A Semi-supervised Learning Algorithm Based on Low Rank and Weighted Sparse Graph for Face Recognition

Tao Zhang$^{(\boxtimes)}$, Zhenmin Tang, and Bin Qian

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
`njustztwork@l26.com`, {`Tzm.cs,3ll062l98`}`@njust.edu.cn`

Abstract. Traditional graph-based semi-supervised learning can not capture both the global and local structures of the data exactly. In this paper, we propose a novel low rank and weighted sparse graph. First, we utilize exact low rank representation by the nuclear norm and Forbenius norm to capture the global subspace structure. Meanwhile, we build the weighted sparse regularization term with shape interaction information to capture the local linear structure. Then, we employ the linearized alternating direction method with adaptive penalty to solve the objective function. Finally, the graph is constructed by an effective post-processing method. We evaluate the proposed method by performing semi-supervised classification experiments on ORL, Extended Yale B and AR face database. The experimental results show that our approach improves the accuracy of semi-supervised learning and achieves the state-of-the-art performance.

Keywords: Semi-supervised learning · Graph construction · Low rank · Sparse

1 Introduction

For many machine learning and pattern recognition applications, it is difficult to get enough labeled samples, while a large number of unlabeled samples are widely available over the Internet. Semi-supervised learning (SSL) can utilize both limited labeled samples and abundant unlabeled samples, which has become a research focus in learning tasks. In the current method, graph-based SSL has attracted much attention because it can effectively capture the structure information hidden in the data and obtain a better performance in the practical application [1].

Graph-based SSL employs a graph to represent data structures, where the set of vertices corresponds to the samples, and the set of edges is associated with an adjacency matrix which measures the pairwise weights between vertices. Label information of the labeled samples can be propagated to the unlabeled samples over the graph by label propagation algorithm, such as local and global consistency (LGC) [2] and Gaussian random field and harmonic function (GHFH) [3]. How to construct a good graph is the difficulty of the algorithms, and it is still an open problem. Liu et al. [4] propose low rank representation (LRR), which constructs a low rank graph by solving the nuclear norm minimization problem. LRR can capture the global structure of the

data and performs well on the subspace clustering problem. Zhuang et al. [5] extend LRR and propose non-negative low rank and sparse graph (NNLRS). Compared with LRR, NNLRS adds a sparse constraint in the objective function, and it can capture the global and local structure of the data. In [6, 7], based on NNLRS, the authors propose weighted sparse constraint, where the sparse regularization term is weighted by different weight matrix and it can effectively protect the local structure of the data.

We observe that the above algorithms use the nuclear norm to estimate the rank of the matrix. Nevertheless, the nuclear norm is a convex relaxation of the rank function, and it can not estimate the rank accurately. Choosing a suitable function to estimate the rank can improve the performance of algorithms. Kang et al. [8] propose a rank approximation based on Logarithm-Determinant and it improves the accuracy of subspace clustering. Inspired by elastic net [9] in learning theory, we use both the nuclear norm and Forbenius norm as a replacement function. The rank can be estimated effectively and we can get a more exact LRR. On the other hand, in order to improve the ability to capture the local structure of the data, we also add a weighted sparse regularization term into the objective function. Different from [6, 7], we utilize the shape interaction information to construct weight matrix, which makes the graph contain more information.

The remainder of this paper is organized as follows. We give an overview of the LRR algorithm in Sect. 2. In Sect. 3, we present the proposed low rank and weighted sparse graph (LRWSG) and its optimization by linearized alternating direction method with adaptive penalty (LADMAP) [10]. The experimental results on three widely used face database are presented in Sect. 4. Finally, we conclude this paper in Sect. 5.

