Semi-supervised feature learning for hyperspectral image classification

Pengfei Zhang[a], Liujuan Cao[*a], Cheng Wang[a] , Jonathan Li[a],
[a] Fujian Key Laboratory of Sensing and Computing for Smart Cities, Department of Computer Science, Xiamen University, Xiamen, China 361005

ABSTRACT

Hyperspectral image has high-dimensional Spectral–spatial features, those features with some noisy and redundant information. Since redundant features can have significant adverse effect on learning performance. So efficient and robust feature selection methods are make the best of labeled and unlabeled points to extract meaningful features and eliminate noisy ones. On the other hand, obtaining sufficient accurate labeled data is either impossible or expensive. In order to take advantage of both precious labeled and unlabeled data points, in this paper, we propose a new semi-supervised feature selection method, Firstly, we use labeled points are to enlarge the margin between data points from different classes; Secondly, we use unlabeled points to find the local structure of the data space; Finally, we compare our proposed algorithm with Fisher score, PCA and Laplacian score on HSI classification. Experimental results on benchmark hyperspectral data sets demonstrate the efficiency and effectiveness of our proposed algorithm.

Keywords: Feature selection, Hyperspectral, local preservation, dimension reduction, semi-supervised

1. INTRODUCTION

Hyperspectral image (HSI), satellite sensors collect imagery simultaneously in hundreds of narrow and contiguously spaced spectral bands, with wavelengths ranging from the visible spectrum to the infrared region (0.4–2.5μm). The Hyperspectral image data is a cube-like data set, which contains two spatial dimensions and one spectral. Hyperspectral image with larger spectral and a much higher spectral resolution[1]. It is common step to feature extraction before classification e.g., texture, shape and spectral features. However, one of the main problems with hyperspectral image classification is the high dimension, and not all of the involved features are significant for classification[2-4], since most of them are often redundant to each other and sometimes noisy sensing data, so dimensionality reduction for hyperspectral image data is important. Dimensionality reduction is to find a way to encode projecting high dimensional data into lower dimensional vector without losing important information process. In general, dimensionality reduction methods for hyperspectral images can be grouped into two classes, feature selection[5] and feature extraction[6]. The first approach is to identify those attributes that have no contribution to the classification task and ignore them. The second approach is to find a transformation from high dimension into a lower dimension feature space.

In general, all dimension reduction methods preserve the physical meaning of HSI. Feature selection methods can be classified into supervised, semi-supervised and unsupervised methods based on whether label of feature is available or not. Dimension reduction methods include unsupervised approaches, such as locally linear embedding (LLE)[7], principal component analysis (PCA)[8,9] and Laplacian score[10], as well as supervised approaches, such as Fisher score[11], least squares regression(LSR)[12]. There are numerous variants of these techniques. Traditional supervised dimensionality reduction methods have been demonstrated to be effective when sufficient data are provided. However, it only focuses on labeled points data and ignores the local unlabeled points geometric structure of data. On the other hand these models often suffer from the problem of lacking sufficient data so that it is difficult to build effective models. Recently, spectral graph theory[13] and manifold learning[14] technique were proposed. Local preservation technique preserves geometric structure of data through a nearest neighbor graph on sample of data points. We call this matrix: local geometric matrix. Local geometric matrix refers to the similarity between every two samples in a training dataset. It can be calculated by using a predefined similarity measure e.g. Gaussian, Spectrum and Mismatch kernels.

Above analysis motivates us to develop a new semi-unsupervised dimensionality reduction approach for dealing with the high-dimensional and small-sized labeled data via preserve local structure. Unlike supervised methods[15] which makes

* caoliujuan@gmail.com

2nd ISPRS International Conference on Computer Vision in Remote Sensing (CVRS 2015), edited by Cheng Wang, Rongrong Ji, Chenglu Wen, Proc. of SPIE Vol. 9901, 99010F · © 2016 SPIE · CCC code: 0277-786X/16/$18 · doi: 10.1117/12.2234855

use of only labeled data points and unsupervised methods[16] which makes use of only unlabeled data points, our proposed algorithm makes use of both labeled and unlabeled data points, which tries to discover both geometrical and discriminant structure in the data. Specifically, we construct two graphs, same-class samples graph and unlabeled samples graph. The same-class samples graph connects data points which share the same label, while the unlabeled samples graph connects data points which are close to each other but there is no label available. The goal is to find both geometrical and discriminant structures of data. The contributions of the proposed solution for performing dimensionality reduction on the high-dimensional and small-size data are presented as follows:

(1) Comparing to different types feature dimension reduction methods like PCA, Fisher score and Laplacian score, the computation has a higher accuracy.

