2011 Sixth International Conference on Image and Graphics

A new Graph constructor for Semi-supervised Discriminant Analysis via Group Sparsity

Haoyuan Gao, Liansheng Zhuang, Nenghai Yu
MOE-MS Key Laboratory of Multimedia Computing and Communication
University of Science and Technology of China,
Hefei 230026, P.R.China
Email: picture@mail.ustc.edu.cn

***Abstract*—Semi-supervised dimensionality reduction is very important in mining high-dimensional data due to the lack of costly labeled data. This paper studies the Semi-supervised Discriminant Analysis (SDA) algorithm, which aims at dimensionality reduction utilizing both limited labeled data and abundant unlabeled data. Different from other relative work, we pay our attention to graph construction, which plays a key role in graph based SSL methods. Inspired by the advances of compressive sensing, we propose a novel graph construction method via group sparsity,which means to constrain the reconstruct data to be sparse for each sample, and constrain the representation in each class to be quite similar. Experimental results show that our method can significantly improve the performance of SDA, and outperform state-of-the-art methods.**

***Keywords*-semi-supervised learning; graph construction; sparsest representation;**

I. INTRODUCTION

Dimensionality reduction is very important in many vision tasks such as face recognition and image retrieval. In these tasks, one is often confronted with high-dimensional data, and suffers from "curse of dimensionality". Fortunately, most high-dimensional data in these tasks often lies in a lower dimensional manifold ("bless of dimensionality") [1], [2], [3]. This leads one to consider methods of dimensionlity reduction that allow one to represent the data in a lower dimensional space.

For classification tasks, linear discriminant analysis (LDA) is one of the most popular methods for dimensionality reduction. It seeks a linear projection that simultaneously maximizes the betweens-class dissimilarity and minimizes the within-class dissimilarity to increase the class separability. To achieve promising performance, LDA requires sufficient labeled training samples. When the number of labeled samples is much smaller than the number of dimensions, LDA will suffer from the so-called small sample size (SSS) problem due to severe under-sampling of the underlying data distribution. As a result, the within-class scatter matrix is not of full rank and hence not invertible. In this case, the performance of LDA will reduce obviously, and its generalization capability on test samples cannot be guaranteed.

In many real applications, it is difficult to obtain enough labeled training data. On the other hand, large number of unlabeled data are available at very low cost. One possible solution to overcome the SSS problem is to exploit unlabeled data. Inspired by semi-supervised learning for classification, many methods have proposed to alleviate the SSS problem of LDA by utilizing both unlabeled data and labeled data. Cai [4] first proposed a semi-supervised dimensionality reduction algorithm, called Semi-supervised Discriminant Analysis (SDA). SDA exploits the local neighborhood information of data points in performing dimensionality reduction. Instead of using local manifold structure of data, Yu and Dit-Yan proposed another method called SSDA [5] using path-based similarity measure to capture global manifold structure of the data. Similarly, SMDA [6] and UDA [7] also perform LDA under semi-supervised setting through manifold regularization. Different from above methods, SSDA$_{CCCP}$ [8] was recently proposed, by exploiting label information from unlabeled data to maximize an optimality criterion of LDA. Besides SSDA$_{CCCP}$, M-SSDA$_{CCCP}$ [8] were proposed by adopting the manifold assumption. However, both SSDA$_{CCCP}$ and M-SSDA$_{CCCP}$ have no structure preserving strategy in performing LDA with the augmented labeled data, which may limit their performance. By considering both label augmenting and local structure preserving, Zhai proposed a spectral based discriminant analysis approach called STSDA [9], and achieve an impressive performance.

Though all of above methods perform semi-supervised LDA in different ways, they all model the geometric relationships between all data points in the form of a graph. Graph plays a key role in these methods. However, far little attention has been paid to the graph constructor methods. In this paper, by taking SDA [4] for an example, we investigate the performance of popular graphs in a systematic way. Furthermore, inspired by the advances of compressive sensing, we construct a novel graph called $\ell_{2,1}$-graph for SDA via group sparsity. Experiments show that our $\ell_{2,1}$-graph outperformed other popular methods for SDA.

