling the relationships between two (or more) sets of variables, and it has been applied with great success on a variety of learning problems dealing with multi-view data.

#### 7.1.1 A REVIEW OF CCA

For  $X \in \mathbb{R}^{D_1 \times N}$  and  $Y \in \mathbb{R}^{D_2 \times N}$ , CCA computes two projection vectors,  $w_x \in \mathbb{R}^{D_1}$  and  $w_y \in \mathbb{R}^{D_2}$ , such that the following correlation coefficient:

$$\rho = \frac{w_x^T X Y^T w_y}{\sqrt{(w_x^T X X^T w_x)(w_y^T Y Y^T w_y)}} \quad (30)$$

is maximized. Since  $\rho$  is invariant to the scaling of  $w_x$  and  $w_y$ , CCA can be formulated equivalently as

$$\begin{aligned} \max_{w_x, w_y} & \quad w_x^T X Y^T w_y \\ \text{s. t.} & \quad w_x^T X X^T w_x = 1, \quad w_y^T Y Y^T w_y = 1. \end{aligned} \quad (31)$$

Assuming  $YY^T$  is nonsingular, then  $w_x$  can be obtained by solving the following optimization problem:

$$\begin{aligned} \max_{w_x, w_y} \quad & w_x^T XY^T (YY^T)^{-1} YX^T w_y \\ \text{s. t.} \quad & w_x^T XX^T w_x = 1 \end{aligned} \quad (32)$$

Both formulations in Eqs. (31) and (32) attempt to find the eigenvectors corresponding to the top eigenvalues of the following generalized eigenvalue problem:

$$XY^T (YY^T)^{-1} YX^T w_x = \eta XX^T w_x, \quad (33)$$

where  $\eta$  is the eigenvalue corresponding to the eigenvector  $w_x$ .

### 7.1.2 KERNEL CCA

Canonical correlation analysis (CCA) is a linear feature extraction algorithm, but for many real world datasets exhibiting non-linearities, it is impossible for a linear projection to capture the properties of the data. Kernel methods provide a way to deal with the non-linearities by mapping the data to a higher dimensional space and then applying linear methods in that space.

Formally given a pair of datasets  $X \in \mathbb{R}^{D_1 \times N}$  and  $Y \in \mathbb{R}^{D_2 \times N}$ , CCA seeks to find linear projections  $w_x \in \mathbb{R}^{D_1}$  and  $w_y \in \mathbb{R}^{D_2}$  such that, after projecting, the corresponding examples in the two datasets are maximally correlated in the projected space. To obtain the kernel formulation of CCA, dual representation is engaged by expressing the projection direction as  $w_x = X\alpha$  and  $w_y = Y\beta$  where  $\alpha$  and  $\beta$  are vectors of size  $N$ . In the dual formulation, the correlation coefficient between  $X$  and  $Y$  can be written as:

$$\rho = \max_{\alpha, \beta} \frac{\alpha^T X^T XY^T Y \beta}{\sqrt{\alpha^T X^T XX^T X \alpha \times \beta^T Y^T YY^T Y \beta}} \quad (34)$$

Now using the fact that  $K_x = X^T X$  and  $K_y = Y^T Y$  are the kernel matrices for  $X$  and  $Y$ , kernel CCA amounts to solving the following problem:

$$\begin{aligned} \max_{\alpha, \beta} \quad & \frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T K_x^2 \alpha \times \beta^T K_y^2 \beta}} \\ \text{s. t.} \quad & \alpha^T K_x^2 \alpha = 1, \quad \beta^T K_y^2 \beta = 1. \end{aligned} \quad (35)$$

KCCA works by using the kernel matrices  $K_x$  and  $K_y$  of the examples in the two views  $X$  and  $Y$  of the data. In contrast to linear CCA, which works by carrying out an eigen-decomposition of the covariance matrix, the eigenvalue problem for KCCA is given by:

$$\begin{pmatrix} 0 & K_x K_y \\ K_y K_x & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \lambda \begin{pmatrix} K_x^2 & 0 \\ 0 & K_y^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \quad (36)$$

For the case of a linear kernel, KCCA reduces to the standard CCA.

KCCA can isolate feature space directions that correlate between the two views and might be expected to represent common relevant information; therefore, experiments have



shown that KCCA could be an effective preprocessing step to improve the performance of classification algorithms such as Support Vector Machine (SVM). Combining KCCA with SVM into a single optimization, Farquhar et al. (2005) proposed a method called SVM-2K, which can be seen as the global optimization of two distinct SVMs, one in each of the two feature spaces. Slightly different from the 2-norm that characterizes KCCA, SVM-2K takes an  $\epsilon$ -insensitive 1-norm using slack variable to measure the amount by which points fail to meet  $\epsilon$  similarity:

$$|\langle \mathbf{w}_A, \phi_A(\mathbf{x}_i) \rangle + b_A - \langle \mathbf{w}_B, \phi_B(\mathbf{x}_i) \rangle - b_B| \leq \eta_i + \epsilon,$$

where  $\mathbf{w}_A, b_A$  and  $\mathbf{w}_B, b_B$  are the weight and bias of the first and second SVM respectively. Then with the usual 1-norm SVM constraints, the objective problem can be written as:

$$\begin{aligned} \min \quad L &= \frac{1}{2} \|\mathbf{w}_A\|^2 + \|\mathbf{w}_B\|^2 & (37) \\ &+ C^A \sum_i \xi_i^A + C^B \sum_i \xi_i^B + D \sum_i \eta_i \\ \text{s. t.} \quad &|\langle \mathbf{w}_A, \phi_A(\mathbf{x}_i) \rangle + b_A - \langle \mathbf{w}_B, \phi_B(\mathbf{x}_i) \rangle - b_B| \leq \eta_i + \epsilon \\ &y_i(|\langle \mathbf{w}_A, \phi_A(\mathbf{x}_i) \rangle + b_A| \geq 1 - \xi_i^A) \\ &y_i(|\langle \mathbf{w}_B, \phi_B(\mathbf{x}_i) \rangle + b_B| \geq 1 - \xi_i^B) \\ &\xi_i^A \geq 0, \quad \xi_i^B \geq 0, \quad \eta_i \geq 0 \quad \forall i. \end{aligned}$$

The final decision function is

$$f(x) = \frac{1}{2}(f_A(x) + f_B(x)). \quad (38)$$

### 7.1.3 THEORETICAL ANALYSIS OF CCA

Canonical correlation analysis (CCA) can be viewed as finding basis vectors for two sets of variables such that the correlations between the projections onto these basis vectors  $x_a = w_a^T \phi_a(x)$  and  $y_b = w_b^T \phi_b(y)$  are mutually maximized. KCCA uses the kernel trick to produce a non-linear version of CCA, by looking for functions  $f \in H_x$  and  $g \in H_y$  such that the random variables  $f(x)$  and  $g(y)$  have maximal correlation. This leads to the kernelised form, KCCA

$$\max \frac{Cov[f(x), g(y)]}{\sqrt{Var[f(x)]^{1/2} Var[g(y)]^{1/2}}}. \quad (39)$$

In practice, we have to estimate the desired function from a finite sample, thus an empirical estimate of Eq. (39) is

$$\max \frac{\widehat{Cov}[f(x), g(y)]}{(\widehat{Var}[f(x)]^{1/2} + \epsilon_n \|f\|_{H_x}^2)(\widehat{Var}[g(y)]^{1/2} + \epsilon_n \|g\|_{H_y}^2)}, \quad (40)$$

where  $\epsilon_n$  is the regularization coefficient and  $n$  is the number of examples. Fukumizu et al. (2007) investigated the general problem of establishing a consistency of KCCA by providing the rates for the regularization parameter, and proved that when

$$\lim_{n \rightarrow \infty} \epsilon_n = 0, \quad \lim_{n \rightarrow \infty} \frac{n^{-1/3}}{\epsilon_n} = 0,$$

for the decay of the regularization coefficient  $\varepsilon_n$ , the convergence in the  $L_2$  norm for kernel CCA is ensured.

Hardoon and Shawe-Taylor (2009) proposed a finite sample statistical analysis of KCCA by using a regression formulation. By computing the empirical expected value of  $g_{a,b}(x, y) := \widehat{E}[\|W_a^T \phi_a(x) - W_b^T \phi_b(y)\|^2]$ , the error bound on new data can be obtained by using Rademacher complexity. Formally, given a paired training set  $S = \{(x_i, y_i)\}$  of size  $\mathcal{L}$  in the feature space defined by the bounded kernels  $k_a$  and  $k_b$  drawn i.i.d according to a distribution  $\mathcal{D}$ , then with probability greater than  $1 - \delta$  over the generation of  $S$ , the expected value of  $g_{a,b}(x, y)$  on new data is bounded by

$$\begin{aligned} E_D[g_{a,b}] &\leq \widehat{E}_D[g_{a,b}] + 3RA\sqrt{\frac{\ln 2/\delta}{2\mathcal{L}}} \\ &+ 4A\frac{1}{\mathcal{L}}\sqrt{\sum_{i=1}^{\mathcal{L}}(k_a(x_i, x_i) + k_b(y_i, y_i))^2}, \end{aligned} \quad (41)$$

where

$$\begin{aligned} R &= \max_{x \in \mathcal{D}}(k_a(x, x) + k_b(y, y)) \\ \|W_a^T W_a + W_b^T W_b\|^2 &\leq A. \end{aligned}$$

This suggests the regularization of KCCA because it shows that the quality of the generalization of the associated pattern function is controlled by the sum of the squares of the norms of the weight vectors.

