A MODIFIED STOCHASTIC NEIGHBOR EMBEDDING FOR COMBINING MULTIPLE FEATURES FOR REMOTE SENSING IMAGE CLASSIFICATION

Lefei Zhang [a, *], Liangpei Zhang [a], Dacheng Tao [b] and Xin Huang [a]

[a] The State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China.
[b] Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney, Sydney, NSW 2007, Australia.

Commission ICWG III/VII

KEY WORDS: Multiple Features, Dimensional Reduction, Classification.

ABSTRACT:

In remote sensing image interpretation, it is important to combine multiple features of a certain pixel in both spatial and spectral domains to improve the classification accuracy, such as spectral signature, morphological property, and shape feature. Therefore, it is essential to consider the complementary property of different features and combine them in order to obtain an accurate classification rate. In this paper, we introduce a multi-feature dimension reduction algorithm under a probabilistic framework, modified stochastic neighbor embedding (MSNE). For each feature, a probability distribution is constructed based on SNE, and then we alternatively solve SNE and learn the optimal combination coefficients for different features in optimization. Compared with conventional dimension reduction strategies, the suggested algorithm can considers spectral, morphological and shape features of a pixel to achieve a physically meaningful low-dimensional feature representation by automatically learn a combination coefficient for each feature adapted to its contribution to subsequent classification. In experimental section, classification results using hyperspectral remote sensing image (HSI) show that this modified stochastic neighbor embedding can effectively improve classification performance.

1. INTRODUCTION

In hyperspectral remote sensing image (HSI) classification, it is important to employ multiple features of different types to represent a pixel's information, such as spectral signature (Plaza, Benediktsson et al. 2009), morphological property (Soille and Pesaresi 2002), and shape feature (Segl, Roessner et al. 2003). Previous literatures have reported that combine multiple features of a certain pixel in both spatial and spectral domains could improve the land cover classification accuracy (Landgrebe 1980; Puissant, Hirscha et al. 2005). Since each feature can be viewed as a vector in a high-dimensional feature space, therefore, it is essential to consider the complementary property of different features and combine them in order to obtain an accurate classification rate. A conventional approach is simply concatenating different features into a long vector and applying a particular dimension reduction technique, such as Principal Component Analysis (PCA) (Jolliffe 2002), Fisher Discriminant Analysis (FDA) (Mika, Ratsch et al. 1999), Locally Linear Embedding (LLE) (Roweis and Saul 2000), Laplacian Eigenmaps (LE) (Belkin and Niyogi 2003), and so on. However, this direct feature concatenation strategy intrinsically assumes that different features are distributed in a unified feature space, although they are not, because they have different physical meanings and statistical properties (Xie, Mu et al. 2011). Therefore, it is unreasonable to use simple concatenation to combine different features for subsequent processing.

To overcome this problem, in this paper, we introduce a multi-feature dimension reduction algorithm under a probabilistic framework, stochastic neighbor embedding (Hinton and Roweis 2003). For each feature, a probability distribution is constructed based on stochastic neighbor embedding (SNE), and then we alternatively solve SNE and learn combination coefficients, i.e., weighting factors for different features in optimization. In summary, this modified stochastic neighbor embedding (MSNE): (1) considers texture, morphological, shape and spectral signature features of a pixel to achieve a physically meaningful low-dimensional feature representation for the subsequent classification, and (2) automatically optimize the combination weighting factors for different features according to their contributions for the subsequent classification, which indicate the complementary property of different features. The remainder of this paper is organized as follows. In Section 2, we introduce the multiple feature combination strategy in detail, including the spectral and spatial features extraction of HSI and the full optimization of modified stochastic neighbor embedding algorithm. Then, the hyperspectral remote sensing image classification results are reported in Section 3, followed by the conclusion.

2. MODIFIED STOCHASTIC NEIGHBOR EMBEDDING ALGORITHM

The proposed multiple feature combination strategy can be divided into two main components. In the first step, three kinds of features of HSI are introduced. Then the MSNE algorithm is employed to obtain the final low dimensional representation.

