A Discriminant Sparse Representation Graph-Based Semi-Supervised Learning for Hyperspectral Image Classification

Yuanjie Shao, Changxin Gao, and Nong Sang(✉)

Science and Technology on Multi-spectral Information Processing Laboratory, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China
nsang@hust.edu.cn

Abstract. The classification of hyperspectral image with a paucity of labeled samples is a challenging task. In this paper, we present a discriminant sparse representation (DSR) graph for semi-supervised learning (SSL) to address this problem. For graph-based methods, how to construct a graph among the pixels is the key to a successful classification. Our graph construction method contains two steps. Sparse representation (SR) method is first employed to estimate the probability matrix of the pairwise pixels belonging to the same class, and then this probability matrix is integrated into the SR graph, which can be obtained by solving an ℓ_1 optimization problem, to form a DSR graph. Experiments on Hyperion and AVIRIS hyperspectral data show that our proposed method outperforms state of the art.

Keywords: Hyperspectral image classification · Graph · Semi-Supervised Learning (SSl) · Sparse Representation (SR)

1 Introduction

Hyperspectral image data contains high-resolution spectral information on land covers, which is attractive for discriminating the subtle differences between classes with similar spectral signatures. However, hyperspectral image classification often faces the issue of limited number of labeled samples, as it is labor intensive and time-consuming to collect large number of training samples [1–3]. Semi-supervised learning (SSL) , which can utilize both small amount of labeled samples and abundant yet unlabeled samples, has recently been proposed to tackle the challenge [4,5]. Due to its practical success and its computational efficiency, graph-based SSL is pretty appealing among the semi-supervised methods.

Graph-based SSL is dependent on a graph to represent the data structures, where each vertex corresponding to one sample and the edge weight denotes the similarity between the pairwise samples. Label information of labeled instances can then be efficiently propagated to the unlabeled samples through the graph. In order to expect desired result, it is critical to construct a good graph for all

H. Zha et al. (Eds.): CCCV 2015, Part I, CCIS 546, pp. 160–167, 2015.
DOI: 10.1007/978-3-662-48558-3_16

graph-based SSL methods. Nevertheless, it is still an open problem about how to construct such a good graph [6–8].

Recently, Cheng and Yan [9,10] proposed an ℓ_1-graph structure based on sparse representation(SR).The latent philosophy is that each sample can be encoded as a sparse linear superposition of the remaining samples via solving an ℓ_1 optimization problem. In this way, the adjacency relationship and the weights of graph are derived automatically and simultaneously. Comparing with the traditional methods, e.g., k-nearest neighbors (kNN) graph and local linear embedding (LLE) graph [8,11], ℓ_1-graph (SR graph) explores higher order relationships among data points, and hence has the natural discriminating powerful. However, it finds the sparse representation of each sample in an unsupervised manner, encoding the similarity between samples ineffectively.

Inspired by above insights, we propose to combine both ℓ_1-graph and partial labeled information to construct a discriminant sparse representation (DSR) graph. It could reduce the weight of two samples if they belong to the different clusters. On top of DSR graph, SSL is then conducted to obtain the final classification results. The experimental results on Hyperion and AVIRIS hyperspectral data clearly show it outperforms the state of the art.

2 Related Works

In the following, we will introduce the graph-based SSL methods. They are all dependent on a graph to represent the data structures, where each vertex corresponding to one sample and the edge weight denotes the similarity between the pairwise samples. Popular methods include Gaussian Harmonic Function (GHF) [6], local and global consistency (LGC) [7], linear neighborhoods propagation (LNP) [8]. These methods usually relay on the assumption label smoothness over the graph. They can be viewed as estimating a function f on the graph, one wants f to satisfy both the label consistency on the labeled samples and label smoothness over the graph, where smoothness can be measured by a graph Laplacian regularization term.

