Transductive Optimal Component Analysis

Yuhua Zhu[1], Yiming Wu[1], Xiuwen Liu[1], and Washington Mio[2]
[1]*Department of Computer Science* [2]*Department of Mathematics*
Florida State University, Tallahassee, FL 32306
{zhu,ywu,liux}@cs.fsu.edu *mio@math.fsu.edu*

Abstract

We propose a new transductive learning algorithm for learning optimal linear representations that utilizes unlabeled data. We pose the problem of learning linear representations as an optimization one on the underlying nonlinear manifold. An additional term is used to prefer representations with large "margins" when classifying unlabeled data in the nearest classifier sense, a generalization of transductive support vector machines to learning representations. Experimental results of the proposed algorithm on face recognition data sets show the potential significant improvement for classification accuracy on test sets.

1. Introduction

In recent years, due to dramatic increase in availability of (unlabeled) data, semi-supervised learning methods, such as EM with generative mixture models [6], self-training [7], co-training [5], transductive support vector machines [8], etc., have been proposed to utilize unlabeled data during training to improve the generalization performance of the resulting classifiers. These methods are attractive for many real-world applications as labeling examples is often expensive while unlabeled data are readily available. A particular example among those methods is transductive support vector machine (TSVM) [8]. Transductive learning can be effective in that the learning algorithm can estimate probability distributions of the unlabeled examples in the testing set and can potentially exploit the structure for better generalization performance. For example, TSVM seeks the largest separation between labeled and unlabeled data through regularization by imposing large margins for both labeled and unlabeled data. TSVM has been used widely in classification, such as in [2, 9]

Optimal component analysis (OCA) [4] is a stochastic gradient algorithm that poses the problem of learning optimal linear representations for a particular recognition task as an optimization one. A recognition performance function is specified based on the nearest neighbor classifier. The search for an optimal representation is to maximize a specified performance function over all subspaces of a Grassmann manifold. The solution is obtained by conducting a stochastic searching algorithm utilizing intrinsic geometry structure of the underlying manifolds. OCA provides a computational framework for finding optimal linear representations for particular applications and its effectiveness has been demonstrated in many applications.

In this paper we propose a transductive version of OCA by generalizing the idea of transductive SVM. In OCA, for the labeled training data, the specified performance measure function, which will be referred as objective performance function in the following sections, remains the same as in [4]. However, we add an additional term for unlabeled test data. Since the labels of the test samples are not known, we will enforce a large margin by forcing the test sample to be close to the class of the closest samples and at the same time to be far away from other classes. This corresponds to transductive SVM, where unlabeled test samples are forced to be away from the decision boundaries.

The rest of the paper is organized as follows. Section 2 gives a brief description on the optimal component analysis and Section 3 shows the formulation of TOCA. Experiment results are presented in Section 4 and Section 5 concludes the paper.

2. Overview of Optimal Component Analysis

Compared to PCA, ICA and LDA, OCA has shown its advantages in solving object recognition problems on commonly used datasets. More specifically, let $U \in \Re^{n\times d}$ be an orthonormal basis of a d-dimensional subspace of $\Re^n$, where n is the size of the input image and d is the desired dimension of the resulting subspace (generally $n \gg$d). For an image I, considered as a column vector of size n, the vector of coefficients is given by $\alpha(I,U) = U^T I \in \Re^d$. The performance function F is defined in the following way: let there be C classes to be recognized from the images, where each class has k_{train} training images (denoted by

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

$I_{c,1}, \ldots, I_{c,k_{train}}$) and k_{cross} cross validation images (denoted by $I'_{c,1}, \ldots, I'_{c,k_{cross}}$), we define a performance function by

$$F(U) = \frac{1}{Ck_{cross}} \sum_{c=1}^{C} \sum_{i=1}^{k_{cross}} h(\rho(I'_{c,i}, U) - 1). \quad (1)$$

where h(·), a monotonically increasing bounded function, is used to control bias with respect to particular classes in measurements of performance; in our implementation, we use $h(x) = 1/(1 + \exp(-2\beta x))$, where β controls the degree of smoothness of $F(U)$. In Eq. (1), ρ is given by