2.1 Spectral and spatial features extraction

(1) Spectral Feature: The spectral feature of a pixel in HSI is obtained by arranging its digital number (DN) in all of *l* bands:

$$\text{Spectral} = \left[v_1, v_2, \cdots, v_l\right]^{\mathrm{T}} \tag{1}$$

in which v_i denotes the DN in band *i*.

(2) Morphological Feature: The Differential Morphological Profiles (DMPs) (Benediktsson, Palmason et al. 2005) are defined as a vector where the measures of the slope of the

[*] Corresponding author. E-mail addresses: zhanglefei@whu.edu.cn.

opening-closing profiles are stored for every step of an increasing SE series:

$$\mathrm{DMP}_{\lambda} = \left\{ \begin{array}{l} \mathrm{DMP}_{\lambda_s} : \mathrm{DMP}_{\lambda_s} = \left| \mathrm{MP}_{\lambda_s} - \mathrm{MP}_{\lambda_{s-1}} \right|, \\ \lambda \in [\gamma, \varphi], s \in [1, S] \end{array} \right\} \qquad (2)$$

in which γ_s and φ_s be the morphological opening and closing operators by reconstruction with structural element $\mathrm{SE} = s$. MP_{γ} and MP_{φ} are the opening and closing profiles of the image *I* (Huang and Zhang 2009).

(3) Shape Feature: The pixel shape index (PSI) based method is adopted to describe the shape feature in a local area (Zhang, Huang et al. 2006):

$$\mathrm{Shape} = \left[d_1, d_2, \cdots, d_p \right]^{\mathrm{T}} \qquad (3)$$

in which d_i is the length of the *i*th direction line measured by the pixel homogeneity of the central pixel and the surrounding pixels.

2.2 Modified Stochastic Neighbor Embedding

The proposed MSNE algorithm finds a low dimension representation $y \in \mathrm{R}^d$ of input multiple features $f^{(k)} \in \mathrm{R}^{L_k} \big|_{k=1}^{m}$, in which *m* is the number of features. In order to deal with out-of-sample problem (Bengio, Paiement et al. 2004), only a subset of samples in the HSI are used as input data of MSNE. Suppose given a multiple features data set of *n* samples, e.g., $F = F^{(k)} \in \mathrm{R}^{L_k \times n} \big|_{k=1}^{m}$, wherein $F^{(k)}$ is the *k*th feature matrix. MSNE first builds a probability distribution for each feature based on SNE, then, we alternatively solve SNE and learn the optimal combination coefficients to obtain the solution of MSNE. Finally, the linear transformation for MSNE feature mapping is solved by linear regression, and the extracted feature representation in reduced feature space is achieved by the such linear transformation for each pixel of HSI, respectively.

(1) Stochastic Neighbor Embedding: for the *k*th feature matrix, suppose that we have input high-dimensional data samples $f_i^{(k)} \in \mathrm{R}^{L_k} \big|_{i=1}^{n}$, SNE defined the the normalized pairwise distances as a joint probability distribution over input sample pairs, which are represented in a symmetric matrix $P \in \mathrm{R}^{n \times n}$ (Hinton and Roweis 2003). Similarly, in the output low-dimensional feature space, we define the probability distribution *Q*:

$$Q_{ij} = \frac{\left(1 + \left\| y_i - y_j \right\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \left\| y_k - y_l \right\|^2\right)^{-1}} \qquad (4)$$

The aim of SNE is to match these two distributions *P* and *Q* as well as possible, which is achieved by minimizing the Kullback-Leibler divergences (Kullback and Leibler 1951) between the two distributions over all data points:

$$\min_y KL(P, Q) = \min_y \sum_i \sum_j P_{ij} \log \frac{P_{ij}}{Q_{ij}} \qquad (5)$$

To find the solution of (5), the gradient with respect to *y* is:

$$\begin{aligned} &\partial KL(P,Q) / \partial y_i \\ &= 4 \sum_j \left(P_{ij} - Q_{ij}\right)\left(y_i - y_j\right)\left(1 + \left\| y_i - y_j \right\|^2\right)^{-1} \end{aligned} \qquad (6)$$

Given the gradient (6), there are many possible ways to minimize (5). In this paper, we employ the method suggested in (Maaten and Hinton 2008).