Given the labeled samples $\mathbf{X}_l = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_l]$ and the unlabeled samples $\mathbf{X}_u = [\mathbf{x}_{l+1}, \mathbf{x}_{l+2}, ..., \mathbf{x}_{l+u}]$, there are c classes denoted as $\mathbf{C} = [1, 2, ..., c]$. Both the labeled and unlabeled samples $\mathbf{X} = [\mathbf{X}_l, \mathbf{X}_u]$ produce a connected graph $G = (V, E)$, where the nodes V corresponding to the $n = l + u$ samples, and the edges E are represented by a weight matrix $\mathbf{W} \in R^{n \times n}$. Then we can obtain the graph Laplacian matrix $\mathbf{L_W} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the diagonal degree matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$. Let $\mathbf{Y} = [\mathbf{Y}_l, \mathbf{Y}_u]^T \in R^{n \times c}$ be a label matrix, where $\mathbf{Y}_{ij} = 1$ if the label of sample $\mathbf{X}_i$ belongs to class j for $j \in [1, 2, ..., c]$ and $\mathbf{Y}_{ij} = 0$ otherwise. The objective of SSL is to obtain the labels of unlabeled samples based on the label matrix $\mathbf{Y}_l$ and the whole data set $\mathbf{X}$.

The graph Laplacian regularization term is denoted as

$$Tr(\mathbf{F}^T \mathbf{L_W} \mathbf{F}) = \frac{1}{2} \sum_{i,j=1}^{n} \mathbf{W}_{ij} \|\mathbf{f}_i - \mathbf{f}_j\|^2 \tag{1}$$

where $\mathbf{F}=[\mathbf{F}_l,\mathbf{F}_u]^T \in R^{n\times c}$ indicates the prediction matrix of data $\mathbf{X}$, and $\mathbf{f}_i \in R^{1\times c}$ and $\mathbf{f}_j \in R^{1\times c}$ are the predictions of samples x_i and x_j.

Since the graph-based SSL methods are similar to each other, we only apply DSR graph to GHF, although it can also be used in other methods. GHF learns a prediction function $\mathbf{F} \in R^{n\times c}$ to realize the label propagation. It constrains the predictions of labeled data to be equal to true label information, and solves the following optimization problem:

$$\begin{aligned} \min_{\mathbf{F}\in R^{n\times c}} \; & Tr(\mathbf{F}^T\mathbf{L}_{\mathbf{W}}\mathbf{F}) \\ s.t \quad & \mathbf{F}_l = \mathbf{Y}_l \end{aligned} \tag{2}$$

We can partition the matrix $\mathbf{L}_{\mathbf{W}}$ into four blocks based on labeled and unlabeled nodes,

$$\begin{pmatrix} \mathbf{L}_{\mathbf{W}_{ll}} & \mathbf{L}_{\mathbf{W}_{lu}} \\ \mathbf{L}_{\mathbf{W}_{ul}} & \mathbf{L}_{\mathbf{W}_{uu}} \end{pmatrix} \tag{3}$$

and we obtain the solution:

$$\mathbf{F}_u = -{\mathbf{L}_{\mathbf{W}_{uu}}}^{-1}\mathbf{L}_{\mathbf{W}_{ul}}\mathbf{Y}_l \tag{4}$$

The predicted label of unlabeled samples is given by:

$$y_i = \arg \max_{j=1,2,...,c} \mathbf{F}_u(i,j) \qquad i = 1,2,...,u \tag{5}$$

3 Discriminant Sparse Representation Graph Construction

In this section we propose a new approach to construct an SR graph with non-uniform class-probability called discriminant sparse representation (DSR) graph. In such a graph structure, each pairwise nodes are treated differently according to the probability that they belong to the same class. Different from SR graph, DSR graph explores class relationships among data samples, hence is more discriminative. Firstly, we provide a method on how to estimate the class-probability of unlabeled samples, and then present our DSR graph definition.

3.1 Estimation of Class-Probability

For labeled samples, they have a certain membership with one class. However, those unlabeled samples have an uncertain class relationship. Fortunately, we can estimate the class-probability of unlabeled samples via partial label information. According to the sparse representation based classification (SRC) [13], a test sample in the unlabeled samples can be encoded as a sparse linear superposition of the training samples, two samples that have non-zero coefficients in the decomposition will be in the same class and the coefficient denotes the similarity of the two samples. For its merit, SRC is applied to estimate the class-probability.