Cai and Sun (2011) gave a convergence rate analysis of kernel CCA. Assuming  $(\mathcal{H}_x, \mathcal{H}_y)$  are RKHS of functions on  $\mathcal{X}$  and  $\mathcal{Y}$  respectively,  $V_{YX}$  is a compact operation from  $\mathcal{H}_x$  to  $\mathcal{H}_y$ , and there exist operators  $W_l, W_r$  such that

$$V_{YX} = W_l \Sigma_{XX}^p \text{ and } V_{XY} = \Sigma_{YY}^p W_r,$$

where  $\Sigma_{XX}^p$  and  $\Sigma_{YY}^p$  are covariance operators. Taken  $\varepsilon_n = \varepsilon_1 n^{-\alpha}$  with  $0 < \alpha < 1/3$ , then with probability at least  $1 - \delta$ , we have

$$\|\Sigma_{XX}^{1/2}(\widehat{f}_n - \widehat{f})\|_{H_x}^2 \leq C_{6,\delta} n^{-\theta}, \quad \|\Sigma_{YY}^{1/2}(\widehat{g}_n - \widehat{g})\|_{H_y}^2 \leq C_{6,\delta} n^{-\theta}, \quad (42)$$

where  $\theta = \min\{1 - 3\alpha, 2p\alpha, \alpha\}$  and  $C_{6,\delta}$  is a constant independent of  $n$ . So when  $0 < p \leq \frac{1}{2}$ , the convergence rate is  $\min\{1 - 3\alpha, 2p\alpha, \alpha\}$ .

#### 7.1.4 RELATED ALGORITHMS WITH CCA

CCA has been widely studied in different fields as a general tool for conducting multi-view dimensional reduction. Recently many new algorithms based on CCA have been proposed to extend the original CCA in different applications.

One popular use of CCA is for supervised learning, in which one view is derived from the data and another view is derived from the class labels. In this setting, the data can be projected into a lower-dimensional space directed by the label information (Yu et al., 2006). However, this algorithm does not actually use the multiple views of the data; it is just a single view approach along with the label information. Sharma et al. (2012) proposed a

Generalized Multi-view Analysis (GMA) which exploits the fact that most popular supervised and unsupervised feature extraction techniques are the solution of a special form of quadratically constrained quadratic program. This algorithm can be seen as a supervised extension of CCA and has the potential to replace CCA whenever classification or retrieval is the purpose and label information is available.

Chaudhuri et al. (2009) exploited CCA to project the data to the subspace spanned by the means, and then applied standard clustering algorithms to this subspace. This subspace is valuable for the subsequent clustering, because, when projected onto this subspace, the means of the distributions are well-separated, yet the typical distance between points from the same distributions is smaller than in the original space. Both traditional CCA and KCCA assume that features across all views are available for examples, but this may not be the case with many multi-view datasets. To apply multi-view clustering on such datasets, Anusua Trivedi (2010) found a way to deal with the lack of data in the incomplete views with an idea from Laplacian regularization. Given the known part of  $K$ , the missing parts of kernel matrix  $K$  can be found by solving an optimization problem; following construction of the full kernel, standard algorithms can conduct the subsequent tasks.

In semi-supervised learning, a number of labeled examples are usually required for training an initial weakly useful predictor which is in turn used to exploit the unlabeled examples. By taking advantage of the correlations between the views using CCA, Zhou et al. (2007) proposed a method which can perform semi-supervised learning with only one labeled training example. With the help of CCA, the similarity between an original unlabeled instance and the original labeled instance can be measured. Thus, several unlabeled examples with highest and lowest similarity scores can be selected as the extra positive and negative examples, respectively. As the number of labeled training examples is increased, the traditional semi-supervised learning algorithm can be performed.

Wang et al. (2008) developed a novel multiple kernel learning algorithm, combined with CCA. Initially the input data is mapped into  $m$  different feature spaces by  $m$  different kernels, where each generated feature space is taken as one view of the input data. Borrowing the motivating argument from CCA that  $m$  views in the transformed coordinates can be maximally correlated, the generalization of classifiers can be improved. Combining CCA with PCA, Zhu et al. (2012) suggested a novel method called MKCCA to implement dimensionality reduction. MKCCA improves the kernel CCA by performing PCA followed by CCA to better remove noises and handle the issue of trivial learning. Furthermore, comparing CCA with least squares for regression and classification, Sun et al. (2008) formulated CCA in multi-label classification as a least square problem.

## 7.2 Multi-view Fisher Discriminant Analysis

In contrast to CCA, which ignores label information, Diethe et al. (2008) generalized Fisher's discriminant analysis to find informative projections for multi-view data in a supervised setting.

### 7.2.1 TWO VIEW FISHER DISCRIMINANT ANALYSIS

Given examples drawn from two views of the same underlying semantic object, denoted as  $X_a$  and  $X_b$  respectively, the two view Fisher discriminant chooses two sets of weights  $w_a$

and  $w_b$  to solve the following optimization problem

$$\rho = \frac{w_a^T X_a^T y y^T X_b^T w_b}{\sqrt{(w_a^T X_a^T B X_a w_a + \mu \|w_a\|^2) \cdot (w_b^T X_b^T B X_b w_b + \mu \|w_b\|^2)}},$$

where  $w_a$  and  $w_b$  are the weight vectors for each view. Since the equation is not affected by rescaling of  $w_a$  or  $w_b$ , the optimization can be subjected to the following constraints

$$\begin{aligned} w_a^T X_a^T B X_a w_a + \mu \|w_a\|^2 &= 1, \\ w_b^T X_b^T B X_b w_b + \mu \|w_b\|^2 &= 1. \end{aligned}$$

The corresponding Lagrangian for this optimization can be written as

$$L = w_a^T X_a^T y y^T X_b^T w_b - \frac{\lambda_a}{2} (w_a^T X_a^T B X_a w_a + \mu \|w_a\|^2 - 1) - \frac{\lambda_b}{2} (w_b^T X_b^T B X_b w_b + \mu \|w_b\|^2 - 1),$$

which can be solved by differentiating with respect to the weight vectors  $w_a$  and  $w_b$ .

### 7.2.2 KERNEL TWO VIEW FISHER DISCRIMINANT ANALYSIS

By introducing two dual weight vectors  $w_a = X_a^T \alpha$  and  $w_b = X_b^T \beta$ , we have

$$\rho = \frac{\alpha X_a X_a^T y y^T X_b^T X_b^T \beta}{\sqrt{(\alpha X_a X_a^T B X_a X_a^T \alpha + \kappa \|w_a\|^2) \cdot (\beta X_b X_b^T B X_b X_b^T \beta + \kappa \|w_b\|^2)}},$$

and Its kernel form

$$\rho = \frac{\alpha K_a y y^T K_b \beta}{\sqrt{(\alpha K_a B K_a \alpha + \kappa \|w_a\|^2) \cdot (\beta K_b B K_b \beta + \kappa \|w_b\|^2)}}.$$

Given the constraints

$$\begin{aligned} \alpha K_a B K_a \alpha + \kappa \alpha K_a \alpha &= 1, \\ \beta K_b B K_b \beta + \kappa \beta K_b \beta &= 1, \end{aligned}$$

the corresponding Lagrangian for this optimization can be written as

$$L = \alpha K_a y y^T K_b \beta - \frac{\lambda_a}{2} (\alpha K_a B K_a \alpha + \kappa \alpha K_a \alpha - 1) - \frac{\lambda_b}{2} (\beta K_b B K_b \beta + \kappa \beta K_b \beta - 1).$$

Differentiating with respect to the weight vectors  $\alpha$  and  $\beta$ , the above problem can then be solved.

## 7.3 Multi-view Embedding

Since high dimensionality, i.e. a large amount of input features, may lead to a large variance of estimates, noise, over-fitting, and in general, higher complexity and inefficiency in the learners, it is necessary to conduct dimensional reduction and generate low-dimensional representations for these features. When faced with multiple features, however, performing

a dimensional reduction for each feature is not an ideal solution, considering the underlying connections between them. Thus it may be necessary to resort to advanced methods to conduct embedding for multiple features simultaneously and to output a meaningful low-dimensional embedding shared by all features.

Existing spectral embedding algorithms assume that samples are drawn from a vector space and thus cannot deal straightforwardly with multi-view data. Xia et al. (2010) developed a new spectral embedding algorithm, namely, multi-view spectral embedding (MSE), which encodes multi-view features to achieve a physically meaningful embedding. Based on their previous work of patch alignment (Zhang et al., 2009), MSE can be described as follows. MSE first builds a patch for a sample on a view, then given the patches from different views, part optimization is performed to obtain the optimal low-dimensional embedding for each view. All low-dimensional embeddings from different patches are then unified into one whole by global coordinate alignment. More formally, given the  $i$ -th view  $X^i = [x_1^i, \dots, x_n^i]$ , consider an arbitrary point  $x_j^i$  and its  $k$  nearest neighbors,  $x_j^i$  is defined as  $X_j^i = [x_j^i, x_{j_1}^i, \dots, x_{j_k}^i]$ . For  $X_j^i$ , we want to find a part mapping  $f_j^i : X_j^i \rightarrow Y_j^i$ , where  $Y_j^i = [y_j^i, y_{j_1}^i, \dots, y_{j_k}^i]$ . The part optimization for the  $j$ -th patch on the  $i$ -th view is defined as

$$\min_{Y_j^i} \sum_{l=1}^k \|y_j^i - y_{j_l}^i\|^2 (w_j^i)_l, \quad (43)$$

where  $(w_j^i)_l = \exp(-\|x_j^i - x_{j_l}^i\|^2/t)$ . Eq. (43) can be reformulated to

$$\min_{Y_j^i} \text{tr}(Y_j^i L_j^i (Y_j^i)^T), \quad (44)$$

where  $\text{tr}(\cdot)$  is the trace operator and  $L_j^i$  encodes the objective function for the  $j$ -th patch on the  $i$ -th view. To explore the complementary property of multiple views, a set of non-negative weights  $\alpha = [\alpha_1, \dots, \alpha_m]$  is imposed on part optimizations, thus the multi-view part optimization for the  $j$ -th patch is

$$\min_{\{Y_j^i\}_{i=1}^m} \sum_{i=1}^m \alpha_i \text{tr}(Y_j^i L_j^i (Y_j^i)^T). \quad (45)$$

To ensure that low dimensional embeddings in different views are globally consistent with each one another, assume that the coordinate for  $Y_j^i = [y_j^i, y_{j_1}^i, \dots, y_{j_k}^i]$  is selected from the global coordinate  $Y = [y_1, \dots, y_n]$ , which then gives  $Y_j^i = Y S_j^i$ , where  $S_j^i$  is the selection matrix for encoding the relationships of samples in a patch in the original high dimensional space. By summing over all part optimizations, the global coordinate alignment can be written as

$$\min \sum_{j=1}^n \sum_{i=1}^m \alpha_i \text{tr}(Y S_j^i L_j^i (S_j^i)^T Y^T). \quad (46)$$

From Eq. (46), the alignment matrix for the  $i$ -th view can be written as

$$L^i = \sum_{j=1}^n S_j^i L_j^i (S_j^i)^T. \quad (47)$$

To make sure that each view makes a particular contribution to the final low dimensional embedding, and considering some constraints on the variants, the final objective function is defined as

$$\begin{aligned} \min_{Y, \alpha} \quad & \sum_{i=1}^m \alpha_i^\gamma \text{tr}(Y L^i Y_T) \\ \text{s. t.} \quad & Y Y^T = I, \sum_{i=1}^m \alpha_i = 1, \alpha_i \geq 0, \gamma > 1. \end{aligned} \quad (48)$$

Finally, MSE can generate a low dimensional sufficiently smooth embedding by preserving the locality of each view simultaneously.