(2) MSNE: We assume that the final probability distribution of the input multiple features is a linear combination of all the joint probability distribution matrices, i.e.,

$$P = \sum_{k=1}^{m} \omega_k P^{(k)} \qquad (7)$$

where ω_k is the nonnegative weight of each features with conditions that $\omega_k > 0$ and $\sum_k \omega_k = 1$, and $P^{(k)}$ is a joint probability distribution matrix computed by the *k*th input feature. The larger ω_k is, the more important is the role of the *k*th feature in constructing final probability distribution (7). In order to automatically optimize ω_k for each feature according to its unique contribution, we adopt alternating optimization to optimize the objective function with respect to both *y* and ω simultaneously. The final objective function of MSNE is given by:

$$\begin{aligned} &\min_{y,\omega} \sum_i \sum_j P_{ij} \log \frac{P_{ij}}{Q_{ij}} \\ &P_{ij} = \sum_{k=1}^{m} \omega_k P_{ij}^{(k)} \end{aligned} \qquad (8)$$

In every round of iteration, we first fix ω to find low-dimensional embedding *y*. By constructing the final probability distribution (7), we can use *t*-SNE (Maaten and Hinton 2008) to find low-dimensional embedding.

Then we first fix *y* to optimize ω. Here we can see that the current objective function is a linear programming (LP) with respect to ω. Since the optimal solution of LP will be always at the vertex of the linear feasible region, the solution of ω must be only one of ω_k equal to 1 and others equal to zeros. To avoid this problem, we add an l_2 norm regularization term into the current objective function:

$$\min_{\omega} \sum_i \sum_j P_{ij} \log \frac{P_{ij}}{Q_{ij}} + r \left\| \omega \right\|^2 \qquad (9)$$

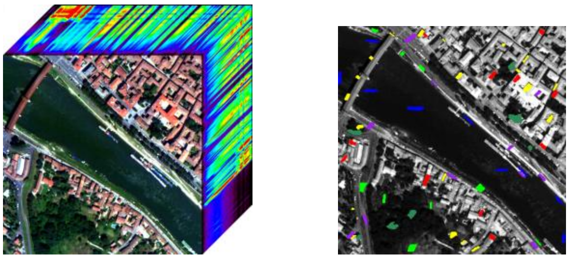
The optimization (9) is convex and could be minimised by using Nesterov’s accelerated first-order method (Nesterov 2005).

(3) Linearization. MSNE tries to train an optimal subspace for original multiple features. However, this feature mapping is always nonlinear and implicit (Zhang, Zhang et al. 2012). In HSI classification, it’s impossible to train such low-dimensional subspace using all the pixels features, because the size of the joint probability distribution matrices $P^{(k)}$ scale with the number of input samples, thus the suggested MSNE suffers from the out-of-sample problem. In this paper, only a subset of samples in the HSI are used as input data of MSNE, then, a explicit linear projection matrix trained by MSNE is applied to approximately construct the low-dimensional representation. Based on this subset of samples $F = F^{(k)} \in \mathrm{R}^{L_k \times n} \big|_{k=1}^{m}$ and the MSNE output low-dimensional representation $Y \in \mathrm{R}^{d \times n}$, the linear transformation for MSNE feature mapping is solved by linear regression:

$$U = Y\left(F^{\mathrm{T}} F\right)^{-1} F^{\mathrm{T}} \qquad (10)$$

3. EXPERIMENT AND ANALYSIS

The experiment and analysis were conducted on a publicly available airborne hyperspectral data set, which was acquired by the sensor ROSIS on July 8, 2002, of the urban test area of Pavia, Northern Italy (45.11N, 9.09E). The subset of the Pavia city data is shown in Fig. 1; its size is 400×400 pixels. Some channels were removed due to noise and the remaining 102 spectral dimensions from 0.43 to 0.83 um were processed. This data set was provided by the Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society.