(2) Unlike previous feature selection methods such as Fisher score and Laplacian score, our approach makes use of both of labeled and unlabeled data points.

(3) Comparing to feature extraction methods like PCA and LPP, our method no need extra computation for original test sample.

The remainder of this paper is organized as follows. In Section 2, we provide the proposed algorithm in detail, including the extraction three kinds of hyperspectral image features. In Section 3, we discusses the experimental results on hyperspectral data set. And followed by the conclusion in Section 4.

2. THE ALGORITHM

This algorithm can be divided into three main steps. Firstly, three kinds of hyperspectral image feature extraction methods are introduced. Secondly, the proposed feature selection method is employed to form new low-dimensional vector. Final, SVM classifier is used for test sample classification.

2.1 Hyperspectral image feature extraction

In this paper, three kinds of features are introduced. Each feature is represented as a single vector. Three kinds of feature comprise the spectral value feature, the Gabor texture feature, and the shape feature.

Spectral Value Feature: The spectral feature of a hyperspectral image is acquired by arranging the digital number (DN) of all of the l bands.

$$\mathrm{S}_{Spectral} = [\mathrm{s}_1, \mathrm{s}_2, \ldots \mathrm{s}_k, \ldots, \mathrm{s}_l]^{'} \in R^l \tag{1}$$

where s_k denotes the value in band i.

Gabor Texture Feature: The Gabor wavelet filter, whose impulse response is defined by a Gaussian envelope and a complex plane wave, has been widely used in HIS analysis. In this paper, we perform a 2-D Gabor wavelet transform corresponding to the orientation and scale of the physical structures on the first principal component analysis (PCA) image, denoted as I of the hyperspectral image , to extract the Gabor texture feature, the generalized 2-D Gabor function can be defined as follows:

$$G_{s,d}(\mathrm{x},\mathrm{y}) = \mathrm{G}_{\bar{v}}(\bar{x}) = \frac{\|\bar{v}\|}{\sigma^2}.\mathrm{e}^{-\frac{\|\bar{v}\|^2.\|\bar{x}\|^2}{2\sigma^2}}.\left[e^{i.\bar{v}.\bar{x}} - e^{-\frac{\sigma^2}{2}}\right] \tag{2}$$

where $\bar{x} = (\mathrm{x},\mathrm{y})$ is the spatial domain variable and the frequency vector $\bar{v}(\mathrm{s},\mathrm{d}) = (\pi / 2 f^s).\mathrm{e}^{i(\pi d/8)}$, in which f = 2, s = 0, 1, . . . , 4 and d = 0, 1, . . . , 11, which determines the 5 scales and the 12 directions of the Gabor function. The number of oscillations under the Gaussian envelope is determined by $\delta = 2\pi$. The Gabor texture feature contains the magnitude information in the first PCA[8] image I with the Gabor function of the specific scale parameter s and direction parameter d

$$F_{s,d}(x,y) = G_{s,d}(x,y) * I(x,y) \tag{3}$$

The texture feature of a pixel located on (x, y) is obtained by

$$s_{Texture} = \left[\mathrm{F}_{1,1}(\mathrm{x},\mathrm{y}),...,\mathrm{F}_{s,d}(\mathrm{x},\mathrm{y}) \right]' \in R^{s*d} \tag{4}$$

PSI: Shape Feature: The pixel shape index method (PSI) is adopted to describe the shape feature in a local area surrounding a certain pixel. Pixel shape feature extraction of a specific pixel consists of the following two steps.