2   Related Work

This section briefly introduces LRR. Let $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$ be a matrix whose columns are n data samples in the d dimensional space. LRR seeks the coefficient matrix $Z = [z_1, z_2, \ldots, z_n] \in \mathbb{R}^{n \times n}$ which is the lowest rank representation that can represent X as a linear combination of itself. The LRR problem is defined as follows:

$$\min_{Z} ||Z||_* + \lambda ||E||_{2,1}, s.t. X = XZ + E. \tag{1}$$

where $|| \cdot ||_*$ is the nuclear norm of a matrix (the sum of the singular values of a matrix). $||E||_{2,1} = \sum_{j=1}^{n} (\sum_{i=1}^{d} E_{ij}^2)^{1/2}$ is 2,1-norm and it is used to represent noise. The parameter λ is used to balance the effect of noise. The inexact augmented Lagrange multiplier (IALM) [11] method is employed to solve the problem (1), and we can get the optimal solution (Z^*, E^*). The adjacency matrix of the low rank graph can be calculated as follows:

$$G = (|Z^*| + |Z^*|^T)/2 \tag{2}$$

After we get the adjacency matrix, LGC or GHFH algorithm is used to propagate label information and obtain the results of semi-supervised classification.

3   The Proposed Method

3.1   Problem Formulation

Elastic net which utilizes both the 1-norm and 2-norm as penalty function is an effective model in statistical learning [9]. The 1-norm guarantees the sparsity of the solution, while the 2-norm guarantees the stability of the solution. And the model performs well on the low rank matrix completion problem [12].

We observe that $||Z||_*$ in Eq. (1) can be represented as $\sum_{i=1}^{r} |\sigma_i|$, where σ_i is the ith singular value of Z, r is the rank of Z. Obviously, $\sum_{i=1}^{r} |\sigma_i|$ is the 1-norm penalty of the singular value of Z. To improve the stability of the algorithm, we introduce $\sum_{i=1}^{r} |\sigma_i|^2$ as a 2-norm penalty of the singular value of Z. Actually, $||Z||_F^2 = Tr(V\Lambda U^T U\Lambda V^T) = Tr(\Lambda^2) = \sum_{i=1}^{r} |\sigma_i|^2$, where $Z = U\Lambda V^T$ is SVD of Z. By combining the 1-norm penalty and 2-norm penalty, we can rewrite Eq. (1) as follows:

$$\min_Z ||Z||_* + \alpha||Z||_F^2 + \lambda||E||_{2,1}, s.t. \, X = XZ + E. \tag{3}$$

where the parameter α is used to trade off the effect of 1-norm penalty and 2-norm penalty. Compared with Eqs. (1), (3) is a stable model which can estimate the rank of Z and capture the global subspace structure more exactly.

In order to capture the local linear structure, $||Z||_1$ is added into Eq. (1) [5]. Later on, [6, 7] propose weighted sparse constraint $||W \odot Z||_1$, where $\odot$ denotes the Hadamard product, if $M = A \odot B$, then $M_{ij} = A_{ij} \times B_{ij}$. Constructing a weight matrix W with more information can protect the local structure of the data. Inspired by [13], we utilize shape interaction information to construct W. Let $X = U_r\Lambda_r V_r^T$ be the skinny SVD of X, where r is the rank of X. The shape interaction representation of each data sample x_i is $R_i = \Lambda_r^{-1} U_r^T x_i$. Normalize all column vectors of R_i by $R_i^* = R_i/||R_i||_2$, and the shape interaction weight matrix $\mathbb{W}$ can be defined as follows:

$$\mathbb{W}_{ij} = || R_i^* - R_j^* ||_2 \tag{4}$$

In summary, we formulate the objective function of LRWSG as follows:

$$\min_Z || Z ||_* + \alpha || Z ||_F^2 + \beta || \mathbb{W} \odot Z ||_1 + \lambda || E ||_{2,1}, \quad s.t. \, X = XZ + E, Z \geq 0. \tag{5}$$

3.2   Optimization

Similar to [5], we utilize LADMAP to solve problem (5). We first introduce an auxiliary variable J to separate the variable in the objective function. Thus Eq. (5) can be rewritten as follows:

$$\min_{Z} \| Z \|_* + \alpha \| Z \|_F^2 + \beta \| W \odot J \|_1 + \lambda \| E \|_{2,1}, \quad s.t. \ X = XZ + E, Z = J, J \geq 0. \quad (6)$$

