Learning with Weak Views Based on Dependence Maximization Dimensionality Reduction

Qing Zhang[1,2], De-Chuan Zhan[2,3,⋆], and Yilong Yin[1]

[1] School of Computer Science and Technology,
Shandong University, Jinan, 250101, China
[2] National Key Laboratory for Novel Software Technology,
Nanjing, 210046, China
[3] Shenzhen Key Laboratory of High Performance Data Mining,
Shenzhen, 518055, China
{zhangqing2008,ylyin}@sdu.edu.cn,
zhandc@nju.edu.cn

Abstract. Large number of applications involving multiple views of data are coming into use, e.g., reporting news on the Internet by both text and video, identifying a person by both fingerprints and face images, etc. Meanwhile, labeling these data needs expensive efforts and thus most data are left unlabeled in many applications. Co-training can exploit the information of unlabeled data in multi-view scenarios. However, the assumptions of co-training, i.e., sufficient and redundant are so strong to be held in most situations. It is notable that different views often have different discrimination ability, while views with strong discrimination ability are usually hard to be obtained. As a consequence, it is a promising way to exploit unlabeled multi-view training data to integrate the information of the strong view into the weak view so that the weak view's discrimination ability can get improved. Only classifiers trained on the weak view will be used to do the classification tasks afterwards. In this paper, based on dependence maximization, we propose a framework to inject the information of strong views into weak ones. Experiments show that the framework outperforms co-training in improving the performances of classifiers trained on the weak view.

Keywords: multi-view, dimensionality reduction, semi-supervised learning, co-training.

1 Introduction

Many real applications involve more than one modal of data, and abundant data with multiple views are at hand. A representative example is that the Internet news on web pages are always presented by text, audio, video, simultaneously. For real applications, unlabeled training data are readily available but labeled ones are fairly expensive to be obtained because labeling unlabeled data requires expensive human efforts and time. Semi-Supervised Learning (SSL) [1] methods were proposed to make use of the unlabeled data. With single view data, the classifier trained on one view tries to exploit the

⋆ Corresponding author.

J. Yang, F. Fang, and C. Sun (Eds.): IScIDE 2012, LNCS 7751, pp. 557–564, 2013.

unlabeled data by itself. In semi-supervised scenario, when data are presented by multiple views, co-training [2] is proposed to exploit the disagreements between the multi-views so as to improve views' generalization abilities. However, Blum and Mitchell [2] pointed out that in co-training style methods, the data should be assumed to have two sufficient and redundant views, where each view is sufficient for training a strong classifier and the views are conditionally independent given the class label. Nevertheless, in real-world applications, it is rarely that the two views are conditionally independent given the class label.

There is a ubiquitous fact in multi-view data to which researchers have never paid much attention, that is, different views always have different discrimination ability. Views can be categorized into strong views which have strong discrimination ability and weak views which are with weak discrimination ability. Classifiers trained on strong views usually have better performance (generalization ability) than those trained on weak views. For example, fingerprints are accurate and robust in identifying a person, while face images show less discrimination ability and are vulnerable to variances. It seems that we can rely on strong views only, however, in real applications, strong view data are always harder to be acquired. E.g., to capture a person's fingerprints needs expert devices and more human efforts than to take the person's face photographs. As a consequence, it is expected that only weak view data are used for the classification task, e.g., face recognition is preferred in human identification. Many efforts have been made to raise the performance of the weak view, e.g. [5-6]. However, there has hardly been any researches so far concerning integrating information of the strong view into the weak view.

In this paper, we propose a framework to improve the weak view in semi-supervised scenario. This framework attempts to exploit unlabeled multi-view training data to integrate information of the strong view into the weak view. Dependence maximization is used to make information of strong and weak views mostly aligned. A dimensionality reduction method DMDR (Dependence Maximization Dimensionality Reduction method) is proposed to project the weak view data to a lower-dimensional space in which the dependence of information from both strong and weak views is maximized. We verify the effectiveness of the proposed framework by experiments on real and synthetic datasets. The proposed framework outperforms co-training in promoting the classifiers trained on the weak view.

This paper is organized as follows. Section 2 shows some related works. Section 3 describes the proposed framework in detail. Section 4 shows the experiments on real and synthetic datasets. Section 5 concludes the paper and points out the future work.