The main idea of Stochastic Neighbor Embedding (SNE) is to construct probability distributions from pair wise distances wherein larger distances correspond to smaller probabilities and vice versa. Formally, suppose we have high-dimensional data points  $\{x_i\}_{i=1}^n$ , the joint probability distribution over sample pairs can be represented in a symmetric matrix  $P \in \mathbb{R}^{n \times n}$ , where  $p_{ii} = 0$  and  $\sum_{i,j} p_{ij} = 1$ . Let  $y_i$  be the low dimensional data corresponding to  $x_i$ , then the probability distribution  $Q$  in low dimensional embedding is defined as

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}. \quad (49)$$

This embedding can be acquired by minimizing the KL divergence of the two probability distributions,

$$\text{KL}(P|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \quad (50)$$

Xie et al. (2011) proposed the m-SNE algorithm to generalize SNE to handle multi-view data by introducing one combination coefficient to each view. The final probability distribution on the high dimensional space is then

$$p_{ij} = \sum_{t=1}^v \alpha^t p_{ij}^t, \quad (51)$$

where  $\alpha^t$  is the combination coefficient for view  $t$  and  $p_{ij}^t$  is the probability distribution on view  $t$ . This combination coefficient plays an important role in utilizing the complementary information and suppressing noise in multi-view data. Additionally, the original objective function contains only KL divergence; a 2-norm regularization term is added to balance the coefficients over all views

$$g(\alpha) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} + \lambda \|\alpha\|^2, \quad (52)$$

where  $\lambda$  is the tradeoff coefficient.

Han et al. (2012) proposed a new framework of sparse unsupervised dimensionality reduction for multi-view data. Considering the specific statistical property of each view, this algorithm first learns low-dimensional patterns from these views using the principal component analysis (PCA) algorithm. After combining the learned low-dimensional pattern of each view into one unified pattern, the construction of the low-dimensional consensus

representation can be formulated to approximate the matrix of patterns by means of a low-dimensional consensus base matrix and a loading matrix. To select the most discriminative feature for the spectral embedding of multiple views, a 1-norm is added into the loading matrix's columns and orthogonal constraints are imposed on the base matrix. A novel method called Spectral Sparse Multi-View Embedding (SSMVE) was subsequently developed to efficiently obtain the solution. Furthermore, since each row of the loading matrix is a vector concatenated by several parts which correspond to the different patterns learned from different views, a novel structured sparsity-inducing norm penalty was imposed on the loading matrix's rows to gain flexibility in sharing information across subsets of the views. Consequently, another approach for multi-view dimensionality reduction with structured sparsity penalty, namely, Structured Sparse Multi-View Dimensionality reduction (SSMVD), was proposed.

#### 7.4 Multi-view Metric Learning

The goal of metric learning for multi-view data is to construct embedding projections from the data in different representations into a shared feature space, so that the Euclidean distance in this space is meaningful not only within a single view, but also between different views.

Motivated by cross-media retrieval tasks, Quadrianto and Lampert (2011) studied the metric learning problem to find the joint Euclidean distance function to allow nearest neighbor queries. Following the classical principle of pulling samples together if they are related and pushing them apart if they are not, multi-view metric learning is formulated as follows. Suppose there are two sets of  $m$  data points,  $X = \{x_1, \dots, x_m\}$  and  $Y = \{y_1, \dots, y_m\}$  describing the same objects from two different views, and for each  $x_i \in X$  there exists a set  $S_{x_i}$  of data points from  $Y$  which are similar to  $x_i$ . Given  $X = \mathbb{R}^{d_1}$  and  $Y = \mathbb{R}^{d_2}$ , we seek the projection functions,

$$g_1 : \mathbb{R}^{d_1} \longrightarrow \mathbb{R}^D \quad \text{and} \quad g_2 : \mathbb{R}^{d_2} \longrightarrow \mathbb{R}^D,$$

with  $D \ll \min(d_1, d_2)$  that respects the neighborhood relationship  $\{S_{x_i}\}_{i=1}^m$ . Considering a linear parameterization of the functions  $g_1(x_i) = \langle w_1, \phi(x_i) \rangle$  and  $g_2(y_i) = \langle w_2, \phi(y_i) \rangle$ , then the metrics  $w_1$  and  $w_2$  are the goal of the learning, and the objective function can be written as

$$L(w_1, w_2, X, Y, S) = \sum_{i,j=1}^m L^{i,j}(w_1, w_2, x_i, y_j, S_{x_i}) + \eta\Omega(w_1) + \gamma\Omega(w_2), \quad (53)$$

where  $L^{i,j}(\cdot)$  is the loss function,  $\Omega(\cdot)$  is a regularizer on the parameters and  $\eta$  and  $\gamma$  are trade-off variables. By choosing the loss function appropriately, the properties the projected data are expected to have can be expressed. In particular, if it is hoped to ensure that similar objects across different views are mapped to nearby points, whereas dissimilar objects across different views are to be pushed apart, the loss function can be designed as the union of two different parts,

$$L(w_1, w_2, X, Y, S) = \frac{\mathbf{I}_{y_i \in S_{x_i}}}{2} \times L_1^{i,j} + \frac{1 - \mathbf{I}_{y_i \in S_{x_i}}}{2} \times L_2^{i,j}, \quad (54)$$

where the similarity term  $L_1^{i,j}$  forces similar objects to be at proximal locations in the latent space and the dissimilar term  $L_2^{i,j}$  pushes dissimilar objects away from one another. This objective function can be decomposed into a difference of two concave functions, thus it can be solved efficiently by the concave convex procedure (CCCP).

Since various different low-level visual features can be extracted to comprehensively represent the image in image processing, it is difficult to choose which feature to depend on to measure the similarity between images. Thus Yu et al. (2012b) proposed a semi-supervised multi-view distance metric learning (SSM-DML) algorithm to construct an accurate metric to precisely measure the dissimilarity between different examples associated with multiple views. Formally define a matrix  $\mathbf{F} = [\mathbf{F}_1^T, \dots, \mathbf{F}_N^T]^T$ , where  $F_{ij}$  is the confidence of  $x_i$  with the label  $y_j$ , and then this matrix  $\mathbf{F}$  can be obtained by minimizing the following objective function:

$$Q = \sum_{i,j=1}^N \mathbf{W}_{ij} \left\| \frac{\mathbf{F}_i}{\sqrt{\mathbf{D}_{ii}}} - \frac{\mathbf{F}_j}{\sqrt{\mathbf{D}_{jj}}} \right\|^2 + \mu \sum_{i=1}^N \|\mathbf{F}_i - \mathbf{Y}_i\|^2, \quad (55)$$

where  $\mathbf{W}$  is an affinity matrix with  $W_{ij}$  indicating the dissimilarity measure between  $x_i$  and  $x_j$ , and  $\mathbf{D}$  is a diagonal matrix with  $D_{ii}$  equal to the sum of the  $i$ -th row of  $\mathbf{W}$ . The first term in Eq. (55) implies the smoothness of the labels on the graph and the second term indicates the constraint of the training data. Suppose  $X^i$  represents the  $i$ -th view of the example; by linearly combining the graphs constructed from multi-view features sets through the weights  $\alpha$ , Eq. (55) can be extended to the multi-view feature sets

$$Q = \sum_{k=1}^K \sum_{i,j=1}^N \alpha_k \mathbf{W}_{ij}^k \left\| \frac{\mathbf{F}_i}{\sqrt{\mathbf{D}_{ii}^k}} - \frac{\mathbf{F}_j}{\sqrt{\mathbf{D}_{jj}^k}} \right\|^2 + \mu \sum_{i=1}^N \|\mathbf{F}_i - \mathbf{Y}_i\|^2 + \lambda \|\alpha\|^2 \quad (56)$$

s. t.  $\sum_{k=1}^K \alpha_k = 1.$

Then through adopting alternating optimization to solve the above optimization problem, SSM-DML can learn the multi-view distance metrics from multiple feature sets and the labels of unlabeled data simultaneously.

Zhai et al. (2012) also studied the multi-view metric learning problem in the semi-supervised learning setting, and proposed a new method called Multi-view Metric Learning with Global consistency and Local smoothness (MVML-GL), which jointly considers global consistency and local smoothness. This algorithm is accomplished in two steps: (1) seek a shared latent feature space to establish the relationship between data from multi-view observation spaces according to pairs of labeled instances; (2) learn the relationships between the input space of each observation and the shared latent space for unlabeled and test data. It is worth noting that this first step is globally consistent, as it simultaneously considers the geometric structures contained in each view and connections between the data from different views, and the second step is locally smooth, which enables each instance to have its own specific distance metric instead of applying a uniform metric for all instances. Additionally, both steps can be formulated as convex optimization problems with closed form solutions, thus they can be efficiently solved.