(a) (b)

Fig. 1. Pavia city data set and reference data.

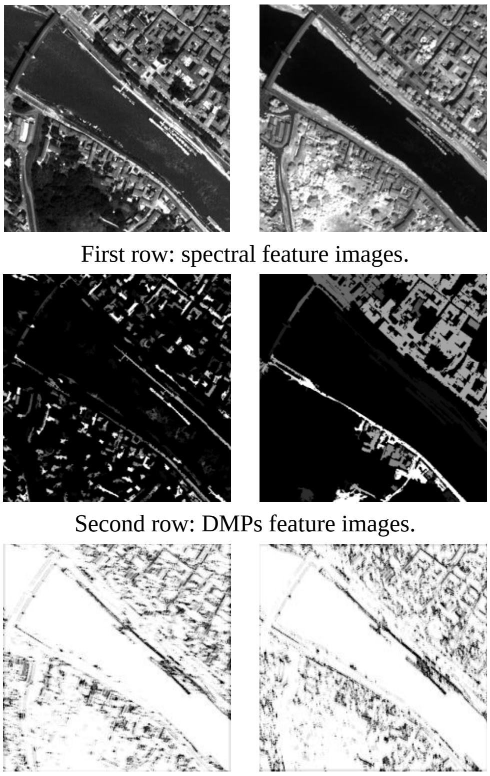
Based on the spectral and spatial features extraction mentioned in section 2.1, we have the following multiple features: 102-dimensional spectral feature vector, 40-dimensional DMPs feature vector and 20-dimensional shape feature vector for each pixel in HSI, respectively. Some of the feature images are shown in Fig. 2. The total number of samples in the data set is *N*=400×400 pixels; *n*=1200 samples (0.75% of all samples) were randomly sampled from *N* and were used to construct the input feature matrix for MSNE. The proposed MSNE as well as PCA (Jolliffe 2002), LPP (He and Niyogi 2004) and SNE algorithm (Hinton and Roweis 2003) are conducted to obtain the low dimension feature representation of multiple features. The support vector machine (SVM) classifier (Mountrakis, Im et al. 2011) was used to interpret the above processed feature data. In SVM classification step, the training samples were randomly selected from the reference data, while we use the rest of reference data as test samples. The numbers of train and test samples are listed in Table I.

TABLE I

NUMBERS OF REFERENCE DATA AND CLASS SPECIFIC ACCURACIES

	Train	Test	PC	LPP	SNE	MSNE
Water	30	756	99.25	99.13	**99.39**	99.13
Road	30	767	90.03	95.66	94.63	**96.69**
Roof	30	794	82.98	91.70	88.41	**95.76**
Shadow	30	698	88.28	92.34	92.98	**96.72**
Grass	30	907	94.16	93.93	95.66	**96.33**
Tree	30	949	87.04	85.62	87.57	**93.48**
OA	180	4871	91.27	92.12	92.71	**95.46**
Kappa	-	-	0.8951	0.9053	0.9124	**0.9455**

We first investigated the complementary property of the above multiple features on Pavia city data set. Fig. 2 shows the spectral, DMPs and shape feature for different pixels, these pixels correspond to varies classes, e.g., road, roof, grass and tree, respectively. Usually, spectral signature is the most discriminative feature in HSI classification, however, in Fig. 2, pixel pair road and roof have a very similar spectral signature. We might still distinguish them according to DMPs and shape features. So this complementary property of the multiple features on HSI data set provides the information to potentially improve the classification performance. The same phenomenon could be observed based on the pixel pair grass and tree.