2 Related Works

Researches on multi-view data learning are mostly concentrated on how to exploit unlabeled information in semi-supervised scenarios. A representative multi-view semi-supervised learning approach is the co-training approach [2] which works with two views. Co-training initializes a classifier on each view, respectively, using the original labeled data. Then, each of the two classifiers selects and labels a certain number of highly-confident unlabeled instances to refine the other classifier. Blum and Mitchell

[2] proved that if the two views are sufficient and redundant, the predictive accuracy of an initial weak classifier can be boosted to arbitrarily high using unlabeled data by co-training. Zhou et al. [7] also showed that, with sufficient and redundant views, it is possible to execute an effective semi-supervised learning with a single labeled training example. In real-world tasks, however, the requirement of sufficient and redundant views is too luxury. Thus, researchers tried to find relaxed conditions for co-training style methods that work in real situations. Abney [3] showed that the two views are not needed to be conditionally independent, and a weak independence assumption is sufficient. Balcan et al. [4] proved that even the weak independence is not needed if PAC classifiers can be obtained on each view, and a weaker assumption of expansion of the underlying data distribution is sufficient. In spite of these findings, co-training can only get effective results under certain assumptions.

In this paper, we emphasize the fact that in multi-view data, different views always have different discrimination ability. Since strong view data are harder to be obtained, it is necessary and important to improve the weak view. Our work is related to dimensionality reduction as the core of the proposed framework to improve the weak view is a dimensionality reduction method for the weak view. Traditional dimensionality reduction methods can be classified into supervised or unsupervised, depending on whether the label information is used or not. A representative unsupervised method is PCA [8], and advances include ISOMAP [9] , ICA [10] , LPP [11] , LLE [12] , etc. Representative supervised dimensionality reduction methods are LDA [13] , PLS [14] , etc. All these methods work on a single view, while the proposed method in this paper takes two views into consideration. The dimensionality reduction method CCA [15] also concerns aligning information of two views. However, since our work concentrates on the promotion of the weak view, it seems more straightforward to only inject information of the strong view into the weak one. Actually, we testified in our experiments that injecting information of both views into each other somehow inhibits the promotion of the weak view. The proposed dimensionality reduction method can be categorized into unsupervised method since no label information is needed.

3 Proposed Framework

In this section, we firstly introduce the DMDR method followed by the full picture of the framework.

3.1 DMDR Method

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the original feature space of the weak view and the strong view, respectively. Labeled data are presented by$\{(\boldsymbol{x}_1, \boldsymbol{y}_1, l_1), (\boldsymbol{x}_1, \boldsymbol{y}_2, l_2), \cdots, (\boldsymbol{x}_n, \boldsymbol{y}_n, l_n)\}$, where $\boldsymbol{x}_i$ denotes an instance of the weak view, and $\boldsymbol{y}_i$ denotes an instance of the strong view, l_i is the corresponding label, $i = 1, 2, \cdots, n$. Unlabeled data are presented by $\{(\boldsymbol{x}_{n+1}, \boldsymbol{y}_{n+1}), (\boldsymbol{x}_{n+2}, \boldsymbol{y}_{n+2}), \cdots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$, where n is the number of labeled data, and N is the total number of labeled and unlabeled data. The total dataset is presented by $Dn = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \cdots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$. By assuming that the weak view contains discriminative information implicitly, we need to extract these discriminations with the

supervision of the strong view. In detail, we attempt to find a lower-dimensional feature space for the weak view in which the dependence of information of both strong and weak views is maximized. So that, the lower-dimensional feature space can stress the discriminations more explicitly and classifiers built on the feature space can be with better performances. By denoting the projection vector of the weak view data as $\boldsymbol{p}$, an instance $\boldsymbol{x}$ is projected into a new space $\mathcal{F}$ by $\phi(\boldsymbol{x}) = \boldsymbol{p}^\top \boldsymbol{x}$ and the deduced kernel function is $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle = \langle \boldsymbol{p}^\top \boldsymbol{x}_i, \boldsymbol{p}\top \boldsymbol{x}_j \rangle$. For instances of the strong view, we define the kernel function $\ell(\boldsymbol{y}_i, \boldsymbol{y}_j) = \langle \boldsymbol{y}_i, \boldsymbol{y}_j \rangle$. Given the dataset Dn with joint distribution $P_{\boldsymbol{xy}}$, we define the kernel matrix for the weak view and the strong view as $\mathbf{K} = [\kappa_{ij}]_{N\times N}, \kappa_{ij} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\mathbf{L} = [\ell_{ij}]_{N\times N}, \ell_{ij} = \ell(\boldsymbol{y}_i, \boldsymbol{y}_j)$, respectively. Then, we try to maximize the dependence of the weak view data in the projected feature space $\mathcal{F}$ with the strong view data.