## 7.5 Latent Space Models

Besides the aforementioned methods, which aim to conduct meaningful dimensional reduction for multi-view data, there are also works that concentrate on analyzing the relationships between different views. These methods are used to build latent space models, with which multiple views can be connected with one another through latent variables, and the information can be propagated from one view to another view.

### 7.5.1 SHARED GAUSSIAN PROCESS LATENT VARIABLE MODEL

Gaussian processes (GPs) are powerful models for classification and regression that subsume numerous classes of function approximators, such as single hidden-layer neural networks and RBF networks. Lawrence (2004) first proposed the Gaussian process latent variable model (GPLVM) as a new technique for non-linear dimensional reduction. Shon et al. (2006) proposed the shared GPLVM (SGPLVM) as a generalization of the GPLVM model that can handle multiple observation spaces, where each set of observations is parameterized by a different set of kernel parameters.

Let  $Y, Z$  be matrices of observations drawn from spaces of dimensionality  $D_Y, D_Z$  respectively, and  $X$  be a latent space of dimensionality  $D_X \ll D_Y, D_Z$ . Assume that each latent point  $x_i$  generates a pair of observations  $y_i, z_i$  via GPs parameterized non-linear functions  $f_Y : X \rightarrow Y$  and  $f_Z : X \rightarrow Z$ . By using an exponential (RBF) kernel to define the similarity between two data points  $x, x'$

$$k(x, x') = \alpha_Y \exp(-\frac{\gamma_Y}{2} \|x - x'\|^2) + \delta_{x, x'} \beta_Y^{-1}, \quad (57)$$

the priors  $P(\theta_Y), P(\theta_Z), P(\theta_X)$  ( $\theta = \{\alpha, \beta, \gamma\}$ ) and the likelihoods  $P(Y), P(Z)$  for the  $Y, Z$  observation spaces are given by

$$P(Y|\theta_Y, X) = \frac{|W|^N}{\sqrt{(2\pi)^{ND_Y} |K|^{D_Y}}} \exp(-\frac{1}{2} \sum_{k=1}^{D_Y} w_k^2 Y_k^T K_Y^{-1} Y_k), \quad (58)$$

$$P(Z|\theta_Z, X) = \frac{|W|^N}{\sqrt{(2\pi)^{ND_Z} |K|^{D_Z}}} \exp(-\frac{1}{2} \sum_{k=1}^{D_Z} w_k^2 Z_k^T K_Z^{-1} Z_k), \quad (59)$$

$$P(\theta_Y) \propto \frac{1}{\alpha_Y \beta_Y \gamma_Y} \quad P(\theta_Z) \propto \frac{1}{\alpha_Z \beta_Z \gamma_Z}, \quad (60)$$

$$P(X) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} \sum_i \|x_i\|^2), \quad (61)$$

then the joint likelihood can be written as

$$P_{GP}(X, Y, Z, \theta_Y, \theta_Z) = P(Y|\theta_Y, X)P(Z|\theta_Z, X)P(\theta_Y)P(\theta_Z)P(X). \quad (62)$$

By using a conjugate gradient solver to maximize Eq. (62), the model can learn a separate kernel for each observation space and a single set of common latent points.

Given a trained SGPLVM, we would like to infer the parameters in one observation space given the parameters in the other observation space. This problem can be solved in two steps. First, we determine the most likely latent coordinate  $x$  given the observation  $y$  using  $\max_x L_x(x, y)$ . Once the correct latent coordinate  $x$  has been inferred for a given  $y$ , the model uses the trained SGPLVM to predict the corresponding observation  $z$ .

### 7.5.2 SHARED KERNEL INFORMATION EMBEDDING

Given samples drawn from a distribution  $p(x)$ , Kernel Information Embedding (Memisevic, 2006) aims to find a low-dimensional latent distribution,  $p(z)$ , that captures the structure of the data, along with explicit bidirectional probabilistic mappings between the latent space and the data space. In particular, KIE finds the joint distribution  $p(x, z)$  that maximizes the mutual information between the latent distribution and the data distribution:

$$\begin{aligned} I(x, z) &= \int p(x, z) \log \frac{p(x, z)}{p(x)p(z)} dx dz \\ &= H(x) + H(z) - H(x, z), \end{aligned} \quad (63)$$

where  $H(\cdot)$  is the usual Shannon entropy, which can be estimated by kernel density.

The shared KIE (sKIE) (Sigal et al., 2009; Memisevic et al., 2012), which can be seen as the extension of KIE, constructs the joint embedding for two views by maximizing the mutual information  $I((x, y), z)$ . Assuming the conditional independence of  $x$  and  $y$  given  $z$ ,  $I((x, y), z)$  can be expressed as a sum of two mutual information terms,

$$I((x, y), z) = I(x, z) + I(y, z), \quad (64)$$

where  $I(x, z)$  and  $I(y, z)$  can be formulated as KIE.

An application of this algorithm is human pose inference. For discriminative pose inference, the aim is to find likely poses  $y$  conditioned on input image features  $x^*$ . Then the conditional pose distribution is:

$$p(y|x^*) = \int_z p(y|z)p(z|x^*)dz. \quad (65)$$

Alternatively, the focus can be on identifying the principal modes of  $p(y|x^*)$ . To this end, it is assumed that the principal modes of  $p(y|x^*)$  coincide with the principal modes of the conditional latent distribution  $p(z|x^*)$ . That is, a search is first conducted for local maxima of  $p(z|x^*)$ , denoted  $\{z_k^*\}_{k=1}^K$  for  $K$  modes. From these latent points it is straightforward to perform either MAP inference or take the expectation over the conditional pose distributions  $p(y|z_k^*)$ .

### 7.5.3 FACTORIZED ORTHOGONAL LATENT SPACE

Both sGPLVM and sKIE only consider the shared information in the views of data but ignore the private part in each view. Salzman et al. (2010) proposed a robust approach called FOLS to factorize the latent space into shared and private spaces by introducing orthogonality constraints, which penalize redundant latent representations.

For minimal factorization, the shared and private latent spaces are required to be non-redundant; in other words, it is desirable to penalize the redundancy of different private spaces and thus encourage the representation of common information in the shared space. More formally, define  $Y^i = [y_1^i, \dots, y_N^i]^T$  as the set of observations from a single view  $i$ , with  $1 \leq i \leq V$ . Additionally, let  $X = [x_1, \dots, x_N]^T$  be the latent space shared across different views,  $Z^i = [Z_1^i, \dots, Z_N^i]^T$  be the private space for  $i$ -th view, and  $M^i = [m_1^i, \dots, m_N^i]^T$  be the joint shared-private latent space for each view, with  $m_j^i = [x_j, z_j^i]$ . By imposing the

above mentioned non-redundant constraint as a soft penalty, a FOLS model can be learned by minimizing

$$\begin{aligned} \mathcal{L} = & L + \underbrace{\alpha \sum_i (\|X^T \cdot Z^i\|_F^2 + \sum_{j>i} \|(Z^i)^T \cdot Z^j\|_F^2)}_{\text{orthogonality}} \\ & + \underbrace{\gamma \sum_i \phi(s_i)}_{\text{low dimensionality}} + \underbrace{\eta \sum_i (E_0^i - \sum_j s_{i,j}^2)^2}_{\text{energy conservation}}, \end{aligned} \quad (66)$$

where  $s_i$  are the singular values of  $M^i$ ,  $E_0^i$  is the energy of stream  $i$ , and  $L$  is the loss function of the particular model into which the factorization constraints are introduced. In the sGPLVM and sKIE models,  $L$  represents the square loss, or the negative mutual information between each joint latent space and its corresponding data stream.

#### 7.5.4 FACTORIZED LATENT SPACES WITH STRUCTURED SPARSITY

Inspired by sparse coding techniques, Jia et al. (2010) proposed a novel approach to finding a latent space in which the information is correctly factorized into shared and private parts, while avoiding the computational burden of previous techniques. In particular, this algorithm represents each view as a linear combination of view-dependent dictionary entries. While the dictionaries are specific to each view, the weights of these dictionaries act as latent variables and are the same for all the views.

More formally, to find a shared-private factorization of the latent embedding  $\alpha$  that represents the multiple input modalities, the algorithm adopts the idea of structured sparsity and aims to find a set of dictionaries  $\mathcal{D} = \{D^1, \dots, D^V\}$ . This problem can be formulated as,

$$\min_{\mathcal{D}, \alpha} \frac{1}{N} \sum_{v=1}^V \|X^v - D^v \alpha\|_F^2 + \lambda \sum_{v=1}^V \psi((D^v)^T) + \gamma \psi(\alpha), \quad (67)$$

where the first item measures the loss, the second item encourages each view to only use a limited number of latent dimensions, and the third item indicates a relaxation of rank constraints to discover the dimensionality of the latent space.

At inference, given a new observation  $\{x_*^1, \dots, x_*^V\}$ , the corresponding latent embedding  $\alpha_*$  can be obtained by solving the convex problem

$$\min_{\alpha_*} \sum_{v=1}^V \|x_*^v - D^v \alpha_*\|_2^2 + \gamma \|\alpha_*\|_1, \quad (68)$$

where the regularizer allows us to deal with noise in the observations.

#### 7.5.5 LATENT SPACE MARKOV NETWORK

Chen et al. (2010) constructed a predictive subspace shared by multi-view data based on the generic multi-view latent space Markov network (MN), under the assumption that the data from different views and the response variables are conditionally independent given a set of latent variables.