Given the initial label matrix $\mathbf{Y}_l \in R^{l\times c}$, where $\mathbf{Y}_{ij} = 1$ if the label of data $\mathbf{x}_i$ belongs to class j for $j \in [1, 2, ..., c]$ and $\mathbf{Y}_{ij} = 0$ otherwise. Let $\mathbf{D}$ be the training samples, $\mathbf{x}_i \in \mathbf{X}_u$ be a test sample, we can acquire a sparse vector $\mathbf{A} \in R^{l\times 1}$, which denotes the similarity between test sample $\mathbf{x}_i$ and l training samples, via solving following ℓ_1 minimization:

$$\begin{aligned} &\min \|\mathbf{A}\|_1 \\ &s.t. \quad \mathbf{DA} = \mathbf{x}_i \end{aligned} \tag{6}$$

where $\|\mathbf{A}\|_1$ denotes the ℓ_1 norm, i.e., the sum of the absolute value of all components in $\mathbf{A}$.

The class-probability vector of $\mathbf{x}_i$ then can be calculated by

$$\mathbf{P}_i = \mathbf{A}^T \mathbf{Y}_l \tag{7}$$

where $\mathbf{P}_i = (\mathbf{P}_{i1}, \mathbf{P}_{i2}, ..., \mathbf{P}_{ic}) \in R^{1\times c}$, the entry $\mathbf{P}_{ic}$ of the vector represents the probability of data $\mathbf{x}_i$ belonging to class c. Then we can obtain a class-probability matrix $\mathbf{P}_U \in R^{u\times c}$ of unlabeled samples. For labeled samples, we denote class-probability matrix $\mathbf{P}_L \in R^{l\times c}$ as $\mathbf{Y}_L$.

Therefore, the probability of $\mathbf{x}_i$ and $\mathbf{x}_j$ belonging to the same class can be given by

$$\mathbf{P}_{ij} = \begin{cases} 1 & i = j \\ \mathbf{P}_i {\mathbf{P}_i}^T & i \neq j \end{cases} \tag{8}$$

3.2 Discriminant Sparse Representation Graph

Compared with the kNN graph and LLE graph, SR graph can discover the local relationship and obtain the edge weights simultaneously, and has discriminating power. For each sample $\mathbf{x_i}$, SR can encode it as a sparse linear superposition of the remaining samples by solving following problem:

$$\begin{aligned} &\min \|\alpha_i\|_1 \\ &s.t. \quad \mathbf{B}\alpha_i = \mathbf{x}_i, \quad \alpha \geq 0 \end{aligned} \tag{9}$$

where $\mathbf{B} = \{\mathbf{x} | \mathbf{x} \in \mathbf{X}, \mathbf{x} \neq \mathbf{x}_i\}$ denotes all the data points except $\mathbf{x}_i$. We can construct an SR graph with a norm that an edge connects $\mathbf{x}_i$ and $\mathbf{x}_j$ if the coefficient $\alpha_{ij} \neq 0$, and the edge wight $\mathbf{W}_{(sr)_{ij}} = \alpha_{ij}$.

However, SR graph did not take prior knowledge into account. Sometimes we may know a prior the existence of certain edges and we would like to include those edges in the final graph. Therefore, we construct a DSR graph by considering partial label information, the weight of two samples $\mathbf{x}_j$ and $\mathbf{x}_i$ in which is given by

$$\mathbf{W}_{(dsr)_{ij}} = \mathbf{W}_{(sr)_{ij}} \mathbf{P}_{ij} \tag{10}$$

Different from SR graph, the DSR graph explores the classified information among the samples, and therefore is more powerful and discriminative.

4 Experiments and Analysis

4.1 Experimental Datasets

In our experiments, two hyperspectral images were employed to evaluate the performance of the DSR graph. The first one was collected by the Hyperion instrument on the NASA EO-1 satelite, and the other by the NASA Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). Hyperion acquires 242-band data at 30-m spatial resolution, covering the 357-2576-nm portion of the spectrum in 10-nm bands. Removal of uncalibrated and noisy bands resulted in 145. The Hyperion images utilized in the experiments were acquired over the Okavango Delta, Botswana (BOT) in May 2001. There are 9 classes in BOT images. The 224-band AVIRIS data was collected over Indiana Pine (IND PINE) in 1992, with a 20-m spatial resolution and 10-nm spectral resolution over the range of 400-2500 nm. 220 available bands remained after removal of noisy and water absorption bands. The RGB images and ground reference information are shown in Fig. 1.