For facilitating the dependence maximization and make full use of the potential nonlinearities in both views, we define a dependence between the kernels of different views as

$$\mathfrak{D}(\mathbf{K}, \mathbf{L}) = tr(\mathbf{KL}) \tag{1}$$

by assuming both $\mathbf{K}$ and $\mathbf{L}$ are centralized and normalized. For general kernels of data, if we define $\mathbf{H} = \mathbf{I} - \frac{1}{N} \times \mathbf{e}\mathbf{e}^\top$, where $\mathbf{I}$ is an identity vector and $\mathbf{e}$ is an all-one column vector, the equation above becomes

$$\mathfrak{D}(\mathbf{K}, \mathbf{L}) = tr(\mathbf{HKHL}) \tag{2}$$

where $\mathbf{H}$ can be regarded as a centralized operator. Our dependence criterion is closely related to a kind of independence criterion called Hilbert-Schmidt Independence Criterion [16] . The dependence criterion computes the square of the norm of the cross-covariance operator over the domain $\mathcal{X} \times \mathcal{Y}$ in Hilbert Space. Due to the neat theoretical properties, we maximize the dependence of the information of both views, i.e.,

$$\max \quad \mathfrak{D}(\mathbf{K}, \mathbf{L}) \quad = \quad \max \quad tr(\mathbf{HKHL}) \tag{3}$$

By representing the instances in $\mathcal{X}$ as $\phi(\boldsymbol{x})$, we can rewrite the target function in eq. 3 as

$$\boldsymbol{p}^* = \arg\max_{\boldsymbol{p}} \quad tr(\mathbf{H}\mathbf{X}^\top \boldsymbol{p}\boldsymbol{p}^\top \mathbf{XHL}) \tag{4}$$

To avoid the scaling problem, we add the constraint that the $l_2 - norm$ of $\boldsymbol{p}$ should be bounded. Therefore, we reformulate the optimization problem as

$$\boldsymbol{p}^* = \arg\max_{\boldsymbol{p}} \quad tr(\mathbf{H}\mathbf{X}^\top \boldsymbol{p}\boldsymbol{p}^\top \mathbf{XHL})$$

$$s.t. \qquad \boldsymbol{p}^\top \boldsymbol{p} = 1$$

Notice that

$$tr(\mathbf{H}\mathbf{X}^\top \boldsymbol{p}\boldsymbol{p}^\top \mathbf{XHL}) = \boldsymbol{p}^\top (\mathbf{XHLH}\mathbf{X}^\top)\boldsymbol{p} \tag{5}$$

Since $\mathbf{XHLH}\mathbf{X}^\top$ is symmetric, the eigenvalues are all real. Without any loss of generality, we can assume that the eigenvalues of $\mathbf{XHLH}\mathbf{X}^\top$ are sorted as $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_D$. Thus, if d is the dimensionality of the new feature space, the optimal projection

matrix $\boldsymbol{P}^*$ can be defined as $\boldsymbol{P}^* = [\boldsymbol{p}_1^*, \boldsymbol{p}_2^*, \cdots, \boldsymbol{p}_d^*]$, where $\boldsymbol{p}_i^*$ is the normalized eigenvector corresponding to the i-th largest eigenvalue λ_i, $i = 1, \cdots, d$ $(d \ll D)$. Since the eigenvalues reflect the contribution of the corresponding dimensions, we can control d by setting a threshold $thr (0 \leq thr \leq 1)$ and then choose the first d eigenvectors such that

$$\sum_{i=1}^{d} \lambda_i \geq thr \times \sum_{i=1}^{D} \lambda_i \tag{6}$$

3.2 Framework Summarization

To give a clear picture of the framework, we summarize the framework in Algorithm 1.