The two-view latent space Markov networks consist of two views of input data  $\mathbf{X} : \{X_n\}$  and  $\mathbf{Z} : \{Z_m\}$  and a set of latent variables  $\mathbf{H} : \{H_k\}$ . According to random field theory, the marginal distributions for two views respectively can be written in the exponential forms

$$p(x) = \exp\left\{\sum_i \theta_i^T \phi(x_i, x_{i+1}) - A(\theta)\right\}, \quad (69)$$

$$p(z) = \exp\left\{\sum_j \eta_j^T \psi(z_j, z_{j+1}) - B(\eta)\right\}, \quad (70)$$

where  $\phi$  and  $\psi$  are feature functions,  $A$  and  $B$  are log partition functions. For the latent variables, the marginal distribution is

$$p(h) = \prod_k \exp\{\lambda_k^T \varphi(h_k) - C_k(\lambda_k)\}, \quad (71)$$

where  $\varphi(h_k)$  is the feature vector of  $h_k$ ,  $C_k$  is the log-partition function. By combining the above components in the log-domain, the joint model distribution is defined as

$$\begin{aligned} p(x, z, h) \propto & \exp\left\{\sum_i \theta_i^T \phi(x_i, x_{i+1}) + \sum_j \eta_j^T \psi(z_j, z_{j+1}) + \sum_k \lambda_k^T \varphi(h_k)\right. \\ & \left. + \sum_{ik} \phi(x_i, x_{i+1})^T W_i^k \varphi(h_k) + \sum_{jk} \psi(z_j, z_{j+1})^T U_j^k \varphi(h_k)\right\}. \end{aligned}$$

Additionally considering each input sample is associated with a supervised response variable  $y \in \{1, \dots, T\}$ , we can define

$$p(y|h) = \frac{\exp\{V^T f(h, y)\}}{\sum_{y'} \exp\{V^T f(h, y')\}}, \quad (72)$$

where  $f(h, y)$  is the feature vector whose elements from  $(y-1)K+1$  to  $yK$  are those of  $h$  and all others are 0. Accordingly,  $V$  is a stacking parameter vector of  $T$  sub-vectors  $V_y$ , each of which corresponds to a class label  $y$ .

Although this multi-view latent space MNs can be learned by maximum likelihood estimation (MLE), Chen et al. (2010) estimated the decision boundary directly in a large margin approach. Assume the discriminant function  $F(y, h; V)$  is linear, that is,  $F(y, h; V) = V^T f(h, y)$ , which looks like the discriminant function  $W^T X$  in SVM. Then the objective function is

$$\min_{\Theta, V} L(\Theta) + \frac{1}{2} C_1 \|V\|^2 + C_2 \mathcal{R}_{hinge}(V), \quad (73)$$

where the first item  $L(\Theta) = -\sum_d \log p(x_d, z_d)$  is the negative data likelihood, the second item is the constraint of the decision boundary, and the third item hinge loss acts as the slack variable  $\xi$  in SVM. Since Eq. (73) maximizes the data likelihood and minimizes training loss, it can be expected that by solving this problem we can find a predictive latent space representation  $p(h|x, z)$  and a prediction model parameter  $V$  at the same time.

## 8. Applications

In general, by exploiting the consistency and complement of multiple views, learning models from multi-view data will lead to an improvement in learning performance. Thus multi-view learning has been applied successfully in a number of real-world applications.

Since Blum and Mitchell (1998) first proposed the co-training algorithm and applied it to the web document classification problem, this novel method has caught the attention of many researchers and has been widely applied in the field of natural language processing (Craven et al., 2000; Müller et al., 2002; Phillips and Riloff, 2002). Pierce and Cardie (2001) studied the learning behavior of co-training and showed that given a small set of labeled training data and a large set of unlabeled data, co-training can reduce the difference in error between co-trained classifiers and fully supervised classifiers trained on a labeled version of all available data by 36%. Unlike previous efforts which cope with the task of word sense disambiguation in a supervised way, Mihalcea (2004) suggested combining co-training with majority voting, with the effect of smoothing the learning curves to improve average performance. Maeireizo et al. (2004) investigated the applicability of co-training to train classifiers that predict emotions in spoken dialogues on features pre-processed in a wrapper approach with forward selection. Kiritchenko and Matwin (2001); Kockelkorn et al. (2003) and Scheffer (2004) treated the email classification problem in the framework of semi-supervised learning, so that the cost of labeling unlabeled data could be eliminated, and a co-training method employed to significantly improve learning performance. Besides these applications involving text or natural language processing, co-training has also found application in the field of computer vision. For instance, Liu and Yuen (2011) studied the human action recognition problem and introduced two new confidence measures, i.e. inter-view confidence and intra-view confidence, to address view sufficiency and view dependency issues in co-training. Christoudias et al. (2009a) designed a probabilistic heteroscedastic approach to co-training, which discovers the amount of noise while solving multi-view object recognition tasks. Feng and Chua (2003) and Feng et al. (2004) addressed the image annotating problem by combining co-training with active learning. Thus the requisition for the large labeled training corpus for effective learning is relaxed in co-training and the best examples are selected to label at each stage to maximize the learning objective in active learning. Considering various kinds of visual features, such as color and texture features, as sufficient and uncorrelated views of an image, Zhou et al. (2004) and Cheng and Wang (2007) introduced a co-training algorithm to conduct relevance feedback in content-based image retrieval.

As for multiple kernel learning, Kumar and Sminchisescu (2007); Lin et al. (2007) and Varma and Ray (2007) applied it to object classification by linearly combining similarity functions between images so that the combined similarity function yields improved classification performance. Longworth and Gales (2008) employed multiple kernel learning for object detection with the goal of finding an optimal combination of exponential  $\chi^2$  kernels, each of which would capture a different feature channel, such as the distribution of edges, and visual words. Kembhavi et al. (2009) proposed an incremental multiple kernel learning approach for object recognition. In this case, “incremental” means that the images of objects in poses more commonly observed in the scene as well as the kernel weights will be updated in each iteration, thus further improving the learning performance.

Subspace learning is an important tool for analyzing the relationships between different views of the data and has a number of applications. Donner et al. (2006) introduced a fast active appearance model search algorithm based on CCA. Zheng et al. (2006) used KCCA to solve the facial expression recognition problem. Dhillon et al. (2011) computed the CCA between different views of the data to estimate low dimensional context specific word representations from unlabeled data in NLP tasks. Fu et al. (2008) effectively solved the face recognition task by constructing a linear subspace in which the cumulative canonical correlation between any pair of feature sets is maximized. Zhang et al. (2012) studied the hyperspectral remote sensing image classification problem in the approach of multi-view learning, and introduced the patch alignment framework to linearly combine multiple features in an optimal way and a unified low-dimensional representation of these multiple features for subsequent classification. Considering that the key issue in cartoon character retrieval is proper representation that describes the cartoon character effectively, Yu et al. (2012a) introduced a semi-supervised multi-view subspace learning algorithm which encodes different features in a unified space, as illustrated in Figure 5. In this unified subspace, the Euclidean distance can be straightforwardly used to measure the distance between two cartoon characters. To improve the performance of the ranking and difficulty estimation in image retrieval, Li et al. (2011) applied multi-view embedding (ME) to images represented by multiple features for integrating a joint subspace by preserving the neighborhood information in each feature space, as illustrated in Figure 6. To eliminate the “out of sample” and huge computation cost problem, a linear multi-view embedding algorithm was developed which learns a linear transformation from a small set of data and can effectively infer the subspace features of new data.

## 9. Performance Evaluation

In this section, we introduce some widely used datasets in multi-view learning experiments and make an empirical comparison of several representative multi-view learning algorithms with single-view learning algorithms.

**Data Sets for Multi-view Learning.** So far, several datasets have been widely employed in multi-view learning experiments. Here we give a simple introduction to these datasets.

- WebKB dataset <sup>1</sup> is the most famous dataset used in multi-view learning, on which the co-training algorithm was first evaluated. This dataset consists of 8282 academic web pages collected from computer science department web sites at four universities: Cornell, University of Washington, University of Wisconsin, and University of Texas. These pages can be grouped into six classes: student, staff, faculty, department, course and project. There are two views containing the text on the page and the anchor text of hyperlink respectively.
- Citeseer dataset <sup>2</sup> is a collection of scientific publications which contains 3312 documents belonging to six classes. There are three natural views for each document:

---

1. <http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/>

2. <http://komarix.org/ac/ds/>

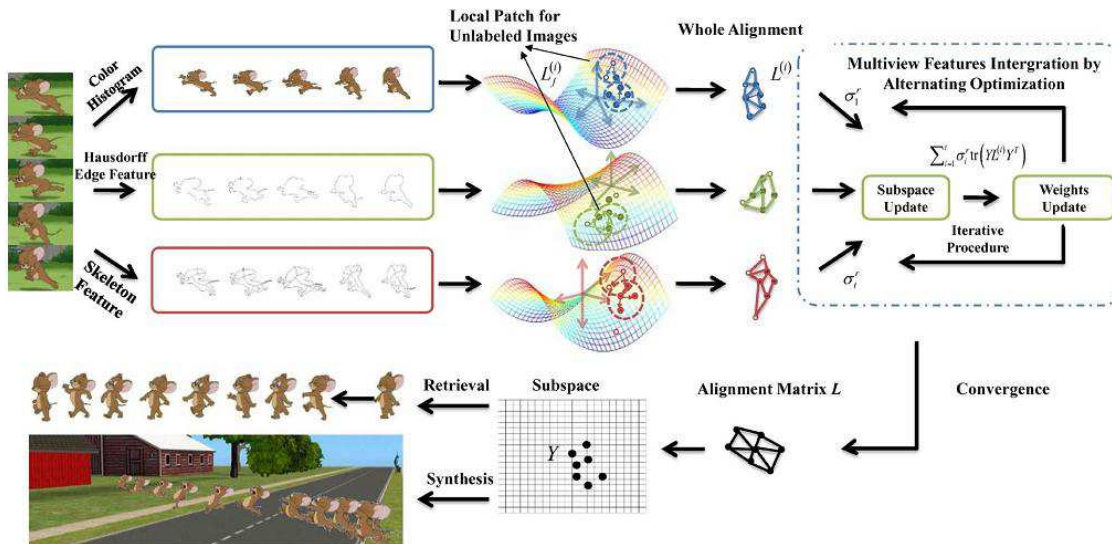


Figure 5: Flowchart of the semi-supervised multi-view subspace learning algorithm (Yu et al., 2012a). The method first extracts multi-view features from cartoon characters. Then, by considering the constraints of each local patch and the complementary characteristics of multi-view features, the low dimensional representation  $Y$  can be obtained through solving an alternating optimization problem. Finally, the cartoon character retrieval and clip synthesis can be conducted by measuring the dissimilarity in the subspace  $Y$ .

the text view consists of the title and abstract of the paper; the two link views are inbound and outbound references.