The augmented Lagrange function of Eq. (6) is

$$
\begin{aligned}
L = \ & \| Z \|_* + \alpha \| Z \|_F^2 + \beta \| W \odot J \|_1 + \lambda \| E \|_{2,1} + Tr \left[Y_1^T \left(X - XZ - E \right) \right] \\
& + Tr \left[Y_2^T \left(Z - J \right) \right] + \frac{\mu}{2} [\| X - XZ - E \|_F^2 + \| Z - J \|_F^2]
\end{aligned}
\quad (7)
$$

where Y_1 and Y_2 are Lagrange multipliers, $\mu > 0$ is a penalty parameter.
Update Z_{k+1} with Z_k, J_k, E_k fixed.

$$
\begin{aligned}
Z_{k+1} = \arg\min_{Z} \ & \frac{1}{\eta \mu_k} \| Z \|_* + \frac{1}{2} \| Z - (Z_k + (X^T (X - XZ_k - E + Y_{1,k}/\mu_k) \\
& - (Z_k - J_k + Y_{2,k}/\mu_k) - (2\alpha/\mu)Z_k)/\eta) \|_F^2
\end{aligned}
\quad (8)
$$

where $\eta = \|X\|_2^2$, Eq. (8) can be solved by singular value thresholding operator [14].
We set $A = (Z_k + (X^T(X - XZ_k - E + Y_{1,k}/\mu_k) - (Z_k - J_k + Y_{2,k}/\mu_k) - (2\alpha/\mu)Z_k)/\eta$,
$A = U\Lambda V^T$ is the SVD of A, the solution of Eq. (8) is $Z_{k+1} = US_{1/\beta}(\Lambda)V^T$, where S is
soft thresholding operator [11], defined as $S_\varepsilon[x] = \max(x - \varepsilon, 0) + \min(x + \varepsilon, 0)$.
Update J_{k+1} with Z_{k+1}, J_k, E_k fixed.

$$J_{k+1} = \arg\min_{J \geq 0} \frac{\beta}{\mu_k} \| W \odot J \|_1 + \frac{1}{2} \| J - (Z_{k+1} + Y_{2,k} / \mu_k) \|_F^2 \quad (9)$$

Equation (9) can be solved by soft thresholding operator and the solution is

$$(J_{k+1})_{ij} = \max(S_{\varepsilon_{ij}}[(Z_{k+1})_{ij} + (Y_{2,k})_{ij}/\mu_k], 0) \quad (10)$$

where $\varepsilon_{ij} = (\beta / \mu_k) W_{ij}$.

Update E_{k+1} with Z_{k+1}, J_{k+1}, E_k fixed.

$$E_{k+1} = \arg\min_{E} \frac{\lambda}{\mu_k} \|E\|_{2,1} + \frac{1}{2} \| E - (X - XZ_{k+1} + Y_{1,k}/\mu_k) \|_F^2 \quad (11)$$

The solution of Eq. (11) is

$$E_{k+1} = \Omega_{\lambda/\mu_k}(X - XZ_{k+1} + Y_{1,k}/\mu_k) \quad (12)$$

where Ω is 2,1-norm minimization operator [4]. If $Y = \Omega_\varepsilon(X)$, then the ith column of Y is

$$Y(:,i) = \begin{cases} \frac{||X(:,i)||_2 - \varepsilon}{||X(:,i)||_2} X(:,i), & \varepsilon < ||X(:,i)||_2 \\ 0, & \varepsilon \ge ||X(:,i)||_2 \end{cases} \quad (13)$$

The complete optimization to LRWSG is summarized in Algorithm 1.

3.3   Graph Construction

Once problem ([5]) is solved, we can obtain an optimal Z^*. Different from the traditional graph-based SSL construct the adjacency matrix by Eq. ([2]), we utilize the method which is used on the subspace clustering problem [4]. Let $Z^* = U^* \Lambda^* (V^*)^T$ be the skinny SVD of Z^*, we define $P = U^* (\Lambda^*)^{1/2}$, the adjacency matrix of LRWSG is calculated as follows:

Algorithm 1: LRWSG

Input: data matrix X, parameters α, β and λ, weight matrix $\mathbb{W}$;