978-0-7695-4541-7/11 $26.00 © 2011 IEEE
DOI 10.1109/ICIG.2011.82

A. Related work

There are many methods proposed for graph construction, including k-nearest neighbors (k-NN) method [10] and ϵ-ball based method. These methods often divide graph construction into two separative steps, graph adjacency construction and graph weight calculation. Graph adjacency construction computes the similarity (or distance) between any two data points using some similarity or kernel function, and generates an adjacency matrix reflecting the neighborhoods of each data point. k-Nearest Neighbors (k-NN) method and ϵ-ball method are usually used to graph adjacency construction. Given the adjacency matrix, Heat-kernel, inverse Euclidean Distance and Locally Linear Reconstruction (LLR)[11] are often used to calculate the graph edge weights. However, traditional methods are mainly based on pair-wise Euclidean distance, which are sensitive to noise. Meanwhile, traditional methods usually use a fixed global parameter to determine the neighbors for all the samples, and thus fail to offer data-adaptive neighborhoods. At last, as graph adjacency construction and graph weight calculation are two interrelated steps, separating them leads to an information loss. These shortcomings limit the performance and efficiency of traditional graphs.

Recently, inspired by advances of compressed sensing, Yan etc.[12] proposed a novel graph called ℓ_1-graph via sparse representation by L_1 optimization. An ℓ_1-graph over a dataset is derived by encoding each data point as the sparse representation of the remaining samples, and automatically selects the most informative neighborhoods for each datum. Compared with traditional graphs, ℓ_1-graph is robust, sparse, and datum-adaptive. Meanwhile, ℓ_1-graph simultaneously learns the graph adjacency structure and graph edge weights. However, ℓ_1-graph computes the coefficients of each data points individually, lacking global structures of data. This drawback can largely reduce the performance of ℓ_1-graph in semi-supervised learning tasks.

However, these methods construct all of the data in a subspace[13], using little information about the other generated coefficients when they make a new reconstruct coefficients of data. As we know, if the samples are in the same class, their coefficients will highly related, and the coefficients should have the largest margin property if they are in different classes. In this paper, we present a novel graph construct methods. For each individual coefficient, we use sparse represent to make it more suitable for the human visual system in the feature represent and more robust to noise and partial image occlusions. In addition, for the mutual coefficient relationship, we think that if data in the same class can be well represent by others, then the represent should be highly similarity. This property is important for the low-label rate problem, where the labeled information is really limited. So we add the ℓ_2 norm limitation to the sum of each reconstruct coefficient to be minimize, which constrains graph weight centered on the same class and presents the adjacency of the more precise. Another contribution is we also give the optimization algorithm of the $\ell_{2,1}$ graph by using the traditional Argument Lagrange Multiplier (ALM). Experiments on semi-supervised face recognition show that our proposed $\ell_{2,1}$ graph can reflect the graph weight more accurately, and it can get a more precise subspace for image classification in semi-supervised learning task. Especially for the low-labeled rate dimension deduce problem, it get a better performance.

The rest of this paper is organized as follows. We first introduce traditional SDA framework in Section II. In Section III, we introduce our new $\ell_{2,1}$ optimization algorithm and graph construct method. The experimental results and our analysis are presented in Section IV Finally, we conclude the paper and provide suggestions for future work in Section V.

II. Overview of SDA

Given a label set $\{x_1, \ldots, x_m, x_m+1, \ldots, x_m+l\}$where $N = m + l$, N vectors for the whole data. m of them are labeled as $\{y_1, \ldots, y_m\}$, and other l are unlabeled. They all belong to c classes. The SDA motivates to present the prior assumption of consistency by the graph.