Step 1) Extension of direction lines based on gray level similarity. The pixel homogeneity is defined using the following method:

$$PH_i = \sum_{s=1}^{n} \left| p_s^{cen} - p_s^{sur} \right| \tag{5}$$

The *i*th direction line is extended if the following conditions are met.

1) PH_i is less than a predefined threshold T1.

2) The total number of pixels in this direction line is less than another predefined threshold T2.

Step 2) Length of direction line: The PSI in the ith direction is calculated by the length of the direction line di. Then, the shape feature is achieved by

$$s_{Shape} = \left[d_1, d_2, ..., d_p \right]' \in R^p \tag{6}$$

in which p is the total number of all directions.

2.2 Local structure preservation framework

Construct the feature matrix X, shown as $\mathrm{X}=\left[S'_{Spectral,1:n}, S'_{\mathrm{Texture},1:n}, S'_{Shape,1:n} \right] \in R^{n\times d}$, whose columns $x_i \in X$ correspond to data instances and rows to features, Let $\mathrm{X}_l=\left[x_1, x_2, ..., x_i \right]' \in R^{l\times d}$ and $\mathrm{X}_u=\left[x_1, x_2, ..., x_u \right]' \in R^{u\times d}$ be the labeled and unlabeled data matrices. Since a few label information is unavailable for semi-unsupervised feature selection, Let $Y_l=\left[y_1, y_2, ..., y_l \right]' \in R^{l\times c}$ denote the labels of X_l , where and $y_i \in \{0,1\}^{1\times c}$ is the class label of $x_i \in X_l$. As the corresponding label matrix. We wish to directly learn a transform matrix $\mathrm{W}^{d\times c}$. By which we transform the high-dimensional data X to low-dimensional data. In order to maximally preserve the global structure of X with XW, an $l_{2,1}$-norm regularizer of W is imposed to enforce row sparsity, which has an effect of feature selection and helps to avoid selecting redundant features. The objective function of the proposed method is defined as follows:

$$\arg\min_{W,b,M} f = \left\| Y - X_l W - 1_n b + B \circ M \right\|_F^2 + \lambda \left\| W \right\|_{2,1} \tag{7}$$

where W is the transform matrix, b is the bias term, and 1_n is an n × 1 constant vector where all the elements are equal to 1. $Y \in R^{l\times c}$ is the class indicator matrix, where only these elements corresponding to the data in the jth class are equal to 1 and other remaining elements are 0. Thus, each column vector of Y actually stipulates a type of binary regression with target "+1" for the jth class and target "0" for the remaining classes. For the 0/1 output, we drag these binary outputs far away along two opposite directions by imposing a positive slack variable ε. we hope the output becomes 1 + ε for 1 and −ε for 0. In this way, the distance between two data points from different classes will be enlarged. Let $B \in R^{l\times c}$ be a constant matrix, in which the ith row and jth column element is defined as follows:

$$B_{ij}=\begin{cases}+1, & \text{if } y_{ij}=1\\ \\ -1, & \text{if } y_{ij}=0\end{cases} \tag{8}$$

Where B_{ij} = +1 means it points to the positive direction and B_{ij} =−1 means it points to the negative direction. The ε matrix $M \in R^{l\times c}$. Each of its elements ε_{ij} is nonnegative, this treatment could help to enlarge the distance between classes after the data points are mapped. Now, we can rewritten (6) model as follows:

$$\arg\min_{W,b} f = \left\|Y - X_l W - 1_n b\right\|_F^2 + \lambda \left\|W\right\|_{2,1} \tag{9}$$

where ∘ is a Hadamard product operator of matrices.

The underlying local structure information is ignored. In order to characterize the underlying local structure, many manifold learning algorithms have been proposed, such as Local Linear Embedding (LLE) and ISOMAP . Many unsupervised feature selection algorithms, use various graphs to capture the local geometry of unlabeled points. For semi-supervised, both labeled and unlabeled dada should be considered. So we employ a new method based on widely used models, combining Gaussian and Mismatch kernels, let G_n denote a graph with n nodes. The i th node corresponds to the data point x_i . Put an edge between nodes *i* and *j* if they share the same label, or if one of them is unlabeled but they are sufficiently close to each other. Define a weight matrix S as follows:

$$S_{i,j}=\begin{cases}1 & \text{if nodes i and j share the same label}\\ \exp\left(-\left\|x_i - x_j\right\|^2 / t\right) & \text{if node i or j is}\\ & \text{unlabeled, but}\\ & \text{node i} \in \text{KNN(j)}\\ & \text{or node j} \in \text{KNN(i)}\\ 0 & \text{otherwise}\end{cases} \tag{10}$$

where KNN(i) denotes the set of k nearest neighbors of node i and KNN(j) denotes the set of *k* nearest neighbors of node *j*. The LPP algorithm aims to minimize the following cost function:

$$\arg\min_{W} f = \sum_{i,j=1}^{n} \left\|W^{'} x_i - W^{'} x_j\right\|^2 S_{ij} \tag{11}$$

Combining above two aspect analysis, we propose the objective of our framework as follows:

$$\begin{aligned}\arg\min_{W,b,M} f &= \left\|Y - X_l W - 1_n b + B \circ M\right\|_F^2 \\ &+\mu \sum_{i,j=1}^{n} \left(W^{'} x_i - W^{'} x_j\right)^2 S_{ij} + \lambda \left\|W\right\|_{2,1}\end{aligned} \tag{12}$$

Where μ and λ are tradeoff parameters (μ , λ >0). Equation (11) can be rewritten as

$$\begin{aligned}\arg\min_{W,b,M} f &= \left\|Y - X_l W - 1_n b + B \circ M\right\|_F^2 \\ &+\mu \mathrm{tr}\left(W^{'} X^{'} L X W\right) + \lambda \left\|W\right\|_{2,1}\end{aligned} \tag{13}$$

where the matrix L is often called Laplacian matrix and

$$L = D - S \tag{14}$$

D is a diagonal matrix, with its (i, i) element equal to the sum of the ith row of the weighted matrix

$$D_{ii} = \sum_{j=1}^{n} S(i, j) \tag{15}$$

Where S is computed by (10).

3. RESULTS AND DISCUSSION

In this experiment, Support vector machine (SVM) classifier[17], which has been reported to be effective in the classification of hyper dimensional feature sets, was used to classify samples based on the selected features. The Washington DC Mall dataset used to evaluate the performance of our proposed method. In this dataset there were 210 bands in the 0.4 to 2.4 μm region of the visible and infrared spectrum. This data set contains 1208 scan lines with 307 pixels in each scan line. This data and reference data are show in Figure1 and Figure2.

Figure 1. RGB composite of the Washington DC Mall data (channels 60, 27 and 17)

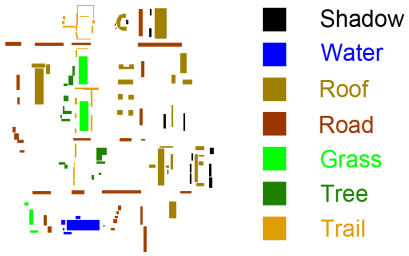

Figure 2. Reference data

Considering that our proposed method belongs to the semi-supervised learning problem, to validate the effectiveness of our algorithm for feature selection, we compare it with the different type dimension reduction methods. The following feature dimension reduction methods are used to be compared.

Table 1. Class-specific accuracies and standard deviation in percentage for various features.

	Method				
	Fisher score	PCA	Laplacian score	Our Method	All Feature
Road	0.9688	0.9728	0.9823	0.9790	0.9741
Grass	0.9888	0.9881	0.9811	0.9965	0.9507
Water	0.9529	0.9283	0.9164	0.9334	0.9358
Trail	0.9350	0.9470	0.9465	0.9663	0.9010
Tree	0.9906	0.9768	0.9941	0.9887	0.9843
Shadow	0.9616	0.9661	0.9656	0.9746	0.9465
Roof	0.9864	0.9995	0.9926	0.9938	0.9981
OA	0.9751	0.9754	0.9738	0.9843	0.9675
Kappa	0.9693	0.9696	0.9676	0.9806	0.9598