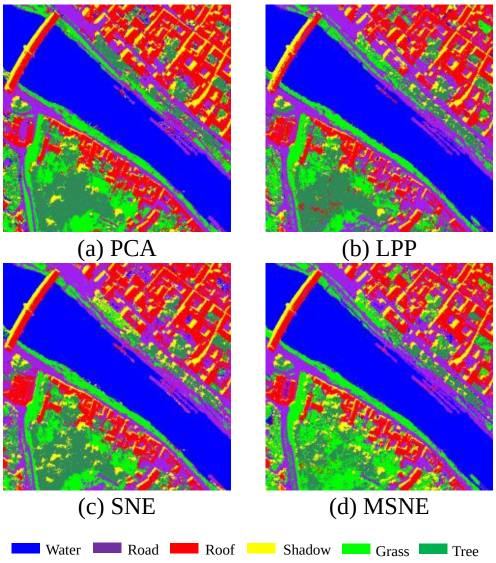


First row: spectral feature images.

Second row: DMPs feature images.

Third row: shape feature images.

Fig. 2. Multiple features of the Pavia city data set.



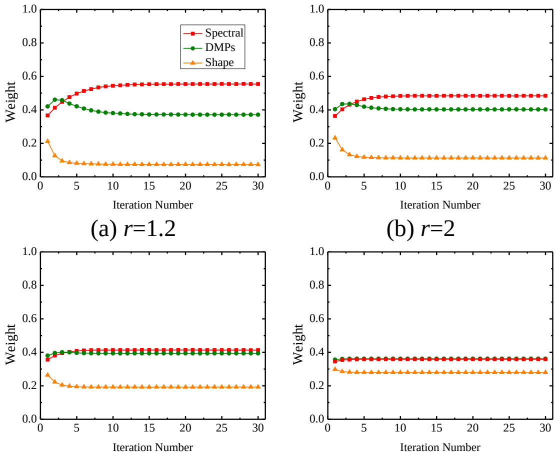
(a) PCA (b) LPP

(c) SNE (d) MSNE

Fig.3. (a)-(d) Classification maps of Pavia city data set obtained using features of PCA, LPP, SNE, and MSNE, respectively.

Four different feature based classification results are compared in Figs. 3 (a)-(d). In all dimensional reduction methods, the size of reduced feature space is fixed at 25. In Fig. 3, the proposed MSNE based classification achieved the best performance. Compared to the other three dimension reduction methods in Figs. 3(a), (b) and (c), the proposed MSNE shows a good classification result. In order to evaluate thoroughly the different feature representations, the averaged classification accuracies of all classes in ten independent classification experiments are also compared in Table I. From Table I, improvements can be observed and MSNE obtains the top classification rate in five classes and achieves the top OA and kappa coefficient.

Here we also investigated the affect of regularization parameter *r* in alternating optimization step. Figs. 4 (a)-(d) describe the relationship of regularization parameter *r* and combination weights in spectral, DMPs and shape feature. We can see the spectral feature is the most discriminating feature for the Pavia city data set. It also can be observed that if *r* is close to 1, the combination weights is very sparse, thus the most discriminative feature will be set to large coefficient. If *r* is increased to infinity, different features will share the similar weights for the subsequent feature combination. Therefore, the selection of regularization parameter *r* should be based on the complementary properties of input features. If the available features are complementary to each other, a larger *r* is preferred to guarantee that all features properly contribute to the subsequent classification; otherwise, we can choose a small *r*.



(a) *r*=1.2 (b) *r*=2

(c) *r*=4 (d) *r*=10

Fig. 4. The affect of regularization parameter *r* and combination weights in each feature.

4. CONCLUSION

In this paper, we introduce a multi-feature dimension reduction algorithm under a probabilistic framework which could considered the spectral, DMPs and shape features of a pixel to achieve a physically meaningful low dimensional representation for an effective and accurate classification. For each input feature, a probability distribution is constructed based on SNE, and then we alternatively solve SNE and learn the optimal combination coefficients for different features in optimization. The linear transformation for MSNE feature mapping is achieved by linear regression in order to deal with out-of-sample problem in HSI classification. Experiment on the classification of ROSIS hyperspectral data sets demonstrate that the proposed approach could explore the complementary properties of different features and find an optimal low dimension representation for classification. The effect of the combination weights of each feature are also investigated. Our future work will explore how to select the optimal parameters in MSNE feature combination and reduction to obtain the best subsequent hyperspectral remote sensing classification accuracy.