By the definition of within class scatter matrix S_w, between class scatter matrix S_b and total class scatter matrix S_t:

$$S_w = \sum_{k=1}^{c}(\sum_{i=1}^{l_k}(x_i^{(k)} - \mu^{(k)})(x_j^{(k)} - \mu^{(k)})^T) \tag{1}$$

$$S_b = \sum_{k=1}^{c} l_k(\mu^{(k)} - \mu)(\mu^{(k)} - \mu)^T \tag{2}$$

$$S_t = \sum_{i=1}^{l}(x_i - \mu)(x_i - \mu)^T \tag{3}$$

where μ is the total sample mean vector, l_k is the number of samples in the k-th class, $\mu^{(k)}$ is the average vector of the k-th class, adn $x_i^{(k)}$ is the i-th sample in the k-th class.

The SDA method [4] want to find a rejection matrix R that maximizes the trace function of S_t and S_w, which is also to find a suitable subspace:

$$R = arg\max_{R} \frac{R^T S_b R}{R^T S_t R + \alpha J(R)} \tag{4}$$

where α is a balance parameter, and $J(R)$ controls the learning complexity of the hypothesis family, which can be generate from a graph matrix. The graph puts an edge between nodes i and j if X_i and X_j are neighbors, which is decided by k nearest neighbor or $\epsilon-ball$ neighbor algorithm. If two data points are linked by an edge, they are likely to

be in the same class. Then the corresponding weight matrix is defined by:

$$S_{ij} = \begin{cases} i, & if\ x_i \in N(x_j)\ or\ x_j \in N(x_i) \\ 0, & otherwise. \end{cases} \tag{5}$$

where $N(x_i)$ denotes the set of neighbors of x_i. then $J(R)$ can be defined as follows:

$$J(R) = \sum_{ij} (R^T x_i - R^T x_j) S_{ij} \tag{6}$$

If we define D is a diagonal matrix where $D_{ii} = \sum_j S_{ij}$, and $L = D - S$, the rejection matrix R is easily obtained:

$$R = arg \max_R \frac{R^T S_b R}{R^T (S_t + \alpha X L X^T) R} \tag{7}$$

III. Group sparse construct algorithm

A. Motivation

For the need of classification problem, how to find a suitable subspace for classification is an important task, which we called dimensionality reduction. We assume that the training sample data are given as $X = [x_1, x_2, \ldots, x_N]$, where $x_i \in \mathbb{R}^d$ and N is the Total number of training samples. In these data l is labeled and m is unlabeled, where $N = l + m$. The dimensionality reduction on the graph consists in finding a labeling of the graph that is consistent with both the initial labeling and the geometry of the data induced by the graph structure(edges and weights W).As the proposed algorithm always analyze the relationship only using the form one-to-others, whatever the most common graph: k-nearest neighbor graph and the $\epsilon - ball$ graph, only for determining the edges and the weight graph should be 1, or the lle-graph and the ℓ_1-graph show the graph structure weights by the limitation of ℓ_2-norm or the ℓ_1-norm. By well considering the strength of the graph weight lle-graph and ℓ_1-graph get a good performance.

But without consider the relation together, above methods are under the assumption that all of the figure an well present on a single linear subspace. So it use the expression as (8)

$$a = arg \min_a \|y - Xa\|_p \tag{8}$$

Where $\|\cdot\|$ is the ℓ_p norm.

Here we think that if the data can be well represent by the same class label, and the represent in the same class are highly similar. If the number of labeled data and unlabeled data is almost the same, this property can be well reflect by their reconstruct coefficient. But when the number of labeled data is much smaller than the whole training data, this property is really helpful. So we added the ℓ_2-norm limitation in the traditional non-negative sparse representation to make the weight matrix more close to limit in the same class. Just like Equation (9).

$$\begin{aligned} &\min_A \|A\|_{2,1} \\ &s.t.\ X = XA, A \geq 0 \end{aligned} \tag{9}$$