- Some popular data sets coming from UCI repository <sup>3</sup> are suitable for evaluating multi-view learning. For example, the internet advertisement dataset contains images from various web pages that are characterized either as advertisements or non-advertisements. The instances are described in terms of six views, which are the geometry of the images, the base url, the image url, the target url, the anchor text and the alt text.
- There are also a number of other multimedia datasets usually employed in experiments on image annotation, image classification and image retrieval, which include TRECVID2003 video dataset <sup>4</sup>, Caltech256 <sup>5</sup>, etc. We extract different visual features to represent multiple views of the data, such as color histogram, edge direction histogram, and wavelet texture.

3. <http://archive.ics.uci.edu/ml/>  
 4. <http://www-nlpir.nist.gov/projects/tv2003/>  
 5. [http://www.vision.caltech.edu/Image\\_Datasets/Caltech256/](http://www.vision.caltech.edu/Image_Datasets/Caltech256/)

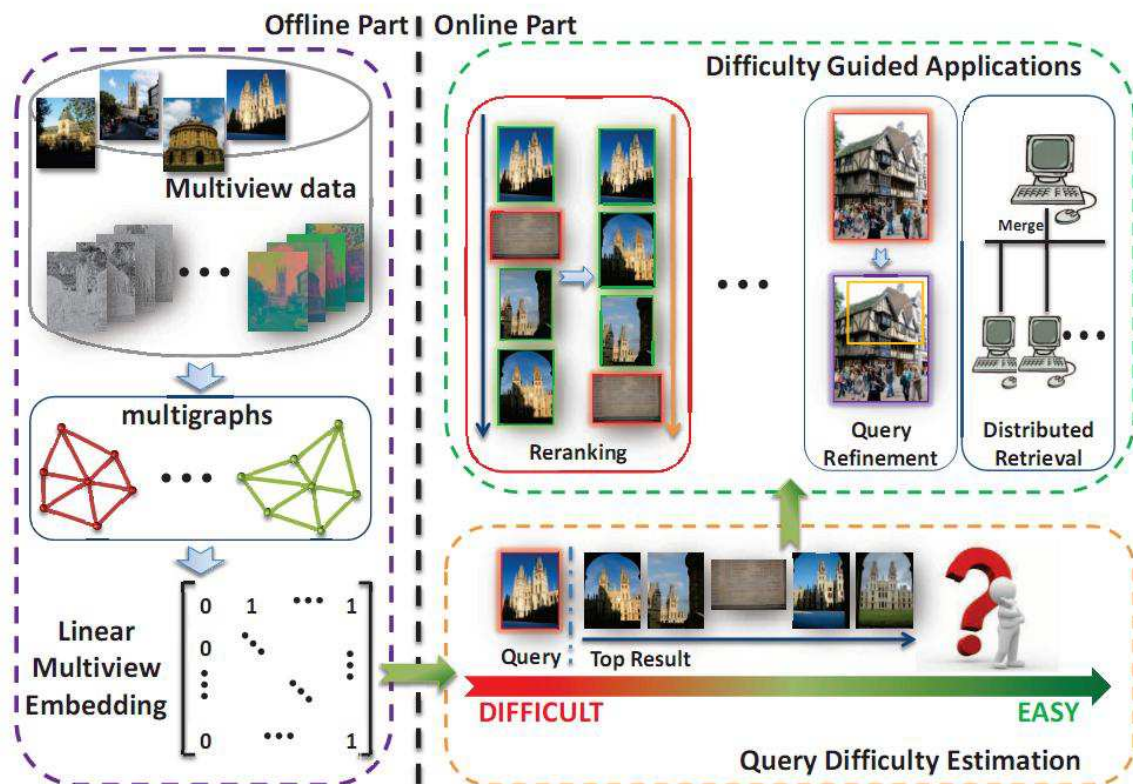


Figure 6: Application of linear multi-view embedding in difficulty-guided image retrieval (Li et al., 2011).

**Empirical Evaluation.** To illustrate the benefits of multi-view learning methods compared to traditional single-view learning, Table 1 presents a list drawn from several published multi-view learning papers. Blum and Mitchell (1998); Nigam and Ghani (2000); Brefeld and Scheffer (2004); Sindhwani et al. (2005); Yu et al. (2011) and Zhu et al. (2012) used the WebKB data as one of the evaluation datasets. Due to the different preprocessing steps of the algorithms by different researchers, it is difficult to make a direct comparison of the proposed methods; thus we denote them as  $WebKB_1, \dots, WebKB_6$  respectively and show the comparison results between the proposed multi-view learning methods and single-view learning methods in the table.

On the  $WebKB_1$  data, Blum and Mitchell (1998) evaluated the co-training algorithm and compared its performance with that of the single-view learning algorithm naive Bayes. On the  $WebKB_2$ , Nigam and Ghani (2000) evaluated the proposed co-EM method. Brefeld and Scheffer (2004) developed a novel co-EM based on SVM and showed its satisfactory performance compared to single-view SVM and co-trained naive Bayes on the  $WebKB_3$ . Sindhwani et al. (2005) evaluated their proposed co-regularization method on the  $WebKB_4$ , and compared it to the single-view regularization method, single-view SVM and co-trained Laplace SVM. Yu et al. (2011) illustrated the co-training algorithm in a graphical way, developed Bayesian



Table 1: Comparison between multi-view learning and single-view learning methods

DataSet (reference)	Data	Single-view		Multi-view	
WebKB <sub>1</sub> (Blum and Mitchell, 1998) Error rate		<b>naive Bayes</b>		<b>Co-trained NB</b>	
	Page	12.9%		6.2%	
	Hyperlink	12.4%		11.6%	
	Page+Hyperlink	11.1%		5.0%	
WebKB <sub>2</sub> (Nigam and Ghani, 2000) Error rate	Page+Hyperlink	13.0%		5.4%	4.3%
		<b>naive Bayes</b>		<b>Co-trained NB</b>	<b>Co-EM NB</b>
WebKB <sub>3</sub> (Brefeld and Scheffer, 2004) Error rate	Page+Hyperlink	13.0%	10.39%	5.08%	0.99%
		<b>naive Bayes</b>	<b>SVM</b>	<b>Co-EM NB</b>	<b>Co-EM SVM</b>
WebKB <sub>4</sub> (Sindhvani et al., 2005) mean PRBEP		<b>SVM</b>	<b>RLS</b>	<b>Co-LapSVM</b>	<b>Co-LapRLS</b>
	Page	77.8%	71.6%	93.3%	92.0%
	Hyperlink	74.4%	72.0%	94.3%	94.4%
	Page+Hyperlink	84.4%	78.3%	94.2%	93.6%
WebKB <sub>5</sub> (Yu et al., 2011) AUC	Page+Hyperlink	<b>GPLR</b>		<b>Co-trained GPLR</b>	<b>Bayesian Co-training</b>
		0.57%		0.56 %	0.58%
WebKB <sub>6</sub> (Zhu et al., 2012) AUC	Page+Hyperlink	<b>KPCA</b>		<b>KCCA</b>	<b>MKCCA</b>
		94.5%		86.6%	94.6%
UCI <sub>1</sub> (Gönen and Alpaydin, 2008) ACC		<b>SVM <math>K_P</math></b>		<b>MKL <math>K_P - K_G</math></b>	<b>LMKL <math>K_P - K_G</math></b>
	Banana	56.51%		81.99%	83.84%
	Heart	72.78%		75.78%	79.44%
	Ionosphere	91.54%		93.68%	93.33%
	Pima	66.95%		98.86%	98.69%
	Sonar	65.29%		80.29%	79.57%
UCI <sub>2</sub> (Varma and Babu, 2009) ACC		<b>LP-SVM</b>		<b>MKL</b>	<b>GMKL</b>
	Ionosphere	93.0%		87.7%	94.1%
	Parkinsons	86.2%		84.7%	92.6%
	Musk	81.5%		87.0%	93.3%
	Sonar	73.7%		79.5%	82.0%
	Wpbc	76.2%		69.4%	78.3%
UCI <sub>3</sub> (Rakotomamonjy et al., 2008) ACC (Time(s))				<b>SILP</b>	<b>Simple MKL</b>
	Liver			65.9% (47.6)	65.9% (18.9)
	Pima			76.5% (224)	76.5% (79.0)
	Ionosphere			91.7% (535)	91.5% (123)
	Wpbc			76.8% (88.6)	76.7% (20.6)
	Sonar			80.5% (2290)	80.6% (163)
UCI <sub>4</sub> (Xu et al., 2010) ACC (Time(s))				<b>Simple MKL</b>	<b>MKLGL</b>
	Ionosphere			91.5% (79.9)	92.0% (12.0)
	Breast			96.5% (110.5)	96.6% (14.1)
	Sonar			82.0% (57.0)	82.0% (5.7)
	Pima			73.4% (94.5)	73.5% (15.1)

co-training, and performed experiments on the WebKB<sub>5</sub>. On the WebKB<sub>6</sub>, Zhu et al. (2012) compared the performances of multi-view approaches and single-view approaches in respect of subspace learning.

Gönen and Alpaydin (2008); Varma and Babu (2009); Rakotomamonjy et al. (2008) and Xu et al. (2010) used the benchmark datasets from the UCI machine learning repository. Thus we use UCI<sub>1</sub>, ..., UCI<sub>4</sub> to denote the respective different experiments of these works. In these experiments, several representative multiple kernel learning methods, such as localized MKL and simple MKL, were evaluated in terms of accuracy and time cost. From these comparison results, we discover that multi-view learning methods designed appropriately for real-world applications can indeed improve performance significantly compared to single-view learning methods.