$$\rho(I'_{c,i}, U) = \frac{\min_{c' \neq c, j} D(I'_{c,i}, I_{c',j}; U)}{\min_j D(I'_{c,i}, I_{c,j}; U) + \epsilon}, \quad (2)$$

which measures the ratio between the smallest distance between $I'_{c,i}$ and all the training images in other classes and the smallest between $I'_{c,i}$ and the ones in the same class. Here $D(\cdot, \cdot; \cdot)$ is given by

$$D(I_1, I_2; U) = \|\alpha(I_1, U) - \alpha(I_2, U)\|, \quad (3)$$

where $\|\cdot\|$ denotes the 2-norm. In Eq. (2), $\epsilon > 0$ is a small number to avoid division by zero. Here ρ measures the separation between clusters given by different classes; note that $\rho(I'_{c,i}, U) > 1$ means that $I_{c,i}$ is closest to a training image in the same class and when we let $\beta \to \infty$, F is precisely the recognition performance of the nearest neighbor classifier after projection to the subspace given by U [4], Note that while Eq. (1) is defined using a separate cross validation set, it can be modified to be defined on the training set only in the leave-one-set sense.

Under this formulation, $F(U) = F(UH)$ for any $d \times d$ orthogonal matrix H as the distance $D(I_1, I_2; U) = D(I_1, I_2; UH)$; the choice of 2-norm in $D(I_1, I_2; U)$ allows for this equality. In other words, F depends on the subspace spanned by U but not on the specific basis chosen to represent that subspace. Therefore, our search for optimal representation is on the space of d-dimensional subspace rather than the basis.

The Grassmann manifold, $\mathcal{G}(n, d)$, is the set of all d-dimensional subspaces of $\Re^n$ [1]. It is a compact, connected manifold of dimension $d(n - d)$, which can be represented either by a basis (non-uniquely) or by a projection matrix (uniquely). Choosing the former, let U be an $n \times d$ matrix whose columns are an orthonormal basis for the given subspace of $\Re^n$ and let $[U]$ denote the set of all the orthonormal bases of $span(U)$, i.e., $[U] = \{UH | H \in \Re^{d \times d}, H^T H = I_d\} \in \mathcal{G}(n, d)$. Unlike the actual recognition performance, $F(U)$ is smooth and thus allows us to use gradient-type algorithm to solve the optimization problem. An optimal d-dimensional subspace for the given classification problem from the viewpoint of the available data is given by

$$\hat{U} = \arg \max_{U \in \mathcal{G}_{n,d}} F_{cross}(U) \quad (4)$$

Since the Grassmann manifold $\mathcal{G}(n, d)$ is a curved space, as opposed to being a (flat) vector-space, the gradient process has to account for its intrinsic geometry. In [4], an optimization algorithm utilizing the geometric properties of the manifold is presented. A Monte Carlo version of a simulated annealing type stochastic gradient-based algorithm is used to find an optimal subspace $\hat{U}$.

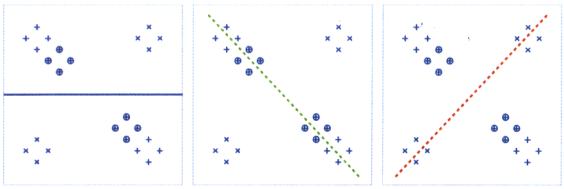
Note that a key advantage of the optimal component analysis compared to principal component analysis and Fisher discriminant analysis is that it allows incorporation of additional constraints on the resulting representation. In this paper we use an additional constraint for utilizing unlabeled data.

3. Transductive Optimal Component Analysis

The key problem in transductive learning is how to make use of unlabeled data effectively. Transductive SVM, for examples, attempts to maximize the margin by forcing the linear decision boundary away from not only labeled data but also unlabeled data. In a loose sense, TSVM attempts to improve the classification accuracy of the test set by imposing a larger margin from the decision boundary. Compared to TSVM OCA algorithm, as described in the previous section, while attempting to learn better representations, can also be interpreted as maximizing the "margin" in the nearest classifier sense. As shown Eq. (1), $F(U)$ attempts to maximize the ratio between the minimum distance to all other classes and the minimum distance within the same class. The larger the ratio, the better confidence for classification and therefore a larger margin for classification.