B. Minimization of the $\ell_{2,1}$ Problem

By concern the relationship of each reconstruct coefficient, we count the graph weight using all of the training data together, and set the diag of the matrix into 0 by considering to minimize the weight of the vector itself. Supposing that there is some noise , we add a Error part E. Then we got Equation (10)

$$\begin{aligned} &\min_{A,E} \|A\|_{2,1} + \lambda \|E\|_1 \\ &s.t.\ X = XA, A \geq 0, Diag(A) = 0 \end{aligned} \tag{10}$$

This problem can be solved by change the reformulate into the below problem (11), which can solved by the Augmented Lagrange Multiplier (ALM)[14] method.

$$\begin{aligned} &\min_{J,E} \|J\|_{2,1} + \lambda \|E\|_1 \\ &s.t.\ X = XA, \\ &\quad J = A, W = A, Z = A \\ &\quad Z \geq 0, Diag(W) = 0 \end{aligned} \tag{11}$$

By using The ALM method ,it minimizes the following augmented Lagrange function (12),and we got Algorithm 1

$$\begin{aligned} &L(J, A, Z, W, E, Y_1, Y_2, Y_3, Y_4, \mu_1, \mu_2, \mu_3, \mu_4) \\ &= \|J\|_{2,1} + \lambda \|E\|_1 + \langle Y_1, X - XA - E \rangle \\ &\quad + \langle Y_2, J - A \rangle + \langle Y_3, Z - A \rangle + \langle Y_4, W - A \rangle \\ &\quad + \tfrac{\mu_1}{2} \|X - AZ - E\|_F^2 + \tfrac{\mu_2}{2} \|J - A\|_F^2 \\ &\quad + \tfrac{\mu_3}{2} \|Z - A\|_F^2 + \tfrac{\mu_4}{2} \|W - A\|_F^2 \end{aligned} \tag{12}$$

And one of the most important part of this algorithm is to solve the optimal problem as

$$J = \operatorname*{argmin}_J \|J\|_{2,1} + \frac{1}{2} \|J - Q\|_F^2$$

Fortunately, we got the method to solve this method from Guangcan Liu [15]. Let $Q = [q_1, q_2, \ldots, q_i, dots]$ the $i-$th column of W is

$$J(:, i) = \begin{cases} \frac{\|q_i\|_2 - 1}{\|q_i\|_2} q_i, & if \lambda < \|q_i\|_2, \\ 0, & otherwise. \end{cases}$$

C. Graph construct

For the training samples, we set the whole matrix as $X = [x_1, x_2, \ldots, x_N]$. Using Algorithm 1 we can got the adjacency structure and the graph weight at the same time by solving Equation (12). So we set the graph weight as : $W = A$. Using the above method, we got the coefficient compared with sparse representation (like ℓ_1-graph). As we see from figure 1, our coefficient are accumulated more from 0 to 30, which is the first class of the training sample. So the weight of the graph by $\ell_{2,1}$ graph is more efficiency.

Algorithm 1 Solving Problem (12) by Inexact ALM

Input: data matrix X, dictionary A, parameter λ

Initialize: $Z = J = W = 0, E = 0, Y_1 = Y_2 = Y_3 = 10^{-2}, \mu_{i,max} = 10^{10}, \rho = 1.1, \varepsilon = 10^{-8}$

1: **while** not converged **do**

2: Update the variables in parallel. Namely, a newly updated variables are not immediately used for updating the next variable:

$$A = [\mu_1 X^T X + (\mu_2 + \mu_3 + \mu_4) I]^{-1} [(X^T Y_1) + Y_2 + Y_3 + Y_4) + \mu_1 X^T (X - E) + \mu_2 (J - A) + \mu_3 (J - A) + \mu_4 (W - A)]$$

$$J = \underset{J}{\operatorname{argmin}} \|J\|_{2,1} + \langle Y_2, J - A \rangle + \frac{\mu_2}{2} \|J - A\|_F^2$$

$$Z = \underset{Z \geq 0}{\operatorname{argmin}} + \frac{\mu_3}{2} \|Z - A\|_F^2$$

$$W = \underset{Diag(W=0)}{\operatorname{argmin}} + \frac{\mu_4}{2} \|W - A\|_F^2$$

$$E = \underset{E}{\operatorname{argmin}} \lambda \|E\|_1 + \langle Y_1, X - XA - E \rangle \frac{\mu_1}{2} \|X - XA - E\|_F^2$$