Initialization: $Z_0 = J_0 = E_0 = Y_{1,0} = Y_{2,0} = 0$, $\mu_0 = 0.1$, $\mu_{max} = 10^{10}$, $\rho_0 = 1.1$,

$\quad\quad\quad\quad \varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 10^{-2}$, $\eta = \|X\|_2^2$, $k = 0$;

Definition: $\varepsilon = \mu_k \max(\sqrt{\eta}\|Z_{k+1} - Z_k\|_F, \|J_{k+1} - J_k\|_F, \|E_{k+1} - E_k\|_F)/\|X\|_F$;

while $\|X - XZ_{k+1} - E_{k+1}\|_F / \|X\|_F \ge \varepsilon_1$ or $\varepsilon \ge \varepsilon_2$ **do**

(1) Update the variables (Z, J, E) as Eq. (8), Eq. (9) and Eq. (11);

(2) Update Lagrangian multipliers as

$\quad Y_{1,k+1} = Y_{1,k} + \mu_k (X - XZ_{k+1} - E_{k+1})$, $\quad Y_{2,k+1} = Y_{2,k} + \mu_k (Z_{k+1} - J_{k+1})$;

(3) Update μ as $\mu_{k+1} = \min(\mu_{max}, \rho\mu_k)$, where

$\quad \rho = \begin{cases} \rho_0, & \varepsilon < \varepsilon_2, \\ 1, & \varepsilon \ge \varepsilon_2, \end{cases}$;

(4) Update k as $k = k + 1$;

end while

Output: optimal solution (Z^*, E^*).

$$(G)_{ij} = (PP^T)^2_{ij} \quad (14)$$

After we obtain the adjacency matrix G, LGC algorithm is employed to solve the semi-supervised classification problem.

4   Experiment

In this section, we evaluate the effectiveness of LRWSG on semi-supervised classification experiments. LRWSG is compared with several LRR related graphs including LRR [4], NNLRS [5], LRRLC [6], SCLRR [7] and CLAR [8]. The classification

accuracy is used to evaluate the semi-supervised classification performance, which is defined to be the percentage of correctly classified samples versus the test samples. The parameters of the compared algorithms are tuned to achieve the best performance. In LRWSG, the parameter α balances the effect of nuclear norm and Forbenius norm. According to a large number of experiments, the classification accuracy is better when we set $\alpha = 1$. The parameter λ is used to describe the noise of data, we set $\lambda = 10$ for our experiments. The parameter β controls the effect of sparse regularization term, we set $\beta = 0.3$ on ORL database and EYaleB database and we set $\beta = 0.1$ on AR database. The experiments are implemented on Intel Core i7 4710MQ CPU with 8 G memory.

4.1   Databases

We select three face databases for our experiments: ORL, Extended Yale B (EYaleB) and AR. The ORL database contains 40 distinct subjects, and each subject has 10 different images. The images were taken at different times, varying the lighting, facial expressions and facial details. There are 64 face images under different illuminations of each of 38 individuals in the EYaleB database. In our experiments, we use the first 20 subjects and each subject chooses the first 50 images. The AR database contains 3120 images of 120 subjects with different facial expressions, lighting conditions and occlusions. The first 50 subjects are chosen for our experiments, and each subject chooses the first 20 images. All the images are resized to 32×32. Several sample images of the three face databases are shown in Fig. 1.

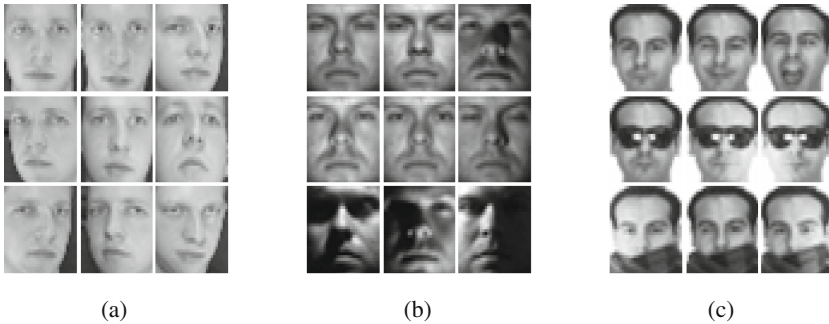(a)   (b)   (c)

Fig. 1. Some sample images from three databases: (a) ORL, (b) EYaleB, (c) AR

4.2   Experimental Results and Analysis

For each database, we randomly choose 10 % to 60 % samples from each class as labeled samples, and the rest samples are used for testing. For each percentage of labeled samples, we repeat the experiment 20 trials for each algorithm. Tables 1, 2 and 3 report the classification accuracies and standard deviations of each algorithm on ORL, EYaleB and AR.