REFERENCES

Belkin, M. and Niyogi, P., 2003. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Neural Computation 15(6): 1373-1396.

Benediktsson, J. A., Palmason, J. A. and Sveinsson, J. R., 2005. Classification of Hyperspectral Data From Urban Areas based on Extended Morphological Profiles. IEEE Transactions on Geoscience and Remote Sensing 43(3): 480-491.

Bengio, Y., Paiement, J. F., Vincent, P., Delalleau, O., Le Roux, N. and Ouimet, M., 2004. Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering. Advances in Neural Information Processing Systems, 177-184, The MIT Press.

He, X. and Niyogi, P., 2004. Locality Preserving Projections. Advances in Neural Information Processing Systems, 153-160, MIT Press.

Hinton, G. and Roweis, S., 2003. Stochastic Neighbor Embedding. Advances in Neural Information Processing Systems, 857-864, MIT Press.

Huang, X. and Zhang, L., 2009. A Comparative Study of Spatial Approaches for Urban Mapping Using Hyperspectral ROSIS Images ver Pavia City, Northern Italy. International Journal of Remote Sensing 30(12): 3205-3221.

Jolliffe, I. T., 2002. Principal Component Analysis. New York, USA, Springer.

Kullback, S. and Leibler, R. A., 1951. On Information and Sufficiency. The Annals of Mathematical Statistics 22(1): 79-86.

Landgrebe, D. A., 1980. The Development of a Spectral-Spatial Classifier for Earth Observational Data. Pattern Recognition 12(3): 165-175.

Maaten, L. v. d. and Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9: 2579-2605.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B. and Mullers, K.-R., 1999. Fisher Discriminant Analysis with Kernels. IEEE Signal Processing Society Workshop 41-48, Madison, WI , USA

Mountrakis, G., Im, J. and Ogole, C., 2011. Support Vector Machines in Remote Sensing: A Review. ISPRS Journal of Photogrammetry and Remote Sensing 66(3): 247-259.

Nesterov, Y., 2005. Smooth Minimization of Non-Smooth Functions. Mathematical Programming 103(1): 127-152.

Plaza, A., Benediktsson, J. A., Boardman, J. W., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Marconcini, M., Tilton, J. C. and Trianni, G., 2009. Recent Advances in Techniques for Hyperspectral Image Processing. Remote Sensing of Environment 113(1): 110-122.

Puissant, A., Hirscha, J. and Webera, C., 2005. The Utility of Texture Analysis to Improve Per-pixel Classification for High to Very High Spatial Resolution Imagery. International Journal of Remote Sensing 26(4): 733-745.

Roweis, S. T. and Saul, L. K., 2000. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290(22): 2323-2326.

Segl, K., Roessner, S., Heiden, U. and Kaufmann, H., 2003. Fusion of Spectral and Shape Features for Identification of Urban Surface Cover Types Using Reflective and Thermal Hyperspectral Data. ISPRS Journal of Photogrammetry and Remote Sensing 58(1): 99-112.

Soille, P. and Pesaresi, M., 2002. Advances in Mathematical Morphology Applied to Geoscience and Remote Sensing. IEEE Transactions on Geoscience and Remote Sensing 40(9): 2042-2055.

Xie, B., Mu, Y., Tao, D. and Huang, K., 2011. m-SNE: Multiview Stochastic Neighbor Embedding. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 41(4): 1088-1096.

Zhang, L., Huang, X., Huang, B. and Li, P., 2006. A Pixel Shape Index Coupled with Spectral Information for Classification of High Spatial Resolution Remotely Sensed Imagery. IEEE Transactions on Geoscience and Remote Sensing 44(10): 2950-2961.

Zhang, L., Zhang, L., Tao, D. and Huang, X., 2012. On Combining Multiple Features for Hyperspectral Remote Sensing Image Classification. IEEE Transactions on Geoscience and Remote Sensing 50(3): 879-893.