3: Update the multiplier, using the newly updated variables:

$$Y_1 = Y_1 + \mu_1 (X - XA - E),$$
$$Y_2 = Y_2 + \mu_2 (J - A),$$
$$Y_3 = Y_3 + \mu_3 (W - A),$$
$$Y_4 = Y_4 + \mu_4 (Z - A),$$

4: Update the parameter μ_i by $\mu_i = \min(\rho \mu_i, \mu_{i,max})$, where $\mu_{i,max}$ is an upper bound of μ_i and $i = 1, 2, 3$

5: Check the convergence conditions:

$$\max(\|X - XA - E\|_\infty, \|J - A\|_\infty, \|Z - A\|_\infty, \|W - A\|_\infty) < \varepsilon$$

6: **end while**

Output: an optimal solution (A^*, E^*)

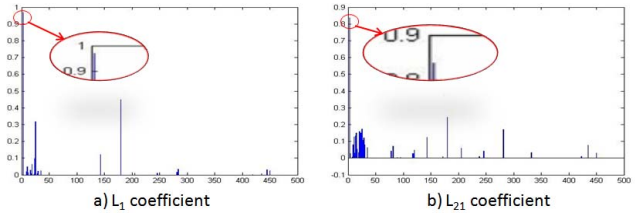

Figure 1. Coefficient of the reconstruct Matrix

IV. Experiment and Analysis

In this section,experiment is performed to test our algorithm. The databases is PIE[1], which contains 41,368 images of 68 people, each person under 13 different poses, 43 different illumination conditions, and with 4 different expressions. All of facial images are normalized to the size of 32-by32 pixel. We only pick the first 15 people for classification, and normalize the data before training and testing. We use 30 images for the training part, and the percentage of the label rate is under this training data. In such a database we trained our rejection matrix by different graph construct algorithms like k-nearest neighbor, $\epsilon - ball$, lle, ℓ_1 and our $\ell_{2,1}$ for graph construction first. We set $\lambda = 1$, and for the other graphs, we vary the values of k and ϵ. We report the classification results for different configuration with $k = 3, 5$ and $\epsilon = 0.2, 0.3$. For the lle-graph, we use the Euclidean Distance for the weight count. Then based on rejection space semi-supervised learned by these graph, we make a image classification experiment on the deduced dimension and use the basic nearest neighbor for classification.

We carry out the classification experiments on the face databases under different deduced dimension. For the whole database, we change 25 different training data and 25 different testing data to get a mean accuracy. For the percentage of the training label, we use a random select algorithm to make sure that all the classes have been selected, but the labeled number may be different for each class. We test the classification accuracy for different labeled percentage, and compare 8 different methods with 5 different algorithms where k=3 or 5 and ϵ =0.2 or 0.3. For ℓ_1-graph and lle-graph, the graph adjacency matrix W is asymmetric and a symmetrization process is used as to mean value of its transpose and itself. Focusing on low-labeled rate, we randomly choose the i percentage of the total training samples where $i = 5\%, 8\%, 10\%$ and 20%. Then we use the Semi-supervised Discriminant Analysis (SDA) is used to evaluate the performances of different graphs. And we use the nearest neighbor for the classification.

The classification error rates for semi-supervised learning based on different graphs generated subspace is shown in Figure 2, form which we can have a set of observation.