## 10. Conclusions

In many scenarios, more than one view can be provided to describe the data. Instead of selecting one view from the corpus or simply concatenating them for learning, we are more interested in algorithms that can learn models from multi-view data by considering the

diversity of different views. In this survey paper, we have therefore reviewed several current trends of multi-view learning and classified these algorithms into three different settings: co-training, multiple kernel learning, and subspace learning. Through analyzing these different approaches to the integration of multiple views, we observe that they mainly depend on either the consensus principle or the complementary principle to ensure their success. Furthermore, we also studied the problems with respect to how to construct multiple views and how to evaluate these views. The experimental results show the extensive development of multi-view learning and its promising performance compared to single-view learning.

Although significant work has been carried out in this field, several important research issues need to be addressed in the future. Since the properties of different views largely influence the performance of multi-view learning, it is necessary to place more emphasis on methods to construct, analyze and evaluate the views. For the three groups of multi-view learning algorithms, each have their own advantages, but they are mainly designed and developed separately. Therefore it would be valuable to develop a general framework of multi-view learning which includes the merits of different multi-view learning methods.

We conclude that multi-view learning is effective and promising in practice, but it has not been well-addressed to date. There is still much work to be done to better process multi-view data in a wide variety of applications.

## References

- Steven Abney. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 360–367, 2002.
- S. Akaho. A kernel method for canonical correlation analysis. *Arxiv preprint cs/0609071*, 2006.
- M.R. Amini and C. Goutte. A co-classification approach to learning from multilingual corpora. *Machine learning*, 79(1):105–121, 2010.
- M.R. Amini, N. Usunier, C. Goutte, et al. Learning from multiple partially observed views—an application to multilingual text categorization. In *NIPS 22:2009*, volume 1, pages 28–36, 2010.
- Hal Daum III Scott L. DuVall Anusua Trivedi, Piyush Rai. Multiview clustering with incomplete views. In *NIPS 2010: Machine Learning for Social Computing Workshop*, 2010.
- F.R. Bach, G.R.G. Lanckriet, and M.I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM, 2004.
- M.F. Balcan, A. Blum, and Y. Ke. Co-training and expansion: Towards bridging theory and practice. *Computer Science Department*, page 154, 2004.
- K.P. Bennett, M. Momma, and M.J. Embrechts. Mark: A boosting algorithm for heterogeneous kernel models. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 24–31. ACM, 2002.

- J. Bi, T. Zhang, and K.P. Bennett. Column-generation boosting methods for mixture of kernels. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 521–526. ACM, 2004.
- S. Bickel and T. Scheffer. Multi-view clustering. In *Proceedings of the IEEE international conference on data mining*, volume 36, 2004.
- A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.
- U. Brefeld and T. Scheffer. Co-em support vector learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 16. ACM, 2004.
- U. Brefeld, C. Büscher, and T. Scheffer. Multi-view discriminative sequential learning. *Machine Learning: ECML 2005*, pages 60–71, 2005.
- U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel. Efficient co-regularised least squares regression. In *Proceedings of the 23rd international conference on Machine learning*, pages 137–144. ACM, 2006.
- L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- J. Cai and H.W. Sun. Convergence rate of kernel canonical correlation analysis. *Science China Mathematics*, pages 1–10, 2011.
- O. Chapelle and A. Rakotomamonjy. Second order optimization of kernel parameters. In *Proc. of the NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, 2008.
- K. Chaudhuri, S.M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136. ACM, 2009.
- M. Chen, K.Q. Weinberger, and Y. Chen. Automatic feature decomposition for single view co-training. In *International Conference on Machine Learning*, 2011.
- N. Chen, J. Zhu, and E.P. Xing. Predictive subspace learning for multi-view data: A large margin approach. *Advances in Neural Information Processing Systems*, 24, 2010.
- J. Cheng and K. Wang. Active learning for image retrieval with co-svm. *Pattern recognition*, 40(1):330–334, 2007.
- C.M. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in the presence of view disagreement. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, 2008.
- C.M. Christoudias, R. Urtasun, A. Kapoorz, and T. Darrell. Co-training with noisy perceptual observations. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2844–2851. IEEE, 2009a.

- M. Christoudias, R. Urtasun, and T. Darrell. Bayesian localized multiple kernel learning. *University of California at Berkeley, Tech. Rep*, 2009b.
- C. Cortes, M. Mohri, and A. Rostamizadeh. Learning non-linear combinations of kernels. *Advances in Neural Information Processing Systems*, 22:396–404, 2009.
- C. Cortes, M. Mohri, and A. Rostamizadeh. Generalization bounds for learning kernels. *Proceedings, 27th ICML*, 2010.
- K. Crammer, J. Keshet, and Y. Singer. Kernel design using boosting. *Advances in neural information processing systems*, 15:537–544, 2002.
- M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the world wide web. *Artificial Intelligence*, 118(1):69–113, 2000.
- S. Dasgupta, M.L. Littman, and D. McAllester. Pac generalization bounds for co-training. *Advances in neural information processing systems*, 1:375–382, 2002.
- P.S. Dhillon, D. Foster, and L. Ungar. Multi-view learning of word embeddings via cca. In *NIPS*, volume 24, pages 199–207, 2011.
- W. Di and M.M. Crawford. View generation for multiview maximum disagreement based active learning for hyperspectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, (99):1–13, 2012.
- T. Diethe, D.R. Hardoon, and J. Shawe-Taylor. Multiview fisher discriminant analysis. In *NIPS workshop on learning from multiple sources*, 2008.
- T.G. Dietterichl. Ensemble learning. 2002.
- R. Donner, M. Reiter, G. Langs, P. Peloschek, and H. Bischof. Fast active appearance model search using canonical correlation analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1690–1694, 2006.
- N. Duffy and D. Helmbold. Leveraging for regression. In *Proceedings of COLT*, volume 13. Citeseer, 2000.
- J. Farquhar, D. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak. Two view learning: Svm-2k, theory and practice. 2005.
- H. Feng, R. Shi, and T.S. Chua. A bootstrapping framework for annotating and retrieving www images. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 960–967. ACM, 2004.
- H.M. Feng and T.S. Chua. A bootstrapping approach to annotating large image collection. In *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 55–62. ACM, 2003.

- Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 148–156. MORGAN KAUFMANN PUBLISHERS, INC., 1996.
- J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2001.
- Y. Fu, L. Cao, G. Guo, and T.S. Huang. Multiple feature fusion by subspace learning. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 127–134. ACM, 2008.
- K. Fukumizu, F.R. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *The Journal of Machine Learning Research*, 8:361–383, 2007.
- S. Goldman and Y. Zhou. Enhancing supervised learning with unlabeled data. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 327–334. Citeseer, 2000.
- M. Gönen and E. Alpaydin. Localized multiple kernel learning. In *Proceedings of the 25th international conference on Machine learning*, pages 352–359. ACM, 2008.
- M. Gönen and E. Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.
- I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.
- Y. Han, F. Wu, D. Tao, J. Shao, Y. Zhuang, and J. Jiang. Sparse unsupervised dimensionality reduction for multiple view data. 2012.
- D.R. Hardoon and J. Shawe-Taylor. Convergence analysis of kernel canonical correlation analysis: theory and practice. *Machine learning*, 74(1):23–38, 2009.
- T.K. Ho. The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844, 1998.
- H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.
- A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(2):153–158, 1997.
- Y. Jia, M. Salzmann, and T. Darrell. Factorized latent spaces with structured sparsity. *Advances in Neural Information Processing Systems*, 23:982–990, 2010.
- S. Kakade and D. Foster. Multi-view regression via canonical correlation analysis. *Learning Theory*, pages 82–96, 2007.
- A. Kembhavi, B. Siddiquie, R. Miezianko, S. McCloskey, and L.S. Davis. Incremental multiple kernel learning for object recognition. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 638–645. IEEE, 2009.

- S. Kiritchenko and S. Matwin. Email classification with co-training. In *Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative research*, page 8. IBM Press, 2001.
- M. Kloft and G. Blanchard. The local rademacher complexity of lp-norm multiple kernel learning. *Arxiv preprint arXiv:1103.0790*, 2011.
- M. Kockelkorn, A. Lüneburg, and T. Scheffer. Using transduction and multi-view learning to answer emails. *Knowledge Discovery in Databases: PKDD 2003*, pages 266–277, 2003.
- A. Kumar and H. Daumé III. A co-training approach for multi-view spectral clustering. In *International Conference on Machine Learning*, 2011.
- A. Kumar and C. Sminchisescu. Support kernel machines for object recognition. In *ICCV 2007.*, pages 1–8. IEEE, 2007.
- A. Kumar, P. Rai, and H. Daumé III. Co-regularized spectral clustering with multiple kernels. In *NIPS 2010 Workshop: New Directions in Multiple Kernel Learning*, 2010.
- A. Kumar, P. Rai, and H. Daumé III. Co-regularized multi-view spectral clustering. 2011.
- G.R.G. Lanckriet, N. Cristianini, L. El Ghaoui, P. Bartlett, and M.I. Jordan. Learning the kernel matrix with semi-definite programming. *Computer*, 2002.
- G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L.E. Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.
- H. Lappalainen and J. Miskin. Ensemble learning. *Advances in Independent Component Analysis*, pages 75–92, 2000.
- N.D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16:329–336, 2004.
- D.P. Lewis, T. Jebara, and W.S. Noble. Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. *Bioinformatics*, 22(22):2753–2760, 2006.
- J. Li, N. Allinson, D. Tao, and X. Li. Multitraining support vector machine for image retrieval. *Image Processing, IEEE Transactions on*, 15(11):3597–3601, 2006.
- Y. Li, B. Geng, Z.J. Zha, D. Tao, L. Yang, and C. Xu. Difficulty guided image retrieval using linear multiview embedding. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1169–1172. ACM, 2011.
- Y.Y. Lin, T.L. Liu, and C.S. Fuh. Local ensemble kernel learning for object category recognition. In *2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.
- C. Liu and P.C. Yuen. A boosted co-training algorithm for human action recognition. *IEEE transactions on circuits and systems for video technology*, 21(9):1203–1213, 2011.