Algorithm 1. The proposed framework

1. Get a training dataset of two views $Dn = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ including labeled and unlabeled data.
2. Perform DMDR and project the weak view data to a lower-dimensional space.
3.1 Train a classifier $\mathcal{A}$ by labeled weak view data.
3.2 Train a classifier $\mathcal{B}$ by labeled strong view data.
4. Use $\mathcal{B}$ to predict labels for unlabeled data.
5. Re-train $\mathcal{A}$.

4 Experiments

In experiments, the framework is compared with co-training in the performances of the classifiers trained on the weak view. Five dimensionality reduction methods PCA, ICA, ISOMAP, LPP, and LLE are used in co-training for comparison. CCA is also adopted by the framework to compare with DMDR. Binary SVM is used as the basic classifier.

4.1 Datasets

We use three multi-view datasets in our experiments - two real datasets and a synthetic dataset constructed from a real single view dataset. For each dataset, two views are selected as the strong view and the weak view according to their discrimination ability. The two real datasets are Ads dataset [17] and WebKB dataset [18] . The Ads dataset has five views from which two views are selected. The WebKB dataset contains two views naturally. The synthetic two-view dataset is constructed from the Newsgroup dataset [19] , in which the first view is selected by PCA, and the second view is constructed by selecting 700 features from all features randomly. Classification problems are confined to be two class problems. The WebKb dataset is processed to include two classes - one class is course page of 230 instances, and the other is non-course page of 821 instances. For Newsgroup dataset, we choose the first two classes among all the 20 classes.

4.2 Configuration

The average accuracies of SVMs trained on different views in each dataset evaluated by 10 times10-Crosses Validation (CV) are listed in Table 1. Results show that no matter

Table 1. Accuracies of SVMs trained on different views in each dataset

Datasets	Views	Accuracy of SVMs		
		Euclidean Space	PCA	LLE
Ads	View1	89.0%	93.8%	89.1%
	View2	86.8%	91.2%	86.3%
	View3	91.1%	95.9%	94.9%
	View4	84.2%	89.0%	89.3%
	View5	81.7%	86.7%	88.6%
WebKB	View1	75.4%	82.4%	86.9%
	View2	89.2%	95.1%	92.9%
Newsgroup	View selected by PCA	82.8%	88.7%	93.6%
	View selected randomly	74.8%	78.0%	72.1%

in the original Euclidean space or space selected by linear and nonlinear dimensionality reduction methods, classifiers trained on some views always have better performances than those trained on some other views. In the Ads dataset, View 3 presents anchor text attached to hyper-links pointing to a web page and shows the best discriminative ability; View 5 presents the caption of a web page and is the weakest view in all situations. The situation is the same for the WebKB dataset. It indicates that strong views are always harder to be obtained. Gathering information from other web pages which link to a web page is always harder than getting information from a web page itself. In Newsgroup dataset, the view selected by PCA is always stronger than the view selected randomly.

Table 2. Configurations in co-training

Datasets	Proportion	Labeled data		Unlabeled data labeled in each round	
		Positive Class	Negative Class	Positive Class	Negative Class
Ads	1:6	5	30	1	6
WebKB	1:5	5	25	1	5
Newsgroup	1:1	50	50	5	5

In our framework, 50 instances of each class in the training set are selected as the labeled data, the rest data are treated as unlabeled. In co-training, the number of labeled data and the number of unlabeled data labeled by each classifier in each round in the training set is decided by the proportion of data in two classes. The numbers are listed in Table 2. Co-training process goes until all unlabeled data are labeled.

4.3 Experimental Results

In the experiment, both the proposed framework and co-training are performed by 10 times 10-CV. Average accuracies of SVMs trained on the weak view of each dataset are listed in Table 3. The listed results in co-training are the results obtained in the last round. Performances obtained in the proposed framework with DMDR are higher than the best results in co-training. For the proposed framework, using CCA results in lower performance than using DMDR which implies injecting information of both views into each other will inhibit the promotion of the weak view.

