ISPRS Journal of Photogrammetry and Remote Sensing 83 (2013) 30–39

Contents lists available at SciVerse ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

A modified stochastic neighbor embedding for multi-feature dimension reduction of remote sensing images

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

[a] *State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, China*
[b] *Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology, 235 Jones Street, Ultimo, NSW 2007, Sydney, Australia*
[c] *Computer School, Wuhan University, China*

ARTICLE INFO

Article history:
Received 3 September 2012
Received in revised form 17 May 2013
Accepted 28 May 2013
Available online 21 June 2013

Keywords:
Hyperspectral image
Multiple features
Stochastic neighbor embedding
Dimension reduction
Classification

ABSTRACT

In automated remote sensing based image analysis, it is important to consider the multiple features of a certain pixel, such as the spectral signature, morphological property, and shape feature, in both the spatial and spectral domains, to improve the classification accuracy. Therefore, it is essential to consider the complementary properties of the different features and combine them in order to obtain an accurate classification rate. In this paper, we introduce a modified stochastic neighbor embedding (MSNE) algorithm for multiple features dimension reduction (DR) under a probability preserving projection framework. For each feature, a probability distribution is constructed based on *t*-distributed stochastic neighbor embedding (*t*-SNE), and we then alternately solve *t*-SNE and learn the optimal combination coefficients for different features in the proposed multiple features DR optimization. Compared with conventional remote sensing image DR strategies, the suggested algorithm utilizes both the spatial and spectral features of a pixel to achieve a physically meaningful low-dimensional feature representation for the subsequent classification, by automatically learning a combination coefficient for each feature. The classification results using hyperspectral remote sensing images (HSI) show that MSNE can effectively improve RS image classification performance.

1. Introduction

In recent years, the advances in earth observation technology, especially in hyperspectral (Landgrebe, 2002) and high-resolution (Acqua et al., 2004) remote sensing, have led to a growing availability of remotely sensed images. These earth observation data provide the opportunity to develop many important applications, which are closely related to the accurate classification of images (Zhu and Blumberg, 2002; Stavrakoudis et al., 2011; Walter and Luo, 2011), e.g., land-cover monitoring, urban planning and growth regulation, environmental damage assessment, military reconnaissance, and so on (Campbell, 2000; Chang, 2003). Image classification is an important issue in remote sensing and other applications. In the remote sensing literature, there are two main groups of approaches for image classification: supervised image classification (Liu et al., 2011; Shao and Lunetta, 2012) and unsupervised image classification (Baraldi and Parmiggiani, 1995). Generally speaking, supervised classification often achieves a higher classification accuracy than unsupervised classification, due o the consideration of discriminative information from the given training samples (Zhong and Zhang, 2012). However, in this case, it is common to perform feature extraction and dimension reduction (DR) (Conese and Maselli, 1993; Harsanyi and Chang, 1994; Zhao and Maclea, 2000; Phillips et al., 2009) before classification, which helps to: (1) remove the redundancy among features, (2) decrease the computational cost, and (3) avoid the Hughes phenomenon (Hughes, 1968).

For the input high-dimensional feature in the original feature space, e.g., the *l*-dimensional feature vector in the spectral domain (*l* is the number of spectral channels of the remote sensing image), the DR algorithm aims to find a feature mapping from the original feature space to a lower-dimensional subspace in which some specific desired information can be preserved as much as is possible. For example, the best-known DR algorithm, principal component analysis (PCA) (Jolliffe, 2002), finds a subspace of principal components in accordance with the maximum variance of the input feature matrix. Another popular DR technology, linear discriminant analysis (LDA) (McLachlan, 1992), finds the low-dimensional subspace where the different classes of samples remain well separated after projection. Considering that PCA and LDA are global linear algorithms, which do not work well in nonlinear distributed data conditions (Zhang et al., 2009), some researchers have also proposed nonlinear DR algorithms for remote sensing data. Such

* Corresponding author.
E-mail address: zlp62@whu.edu.cn (L. Zhang).

0924-2716/$ - see front matter © 2013 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS) Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.isprsjprs.2013.05.009

algorithms include local linear embedding (LLE) (Bachmann et al., 2005), isometric mapping (ISOMAP) (Bachmann et al., 2005), supervised local tangent space alignment (SLTSA) (Ma et al., 2010), local Fisher's discriminant analysis (LFDA) (Li et al., 2012), and spherical stochastic neighbor embedding (SSNE) (Lunga and Ersoy, 2013).