1) The $\ell_{2,1}$ graph generally achieves a highest recognition accuracy compared to those graphs, followed by ℓ_1 graph.
2) In our experiment, the ϵ graph is highly changed by the number of ϵ. You may get a good performance by change it, but when the data changed, it should change too. The k-NN graph is more robust than it.
3) Lle-graph is also stable for the classification, but the performance is not as good as ℓ_1 graph and our method.
4) For the low-labeled rate semi-supervised rejection matrix generation, $\ell_{2,1}$ graph can get a more suitable low-dimension space for classification.

V. Conclusions and Future Work

In this paper, we use a new way to construct graph for semi-supervised learning problems, and give our optimization algorithm for the graph construct. Different from the traditional graph construct method, the graph adjacency structure and the graph weights are derived together. We consider the human vision system in the representation of

[1]Available at http://www.zjucadcg.cn/dengcai/Data/FaceData.html

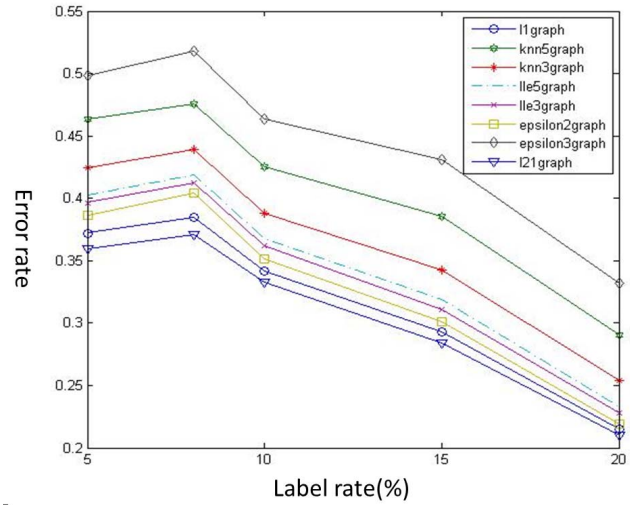

Figure 2. Classification error rate in PIE data

natural scenes by sparse coding, and consider the relationship between the training samples and their coefficient. We find a more precise method to constrain the graph weight in their own class. And the experiment shows it is truly effective for semi-supervised learning. In the further, we want to find a more precise method to construct the graph. And we want to find a more efficiency frame for the semi-supervised learning problem.

ACKNOWLEDGMENT

We would like to thank anonymous reviewers for their comments. This work is partially supported by the National Science Foundation of China (60933013), the National Science and Technology Major Project (2010ZX03004-003), and the Fundamental Research Funds for the Central Universities(WK210023002, WK2101020003).

REFERENCES

[1] R. Vidal, "A TUTORIAL ON SUBSPACE CLUSTERING."

[2] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 90–105, 2004.

[3] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 40–51, 2007.

[4] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. Int. Conf. Computer Vision (ICCV'07)*, 2007.

[5] Y. Zhang and D. Yeung, "Semi-supervised discriminant analysis using robust path-based similarity," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2008.

[6] R. Xiao and P. Shi, "Semi-supervised marginal discriminant analysis based on qr decomposition," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[7] H. Qiu, J. Lai, J. Huang, and Y. Chen, "Semi-supervised discriminant analysis based on udp regularization," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[8] Y. Zhang and D. Yeung, "Semi-supervised discriminant analysis via CCCP," *Machine Learning and Knowledge Discovery in Databases*, pp. 644–659, 2008.

[9] D. Zhai, H. Chang, B. Li, S. Shan, X. Chen, and W. Gao13, "Semi-supervised discriminant analysis via spectral transduction."

[10] J. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, p. 2319, 2000.

[11] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, p. 2323, 2000.

[12] S. Yan and H. Wang, "Semi-supervised learning by sparse representation," in *SIAM International Conference on Data Mining, SDM*, 2009, pp. 792–801.

[13] E. Elhamifar and R. Vidal, "Sparse subspace clustering," 2009.

[14] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Arxiv preprint arXiv:1009.5055*, 2010.

[15] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust Recovery of Subspace Structures by Low-Rank Representation," *Arxiv preprint arXiv:1010.2955*, 2010.