Table 1. The classification accuracies and standard deviations (%) on ORL

Label	LRR	CLAR	NNLRS	LRRLC	SCLRR	LRWSG
10 %	63.83 ± 2.21	68.53 ± 1.88	72.76 ± 2.03	69.78 ± 1.86	75.93 ± 2.22	**82.10 ± 2.32**
20 %	73.19 ± 2.54	80.45 ± 2.48	82.78 ± 2.69	80.98 ± 2.25	84.31 ± 2.77	**87.63 ± 2.22**
30 %	78.00 ± 2.19	86.23 ± 2.71	87.96 ± 1.94	86.41 ± 2.55	88.89 ± 2.00	**91.04 ± 2.10**
40 %	80.56 ± 2.42	89.96 ± 1.69	91.04 ± 1.87	89.98 ± 1.77	92.35 ± 1.64	**93.17 ± 1.98**
50 %	82.03 ± 2.06	91.18 ± 1.47	92.55 ± 1.42	91.53 ± 1.49	93.50 ± 1.47	**94.20 ± 1.66**
60 %	83.75 ± 2.94	93.06 ± 1.89	93.78 ± 1.65	93.28 ± 1.69	94.22 ± 1.61	**94.75 ± 2.08**

Table 2. The classification accuracies and standard deviations (%) on EYaleB

Label	LRR	CLAR	NNLRS	LRRLC	SCLRR	LRWSG
10 %	86.43 ± 1.56	87.57 ± 1.47	92.64 ± 1.02	87.17 ± 1.41	88.90 ± 1.21	**92.99 ± 1.02**
20 %	91.76 ± 1.37	93.01 ± 1.39	95.00 ± 0.79	92.65 ± 1.31	93.21 ± 1.31	**95.36 ± 0.73**
30 %	93.13 ± 0.98	94.69 ± 0.79	96.02 ± 0.61	94.53 ± 0.72	94.98 ± 0.72	**96.45 ± 0.60**
40 %	94.56 ± 0.88	96.25 ± 0.92	96.99 ± 0.61	96.13 ± 0.78	96.43 ± 0.67	**97.32 ± 0.60**
50 %	95.14 ± 1.17	96.85 ± 0.79	97.28 ± 0.59	97.09 ± 0.57	97.41 ± 0.62	**97.69 ± 0.56**
60 %	95.39 ± 0.83	97.50 ± 0.59	97.61 ± 0.70	97.60 ± 0.67	97.60 ± 0.67	**97.96 ± 0.63**

Table 3. The classification accuracies and standard deviations (%) on AR

Label	LRR	CLAR	NNLRS	LRRLC	SCLRR	LRWSG
10 %	85.70 ± 1.78	84.06 ± 2.09	83.68 ± 1.81	85.83 ± 1.88	85.07 ± 1.88	**93.53 ± 1.40**
20 %	92.34 ± 0.89	93.17 ± 1.02	91.63 ± 1.11	93.26 ± 1.01	93.24 ± 1.06	**95.96 ± 0.91**
30 %	94.82 ± 1.04	96.01 ± 0.98	94.96 ± 1.14	95.84 ± 0.97	95.96 ± 0.90	**97.19 ± 0.99**
40 %	95.88 ± 0.72	97.41 ± 0.63	96.99 ± 0.75	97.07 ± 0.58	97.48 ± 0.68	**98.23 ± 0.58**
50 %	96.12 ± 0.88	97.63 ± 0.67	97.37 ± 0.80	97.18 ± 0.83	97.68 ± 0.77	**98.35 ± 0.64**
60 %	96.81 ± 0.64	98.11 ± 0.71	98.24 ± 0.83	97.84 ± 0.80	98.31 ± 0.77	**98.70 ± 0.70**