With this observation, we propose to generalize the OCA algorithm for transductive learning similar to transductive SVM. To illustrate the potential benefit for transductive OCA, Fig. 1 shows a simple example with two classes, where $+$ and $\times$ are labeled data of two different categories and $\oplus$ represents unlabeled data. If we apply PCA or FDA analysis on the training data, the resulting 1-dimensional basis is shown in Fig. 1(a), which gives the worst classification performance using nearest or other classifiers as two classes will be mixed together. If we apply the original OCA algorithm, both the 1-dimensional bases in Fig. 1(b) and (c) will work well, resulting large separations between two classes. However, if we include the unlabeled data, Fig. 1(c) is a better choice since unlabeled data will have a larger "margin" in that sense they will be much closer to the '+' class than to the 'x' class in the one dimensional representation.

To be more specific, in the proposed transductive optimal component analysis algorithm, we make use of both training set and unlabeled testing set. For the labeled train set, the objective performance function uses the leave-one-out version of the function given in Eq.(1). In other words,



(a) (b) (c)

Figure 1. Different 1-dimensional representation for a two-class dataset. (a) 1-dimensional representation by principal component analysis and Fisher discriminant analysis. (b) A potential solution for OCA but not for TOCA. (c) A solution for TOCA

each labeled training example is left one once as the "cross validation" image and this will be done for all the images. In order to utilize the unlabeled data, we introduce an additional constraint term based on the unlabeled test set. This term is given by

$$F_{test}(U) = \frac{1}{Ck_{test}} \sum_{c=1}^{C} \sum_{i=1}^{k_{test}} h(\rho(I''_{c,i}, U) - 1). \quad (5)$$

which is very similar to Eq.(1) except that image $I''_{c,1}$ here belongs to test set. Now the problem is how to calculate ρ in Eq. (2). Here, ρ measures the "margin" in the nearest neighbor sense. As their true labels are unknown, for each test sample, we first find the closest labeled training example and use the dominant label as if it were a true label. To control the relative contribution of the transductive term, the performance functions on the training and test set are combined using

$$F(U) = (1 - w) * F_{train}(U) + w * F_{test}(U), \quad (6)$$

which gives the criterion for transductive OCA. The value of transductive weight w must fall in $[0, 1)$. Eq.(6) is equal to Eq.(1), when w is assigned value 0.

4. Experimental Results

We have applied the transductive OCA algorithm to the search for optimal linear basis on several datasets. Due to the limitation of space, we report the experiments on AR face dataset[1] and ORL face recognition data set[2].

The AR face dataset is a difficult dataset as the images contain significant variations in each class. Fig. 2 shows selected images of a particular subject. It is clear that the variations within the class are very large due to sunglasses, scarf, view, and light changes. The difficulty is also evident that the principal component analysis with the nearest neighbor classifier gives only 67% recognition performance on the test set as shown in Fig. 3(a), where the initial basis is given by principal component analysis.

In the first experiment, we use a subset of 5 classes in AR dataset, where the dimension is reduced from 165 to 5 and transductive weight w is set to 0.2. As shown in Fig. 3(a), the stochastic search process is effective in finding better representations according to $F(U)$. The learning process improves both the performance on the training and unlabeled set as well as the recognition on the test set. The highest $F(U)$ is achieved at iteration 381, with 98% classification accuracy on the training set (in the leave-one-out sense) and 93.3% classification accuracy on the test set. Compared to the initial classification, the improvement is over 25%. Note that the transductive learning only relies on the labeled training set and unlabeled test set. This example shows clearly that the transductive learning improves the generalization performance of the nearest neighbor classifier by learning better and more robust linear representations.

We also have applied the transductive learning on the entire ORL dataset. In this case, 30% of the images are used as traning, the number of which is significantly smaller than that for test. As on the AR dataset, the transductive learning improves significantly the performance on the training (in the leave-one-out sense) and the performance on the test. The classification accuracy on the test is improved from initially 62% to 85% after 500 iterations, an improvement of over 20%. These two examples show clearly the effectiveness of transductive learning.