Table 3. Accuracies of SVMs trained on weak view in each dataset

Datasets	Proposed framework		Co-training				
	DMDR	CCA	PCA	ICA	ISOMAP	LPP	LLE
Ads	**86.7%**	86.6%	86.2%	85.9%	86.1%	85.9%	86.1%
WebKB	**84.2%**	79.2%	79.0%	73.9%	78.1%	78.1%	81.0%
Newsgroup	**71.9%**	59.8%	50.0%	45.5%	45.1%	46.5%	47.5%

5 Conclusion and Future Work

Nowadays, abundant multi-view data are available. Meantime, labeling data needs expensive human efforts and time, and more data at hand are left unlabeled. In this paper, we emphasize a ubiquitous fact in multi-view data that different views always have different discrimination ability. In most situations, more efforts will be paid to collect strong view data, so it is often expected that only weak view data are used for classification. As a result, it is necessary and important to improve the discrimination ability of the weak view. A framework is proposed to improve the weak view through the help of the strong view in semi-supervised mode. We show the superiority of the framework by experiments. In the future work, we will employ the framework on real applications such as biometric recognition and try to extend it for data with more than two views.

Acknowledgements. This research is supported by Shenzhen Key Laboratory for High Performance Data Mining with Shenzhen New Industry Development Fund under grant No.CXB2010052 50021A, NSFC under grant No. 61105043, Baidu Open Research Fund and National Natural Science Foundation of China under Grant No.61070097.

References

1. Chapelle, O., Schölkopf, B., Zien, A.: Semi–Supervised Learning. MIT Press, Cambridge (2006)
2. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co–training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory, Madison, WI, pp. 92–100 (1998)

3. Abney, S.: Bootstrapping. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, pp. 360–367 (2002)
4. Balcan, M.F., Blum, A., Yang, K.: Co-training and expansion: Towards bridging theory and practice. In: Saul, L., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 17, pp. 88–96. MIT Press, Cambridge (2004)
5. Roli, F., Didaci, L., Marcialis, G.L.: Template Co-update in Multimodal Biometric Systems. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 1194–1202. Springer, Heidelberg (2007)
6. Roli, F., Didaci, L., Marcialis, G.L.: Adaptive biometric systems that can improve with use. In: Ratha, N., Govindaraju, V. (eds.) Advances in Biometrics: Sensors, Systems and Algorithms, vol. 3, pp. 447–471. Springer, Heidelberg (2008)
7. Zhou, Z.-H., Zhan, D.-C., Yang, Q.: Semi-supervised learning with very few labeled training examples. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence, Vancouver, Canada, pp. 675–680 (2007)
8. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (1986)
9. Tenenbaum, J.B., Desilva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
10. Comon, P.: Independent component analysis-A new concept? Signal Processing 36(3), 287–314 (1994)
11. He, X., Niyogi, P.: Locality Preserving Projections. In: Sebastian, T., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems 16, Vancouver, Canada, pp. 153–160 (2003)
12. Roweis, S.T., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
13. Fisher, R.: The use of multiple measurements in taxonomic problems. Annals of Eugenics 7(2), 179–188 (1936)
14. Wold, H.: Partial least squares. Encyclopedia of the Statistical Sciences 6, 581–591 (1985)
15. Hardoon, D., Szedmak, S., Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Computing 16(12), 2639–2664 (2004)
16. Gretton, A., Bousquet, O., Smola, A.J., Schölkopf, B.: Measuring Statistical Dependence with Hilbert-Schmidt Norms. In: Jain, S., Simon, H.U., Tomita, E. (eds.) ALT 2005. LNCS (LNAI), vol. 3734, pp. 63–77. Springer, Heidelberg (2005)
17. Kushmerick, N.: Learning to remove internet advertisements. In: Proceedings of the 3rd Annual Conference on Autonomous Agents, Seattle, WA, pp. 175–181 (1999)
18. Craven, M., DiPasquo, D., Freitag, D., McCallm, A., Mitchell, T., Nigam, K., Slattery, S.: Learning to extract symbolic knowledge from the world wide web. In: Proceedings of 15th National Conference on Artificial Intelligence, Madison, WI, USA, pp. 509–516 (1998)
19. Joachims, T.: A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. Computer Science Technical Report CMU-CS-96-118. Carnegie Mellon University (1996)