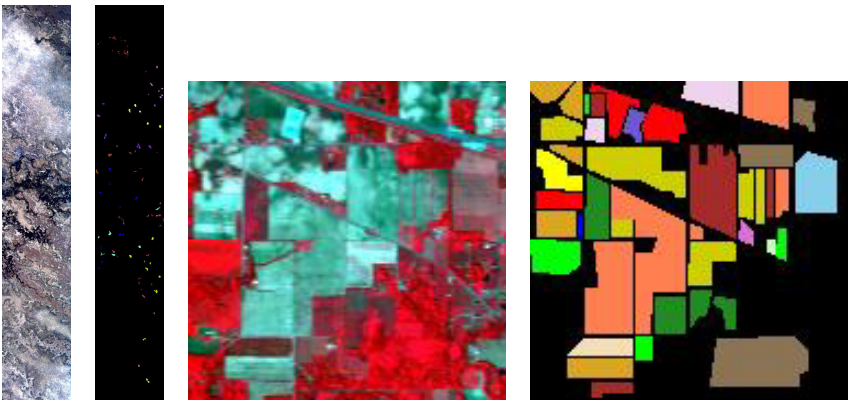


Fig. 1. (a) The BOT scence (band 29, 23, 16 for red, green, and blue, resp.). (b) Ground reference of BOT image. (c) The IND PINE scence (band 57, 27, 17 for red, green, and blue, resp.). (d) Ground reference of IND PINE image.

For IND PINE data set, we selected a sub data set from 16 classes with a modest number of labeled samples. The class names and number of data ponits in the BOT and IND PINESUB data are shown in Table 1.

4.2 Results of Classification Experiments

Four graph construction methods, e.g., (kNN) graph, LLE graph, SR graph, DSR graph, were applied to GHF for comparison. We randomly selected 3, 5, 10, 15,

Table 1. Class names and numbers of samples

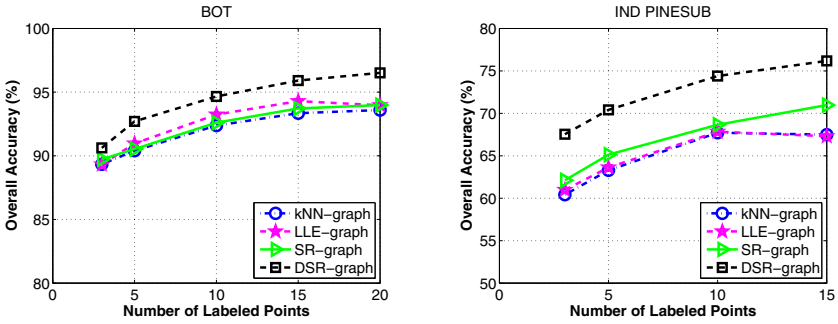
BOT		IND PINESUB	
ID	Class Name	ID	Class Name
1	Water (158)	1	Alfalfa (54)
2	Floodplain (228)	2	Corn - No till (100)
3	Riparian (237)	3	Corn C Min till (270)
4	Firescar (178)	4	Corn (234)
5	Island Interior (183)	5	Grass/pasture (63)
6	Woodlands (199)	6	Grass/trees (101)
7	Savanna (162)	7	Grass/pasture-mowed (26)
8	Short Mopane (124)	8	Hay- windrowed (489)
9	Exposed Soils (111)	9	Oats (20)
		10	Soy C No till (66)
		11	Soy C Min till (122)
		12	Soy C clean (261)
		13	Wheat (212)
		14	Woods (117)
		15	Bldg-grass-trees-drives (291)
		16	Stone-steel towers (95)

20 data points per class as training samples, and the remainder as test samples. We run the algorithms twenty times with the randomly selected samples, and the mean of overall accuracy (OA) were applied to evaluate the classification results. The optimal parameter was obtained by leave-one-out (LOO) [14] methods. For kNN graph, the number of nearest neighbors are each set to 7 and 5, and the gaussian kernel parameter σ are both set to 0.1 in BOT and IND PINESUB data. For LLE graph, the number of nearest neighbors are both set to 7 in BOT and IND PINESUB data.