From the results, we can observe that:

(1) LRWSG achieves the highest accuracies on both databases. LRWSG utilizes the nuclear norm and Forbenius norm to estimate the rank function. Meanwhile, the weighted sparse regularization term with shape interaction information is joined into the objective function. Therefore, LRWSG can capture both the global subspace structure and local linear structure exactly. And the standard deviations of LRWSG are often small, which shows the stability of LRWSG.

(2) CLAR uses the logarithm determinant function to estimate the rank function, which improves the performance of LRR. But compared with the algorithms which consider both low rank and sparse property, CLAR performs worse. Depend on LRR, NNLRS, LRRLC and SCLRR propose different sparsity constraint. Although the weight matrices of sparse regularization term are different, the performance of these

three algorithms are similar. With exact rank estimation and informational weighted sparse matrix, LRWSG performs better than these three algorithms.

(3) With the increase of the number of labeled samples, the classification accuracies of each algorithm are also increased. When given more labeled samples, the label information is more abundant, and each algorithm performs well. When given less labeled samples, the classification becomes more difficult, but LRWSG can still get a higher classification accuracies. For example, with 10 % labeled samples, the accuracy of LRWSG is 82.10 % on ORL database, which is 6.17 % higher than the best result obtained from other algorithms

5   Conclusion

This paper proposes a novel semi-supervised learning algorithm based on low rank and weighted sparse graph (LRWSG), and applies it to face recognition. In order to capture the data structure exactly, LRWSG makes use of the nuclear norm and Forbenius norm to estimate the rank function, and adds a weighted sparse constraint with shape interaction information into the object function. LADMAP is employed to solve the optimization problem. And with an effective post-processing method, the graph is constructed and used for semi-supervised classification. Experimental results on ORL, EYaleB and AR databases show that the proposed approach achieves better classification performance.

References

1. Zhu, X.: Semi-supervised Learning Literature Survey. Technical report, University of Wisconsin, Madison (2006)
2. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Advances in Neural Information Processing Systems, pp. 321–328. MIT Press, Cambridge (2004)
3. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the 20th International Conference on Machine Learning, pp. 912–919. AAAI Press, Washington, DC (2003)
4. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 171–184 (2013)
5. Zhuang, L., Gao, H., Lin, Z., Ma, Y., Zhang, X., Yu, N.: Non-negative low rank and sparse graph for semi-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2328–2335. IEEE Press, Providence (2012)
6. Zheng, Y., Zhang, X., Yang, S., Jiao, L.: Low-rank representation with local constraint for graph construction. Neurocomputing **122**, 398–405 (2013)
7. Tang, K., Liu, R., Su, Z., Zhang, J.: Structure-constrained low-rank representation. IEEE Trans. Neural Netw. Learn. Systems **25**(12), 2167–2179 (2014)
8. Kang, Z., Peng, C., Cheng, Q.: Robust subspace clustering via smoothed rank approximation. IEEE Signal Process. Lett. **22**(11), 2088–2092 (2015)
9. De Mol, C., De Vito, E., Rosasco, L.: Elastic-net regularization in learning theory. J. Complex. **25**(2), 201–230 (2009)

10. Lin, Z., Liu, R., Su, Z.: Linearized alternating direction method with adaptive penalty for low-rank representation. In: Advances in Neural Information Processing Systems, pp. 612–620. Springer Press, Granada (2011)
11. Lin, Z., Chen, M., Ma, Y.: The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. Technical report, UILU-ENG-09-2215 (2009)
12. Li, H., Chen, N., Li, L.: Error analysis for matrix elastic-net regularization algorithms. IEEE Trans. Neural Netw. Learn. Syst. **23**(5), 737–748 (2012)
13. Liu, B., Jing, L., Yu, J., Li, J.: Robust graph learning via constrained elastic-net regularization. Neurocomputing **171**, 299–312 (2016)
14. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**(4), 1956–1982 (2010)