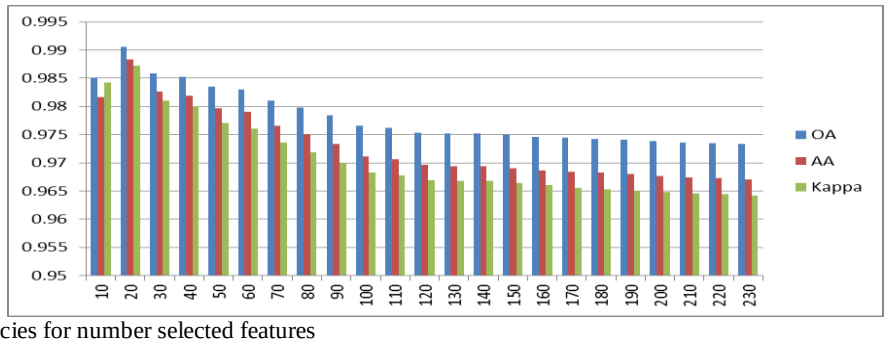

Figure 3. accuracies for number selected features

4. CONCLUSION

In this paper, we propose a novel semi-supervised feature selection approach for dealing with the high-dimensional and small-sized labeled data. This method makes use of both labeled and unlabeled data points to find the local structure of the data. Therefore, it is more effective to select those most discriminative and informative features. As a result, it can select the most representative features. The experimental results validate that the new method achieves significantly higher performance for classification.

ACKNOWLEDGMENT

This work is supported by the Natural Science Foundation of China (No.6402388)

REFERENCES

[1] Chang, C.I., [Hyperspectral Data Exploitation: Theory and Applications], John Wiley & Sons, (2007).

[2] Li, W., Prasad, S., Fowler, J. E., and Bruce, L. M., "Locality-preserving dimensionality reduction and classification for hyperspectral image analysis," IEEE Trans. Geosci. Remote Sens 50(4), 1185–1198 (2012).

[3] Lee, C. and Landgrebe, D., "Analyzing high-dimensional multispectral data," IEEE Trans. Geosci. Remote Sens 31(4), 92–800 (1993).

[4] Shi, Q., Zhang, L. and Du, B., "Semi-supervised discriminative locally enhanced alignment for hyperspectral image classification," IEEE Trans. Geosci. Remote Sens 51(9), 4800–4815 (2013).

[5] Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L. A., [Feature Selection for Knowledge Discovery and Data Mining], Springer, (1998).

[6] Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L. A., [Feature Extraction: Foundations and Applications], Springer, (2006).

[7] Roweis, S. T. and Saul, L. K. , "Nonlinear dimensionality reduction by locally linear embedding," Science 290(22), 2323-2626(2000).

[8] Jolliffe, I., [Principal Component Analysis] Springer, (1986).

[9] Lipovetsky, S. , "PCA and SVD with nonnegative loadings," Pattern Recognit.42 (1) , 68–76(2009).

[10] He, X., Cai, D. and Niyogi, P. , "Laplacian score for feature selection," Advances in Neural Information Processing Systems, (507-514)2005.

[11] Duda, R. O., Hart, P. E. and Stork, D. G., [Pattern Classification]. Wiley-Interscience Publication, (2001).

[12] Xiang, S., Nie, F., Meng, G., Pan, C. and Zhang, C., "Discriminative least squares regression for multiclass classification and feature selection," IEEE Trans. Neural Networks and Learning Systems 23(11), 1738-1754(2013).

[13] Chung, F. R., [Spectral Graph Theory], American Mathematical Soc. (1997).

[14] Belkin, M. and Niyogi, P., "Laplacian eigenmaps and spectral techniques for embedding and clustering, " NIPS 14, 585-591(2001).

[15] Peng, H., Long, F. and Ding, C., "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Analysis and Machine Intelligence 27(8), 1226–1238(2005).

[16] Yang, Y., Shen, H. T., Ma, Z., Huang, Z. and Zhou, X. , "l2, 1 -norm regularized discriminative feature selection for unsupervised learning," Proceedings-International Joint Conference on Artificial Intelligence 22(1), 1589–1594(2011).

[17] Chang, C. C., & Lin, C. J. , "LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology 2(3), 1–27(2011)