5. Conclusion

In this paper we have proposed a new transductive algorithm that utilizes unlabeled data when learning optimal linear representations for recognition and classification. The experimental results on face datasets show the significant improvements. As unlabeled data are more readily available, the proposed methods can be potentially significant in improving performances of content-based image retrieval and other related applications.

One of the limitations of the proposed algorithm is that it is formulated for linear representations. However, nonlinearity can be modeled efficiently by using kernel methods [3]. This will be further investigated for transductive learning of linear representations.

Acknowledgments

This research was supported in part by the National Science Foundation grants CCF-0514743. The authors like to

[1] Available from www.cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html

[2] Available from www.cl.cam.ac.uk/research/dtg/attarchive/facesataglance.html.

Figure 2. Selected images of a particular subject in the AR dataset, showing the significant variations within the class.

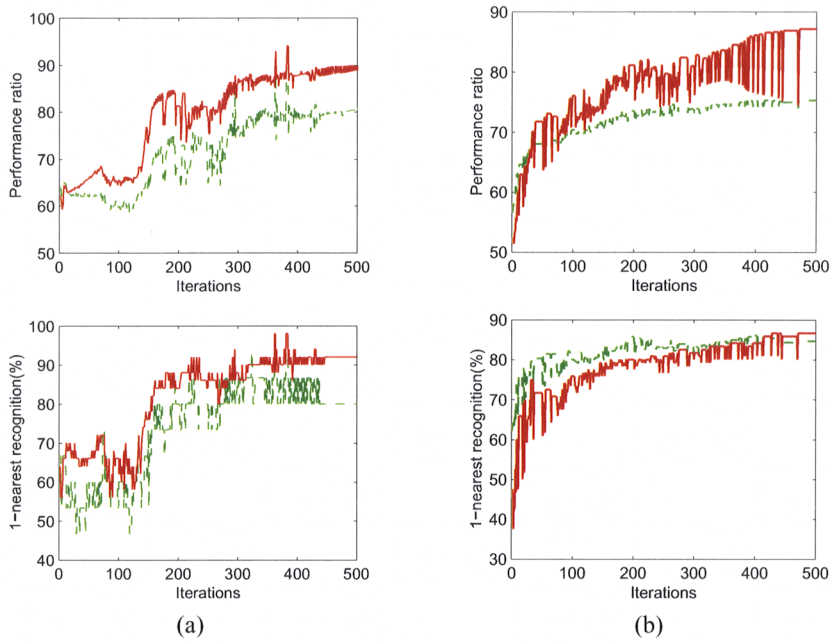


Figure 3. Evolution of performance $F(U)$ and recognition accuracy on (a) AR dataset; (b) ORL dataset. In each column, the top one shows the $F(U)$ with respect to the number of iterations and the bottom the nearest neighbor classifier accuracy. In each panel, the red solid curve shows that for $F(U)$ and classification accuracy on the training set and the green dashed curve shows that on the test set.

thank the producers of AR and ORL datasets for making them publicly available.

References

[1] A. Edelman, T. Arias, and S. T. Smith. The Geometry of Algorithm with orthogonality Constraints. *Matrix Analysis and Applications*, 20(2):303–353, 1998.

[2] T. Joachims. Transductive learning via spectral graph partitioning. *Proceedings of the International Conference on Machine Learning*, pages 290–297, 2003.

[3] X. Liu and W. Mio. "kernel methods for nonlinear discriminative analysis". In *Proceedings of the International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 584–599, 2005.

[4] X. Liu, A. Srivastava, and K. Gallivan. Optimal linear Representation of Images for Object Recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(5):662–666, 2004.

[5] T. Mitchell. The role of unlabeled data in supervised learning. *Proceedings of the Sixth International Colloquium on Cognitive Science*, 1999.

[6] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. *Proceedings of the ninth international conference on Information and knowledge management*, pages 86–93, 2000.

[7] C. Rosenberg, M. Hebert, and H. Schneiderman. Semi-supervised self-training of object detection models. *Seventh IEEE Workshop on Applications of Computer Vision*, 1:29–36, 2005.

[8] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[9] L. Xu and D. Schuurmans. Unsupervised and Semi-supervised Multi-class Support Vector Machines. *The Twentieth National Conference on Artifi cial Intelligence*, 2005.