- C. Longworth and M.J.F. Gales. Multiple kernel learning for speaker verification. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1581–1584. IEEE, 2008.
- B. Maeireizo, D. Litman, and R. Hwa. Co-training for predicting emotions with spoken dialogue data. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 28. Association for Computational Linguistics, 2004.
- E.T. Matsubara, M.C. Monard, and G.E. Batista. Multi-view semi-supervised learning: An approach to obtain different views from text datasets. In *Proceeding of the 2005 conference on Advances in Logic Based Intelligent Systems: Selected Papers of LAPTEC 2005*, pages 97–104. IOS Press, 2005.
- R. Memisevic. Kernel information embeddings. In *Proceedings of the 23rd international conference on Machine learning*, pages 633–640. ACM, 2006.
- R. Memisevic, L. Sigal, and D.J. Fleet. Shared kernel information embedding for discriminative inference. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(4):778–790, 2012.
- R. Mihalcea. Co-training and self-training for word sense disambiguation. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2004)*, 2004.
- C. Müller, S. Rapp, and M. Strube. Applying co-training to reference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 352–359. Association for Computational Linguistics, 2002.
- I. Muslea, S. Minton, and C.A. Knoblock. Selective sampling with co-testing. In *The CRM Workshop on "Combining and Selecting Multiple Models With Machine Learning"*, 2000.
- I. Muslea, S. Minton, and C.A. Knoblock. Active+ semi-supervised learning= robust multi-view learning. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 435–442, 2002a.
- I. Muslea, S. Minton, and C.A. Knoblock. Adaptive view validation: A first step towards automatic view detection. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 443–450. Citeseer, 2002b.
- I. Muslea, S.N. Minton, and C.A. Knoblock. Active learning with strong and weak views: A case study on wrapper induction. In *IJCAI*, volume 18, pages 415–420, 2003.
- I. Muslea, S. Minton, and C.A. Knoblock. Active learning with multiple views. *Journal of Artificial Intelligence Research*, 27(1):203–233, 2006.
- K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 86–93. ACM, 2000.

- W. Phillips and E. Riloff. Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 125–132. Association for Computational Linguistics, 2002.
- D. Pierce and C. Cardie. Limitations of co-training for natural language learning from large datasets. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 1–9, 2001.
- N. Quadrianto and C.H. Lampert. Learning multi-view neighborhood preserving projections. In *Proc. of the the International Conference on Machine Learning (ICML)*, pages 425–432, 2011.
- A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. More efficiency in multiple kernel learning. In *Proceedings of the 24th international conference on Machine learning*, pages 775–782. ACM, 2007.
- A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 2008.
- M. Salzmann, C.H. Ek, R. Urtasun, and T. Darrell. Factorized orthogonal latent spaces. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, 2010.
- T. Scheffer. Email answering assistance by semi-supervised text classification. *Intelligent Data Analysis*, 8(5):481–493, 2004.
- Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.
- H.S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.
- A. Sharma, A. Kumar, H. Daume III, and D.W. Jacobs. Generalized multiview analysis: A discriminative latent space. 2012.
- A. Shon, K. Grochow, A. Hertzmann, and R. Rao. Learning shared latent structure for image synthesis and robotic imitation. *Advances in Neural Information Processing Systems*, 18:1233, 2006.
- L. Sigal, R. Memisevic, and D.J. Fleet. Shared kernel information embedding for discriminative inference. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2852–2859. IEEE, 2009.
- V. Sindhwani and D.S. Rosenberg. An rkhs for multi-view learning and manifold co-regularization. In *Proceedings of the 25th international conference on Machine learning*, pages 976–983. ACM, 2008.
- V. Sindhwani, P. Niyogi, and M. Belkin. A co-regularization approach to semi-supervised learning with multiple views. In *Workshop on Learning with Multiple Views at ICML 2005*. Citeseer, 2005.



- S. Sonnenburg, G. Rätsch, and C. Schäfer. A general and efficient multiple kernel learning algorithm. 2006a.
- S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565, 2006b.
- N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. *Learning Theory*, pages 169–183, 2006.
- N. Subrahmanya and Y.C. Shin. Sparse multiple kernel learning for signal processing applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):788–798, 2010.
- L. Sun, S. Ji, and J. Ye. A least squares formulation for canonical correlation analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 1024–1031. ACM, 2008.
- S. Sun, F. Jin, and W. Tu. View construction for multi-view semi-supervised learning. *Advances in Neural Networks–ISNN 2011*, pages 595–601, 2011.
- M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy. Composite kernel learning. In *Proceedings of the 25th international conference on Machine learning*, pages 1040–1047. ACM, 2008.
- M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy. Composite kernel learning. *Machine learning*, 79(1):73–103, 2010.
- D. Tao, X. Tang, X. Li, and X. Wu. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(7):1088–1099, 2006.
- Nello Cristianini Thorsten Joachims and John Shawe-Taylor. Composite kernels for hyper-text categorisation. In *Proceedings of the Eighteenth International Conference on Machine Learning*, number 250-257, 2001.
- M. Varma and B.R. Babu. More generality in efficient multiple kernel learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1065–1072. ACM, 2009.
- M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *ICCV 2007.*, pages 1–8. IEEE, 2007.
- G. Wahba. Spline models for observational data (philadelphia, pa: Siam). 1990.
- C. Wan, R. Pan, and J. Li. Bi-weighting domain adaptation for cross-language text classification. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

- X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 235–243. Association for Computational Linguistics, 2009.
- W. Wang and Z.H. Zhou. Analyzing co-training style algorithms. *Machine Learning: ECML 2007*, pages 454–465, 2007.
- W. Wang and Z.H. Zhou. A new analysis of co-training. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 1135–1142, 2010.
- Z. Wang, S. Chen, and T. Sun. Multik-mhks: a novel multiple kernel learning algorithm. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):348–353, 2008.
- Z. Wang, S. Chen, and D. Gao. A novel multi-view learning developed from single-view patterns. *Pattern Recognition*, 2011.
- B. Wei and C. Pal. Cross lingual adaptation: An experiment on sentiment classifications. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 258–262. Association for Computational Linguistics, 2010.
- T. Xia, D. Tao, T. Mei, and Y. Zhang. Multiview spectral embedding. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 40(6):1438–1446, 2010.
- B. Xie, Y. Mu, D. Tao, and K. Huang. m-sne: Multiview stochastic neighbor embedding. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 41(4):1088–1096, 2011.
- Z. Xu, R. Jin, I. King, and M.R. Lyu. An extended level method for efficient multiple kernel learning. *Advances in neural information processing systems*, 21:1825–1832, 2009a.
- Z. Xu, R. Jin, J. Ye, M.R. Lyu, and I. King. Non-monotonic feature selection. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1145–1152. ACM, 2009b.
- Z. Xu, R. Jin, H. Yang, I. King, and M.R. Lyu. Simple and efficient multiple kernel learning by group lasso. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1175–1182, 2010.
- R. Yan and M. Naphade. Semi-supervised cross feature learning for semantic concept detection in videos. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 657–663. IEEE, 2005.
- Y. Ying and C. Campbell. Generalization bounds for learning the kernel. 2009.
- J. Yu, D. Liu, D. Tao, and H.S. Seah. On combining multiple features for cartoon character retrieval and clip synthesis. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics: a publication of the IEEE Systems, Man, and Cybernetics Society*, 2012a.

- J. Yu, M. Wang, and D. Tao. Semi-supervised multiview distance metric learning for cartoon synthesis. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 2012b.
- S. Yu, K. Yu, V. Tresp, and H.P. Kriegel. Multi-output regularized feature projection. *Knowledge and Data Engineering, IEEE Transactions on*, 18(12):1600–1613, 2006.
- S. Yu, B. Krishnapuram, R. Rosales, and R.B. Rao. Bayesian co-training. In *NIPS'07*, pages –1–1, 2007.
- S. Yu, B. Krishnapuram, R. Rosales, and R.B. Rao. Bayesian co-training. *The Journal of Machine Learning Research*, 999888:2649–2680, 2011.
- M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.
- D. Zhai, H. Chang, S. Shan, X. Chen, and W. Gao. Multiview metric learning with global consistency and local smoothness. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):53, 2012.
- L. Zhang, D. Tao, and X. Huang. On combining multiple features for hyperspectral remote sensing image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, (99):1–15, 2012.
- T. Zhang, D. Tao, X. Li, and J. Yang. Patch alignment for dimensionality reduction. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1299–1313, 2009.
- W. Zheng, X. Zhou, C. Zou, and L. Zhao. Facial expression recognition using kernel canonical correlation analysis (kcca). *Neural Networks, IEEE Transactions on*, 17(1):233–238, 2006.
- Z.H. Zhou and M. Li. Semi-supervised regression with co-training. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2005a.
- Z.H. Zhou and M. Li. Tri-training: Exploiting unlabeled data using three classifiers. *Knowledge and Data Engineering, IEEE Transactions on*, 17(11):1529–1541, 2005b.
- Z.H. Zhou, K.J. Chen, and Y. Jiang. Exploiting unlabeled data in content-based image retrieval. *Machine Learning: ECML 2004*, pages 525–536, 2004.
- Z.H. Zhou, D.C. Zhan, and Q. Yang. Semi-supervised learning with very few labeled training examples. In *Proceedings of the national conference on artificial intelligence*, volume 22, page 675. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.
- X. Zhu, Z. Huang, H. Tao Shen, J. Cheng, and C. Xu. Dimensionality reduction by mixed kernel canonical correlation analysis. *Pattern Recognition*, 2012.