Fig. 2 shows the the classification results of our algorithm with optimal parameters on two data sets, where the x-axis denotes the number of labeled samples per class, and the y-axis represents the mean of OA, we can observe that:

1) The performance of DSR graph is the best on the two data sets with different proportions of labeled samples, which denotes that the DSR graph can describe the true local linear relationship of the data points, and thus is more discriminative than other three graphs.

2) The DSR graph construction method obtain higher OA than SR graph on both two data sets with different numbers of labeled points, since the latter only considers similarity between data points, whereas the former method calculates the weights by exploiting partial labeled information, which means the lower probability that the pairwise points belong to the same class, the smaller weights are given to them, thus resulting in more discriminative ability for classification.



(a) (b)

Fig. 2. Overall accuracy of four graphs combined with GHF under different proportions of labeled samples. (a) BOT data with 9 classes, (b)IND PINESUB data with 16 classes.

5 Conclusion

This paper has developed a novel discriminative graph, called discriminative sparse representation (DSR) graph, for graph-based SSL. DSR graph has not only the merits of the SR graph, but also exploits partial labeled information. It obtains a more discriminant graph construction by combining above two aspects. The experimental results on Hyperion and AVIRIS hyperspectral data show that, DSR graph is better at reveal the true local linear relationships of the data points, and thus is more discriminative than other graphs for graph-based SSL.

Acknowledgments. This work is supported by the Project of the National Natural Science Foundation of China No.61433007 and No.61401170.

References

1. Kim, W., Crawford, M.: Adaptive classification for hyperspectral image data using manifold regularization kernel machines. IEEE Transactions on Geoscience and Remote Sensing **48**(11), 4110–4121 (2010)
2. Lunga, D., Prasad, S., Crawford, M., Ersoy, O.: Manifold-learning-based feature extraction for classification of hyperspectral data: a review of advances in manifold learning. IEEE Signal Process **31**(1), 55–66 (2014)
3. Gao, Y., Ji, R., Cui, P., Dai, Q., Hua, G.: Hyperspectral image classification through bilayer graph-based learning. IEEE Transactions on Image Processing **23**(7), 2769–2778 (2014)
4. Zhu, X.: Semi-supervised learning literature survey. Computer Sciences. University of Wisconsin-Madison (2009)
5. Camps-Valls, G., Bandos, T., Zhou, D.: Semi-supervised graph-based hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing **45**(10), 3044–3054 (2007)

6. Zhu, X., Lafferty, J., Ghahramani, Z.: Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the 20th International Conference on Machine Learning, pp. 912–919. AAAI Press, California (2003)
7. Zhou, D., Bousquet, O., Lal, T.: Learning with local and global consistency. In: Advances in Neural Information Processing Systems, Massachusetts, pp. 321–328 (2004)
8. Wang, F., Zhang, C.: Label propagation through linear neighborhoods. IEEE Transactions on Knowledge and Data Engineering **20**(1), 55–67 (2008)
9. Cheng, H., Liu, Z., Yang, J.: Sparsity induced similarity measure for label propagation. In: IEEE International Conference on Computer Vision, pp. 317–324. IEEE Press, Kyoto (2009)
10. Yan, S., Wang, H.: Semi-supervised learning by sparse representation. In: SIAM International Conference on Data Mining, pp. 792–801. SIAM Press (2009)
11. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
12. Donoho, D., Elad, M.: Maximal sparsity representation via $\ell 1$ minimization. Proceedings of the National Academy of Sciences of the United States of America **100**(50), 2197–2202 (2003)
13. Wright, J., Yang, A., Ganesh, A.: Robust face recognition via sparse representation. IEEE Transaction on Pattern Analysis and Machine Intelligence **312**(2), 210–227 (2009)
14. Wu, M., Scholkopf, B.: Transductive classification via local learning regularization. In: Proc. 11th Int. Conf. Artif. Intell. Statist., pp. 1529–1536. AAAI Press (2007)