It should be emphasized that, in the aforementioned works, the adopted DR algorithms only deal with a single kind of feature as input, i.e., the spectral feature, which is recognized as the most discriminative feature in remote sensing image classification. Therefore, such an image classification approach processes each pixel independently using its own spectral feature, without considering the spatial relationship of the neighboring pixels. In fact, in remote sensing image classification, it is important to employ multiple features from both the spatial and spectral domains to effectively represent a pixel's information (Segl et al., 2003; Yang and Wang, 2012; Zhang et al., 2012, 2013). Such features include the spectral signature (Vaiphasa, 2006), the morphological property (Benediktsson et al., 2005), the shape feature (Jiao et al., 2012), and so on. Previous studies have reported that combining the multiple features of a certain pixel can improve land-cover classification accuracy (Landgrebe, 1980; Puissant et al., 2005). Since each feature can be viewed as a vector in a high-dimensional feature space, it is essential to consider the complementary properties of different features and combine them in order to obtain an accurate classification rate. A conventional approach is vector stacking (VS) (Huang et al., 2011), which simply concatenates different feature vectors into a long vector, then applies one of the aforementioned DR techniques before the subsequent classification. However, theoretically speaking, these DR technologies can only deal with a single kind of feature as input. In contrast, the direct VS strategy of multiple features intrinsically assumes that the different features are distributed in a unified feature space, although they are not, because they have different physical meanings and statistical properties (e.g., mean and variance). Therefore, it is unreasonable to use simple VS and DR to combine different features for the subsequent classification (Xia et al., 2010).

To overcome this problem, in this paper, we introduce a multiple features dimension reduction algorithm under a probability preserving projection framework, i.e., t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008). For each feature, a probability distribution is constructed based on t-SNE, and we then alternately solve t-SNE and learn the combination coefficients, i.e., the weighting factors for different features in the optimization. In summary, this modified stochastic neighbor embedding (MSNE): (1) considers multiple features of a pixel to achieve a physically meaningful low-dimensional feature representation for the subsequent classification; and (2) automatically optimizes the combination weighting factors for different features according to their contributions to the subsequent classification, which indicates the complementary properties of different features.

The remainder of this paper is organized as follows. In Section 2, we introduce the proposed multiple features dimension reduction strategy in detail. The experimental results are reported in Section 3, including the description of the study area and dataset, the spatial and spectral feature extraction of the remote sensing image, and the image classification results and analysis. Finally, Section 4 concludes the paper.

2. Modified stochastic neighbor embedding algorithm

The principle of the proposed multiple features dimension reduction strategy is shown in Fig. 1. The MSNE algorithm finds a low-dimensional representation $y \in R^d$ of input multiple features $\{f^{(k)} \in R^{L_k}\}_{k=1}^{m}$, in which m is the number of features and k is a specific feature within a population of m features ($k = 1, \ldots, m$), and L_k is the length of the kth feature vector. In order to deal with the out-of-sample problem (Bengio et al., 2004) (see Section 2.3 for a detailed discussion of this issue), only a subset of samples in the image are used as the input data of MSNE. Suppose we are given a multiple features dataset of n samples, e.g., $F = \{F^{(k)} \in R^{L_k \times n}\}_{k=1}^{m}$, wherein $F^{(k)}$ is the kth feature matrix. In MSNE, we first build a probability distribution $P^{(k)}$ for each feature based on t-SNE. We then alternately solve t-SNE and learn the optimal combination coefficient vector ω to obtain the solution of MSNE. Finally, the linear transformation for MSNE feature mapping is solved by linear regression, and the optimized feature representation in reduced feature space is achieved by such a linear transformation for each pixel of the remote sensing image, respectively.

2.1. t-distributed stochastic neighbor embedding

t-SNE is extended from standard SNE (Hinton and Roweis, 2003), which is designed for single feature nonlinear dimension reduction. Suppose that we have input high-dimensional data samples $X = \{x_1, \cdots, x_n\} \in R^{L \times n}$, in which n is the number of samples and L is the length of feature vector, respectively. SNE defines the normalized pairwise distances as a joint probability distribution over the input sample pairs, which are represented in a matrix P^s:

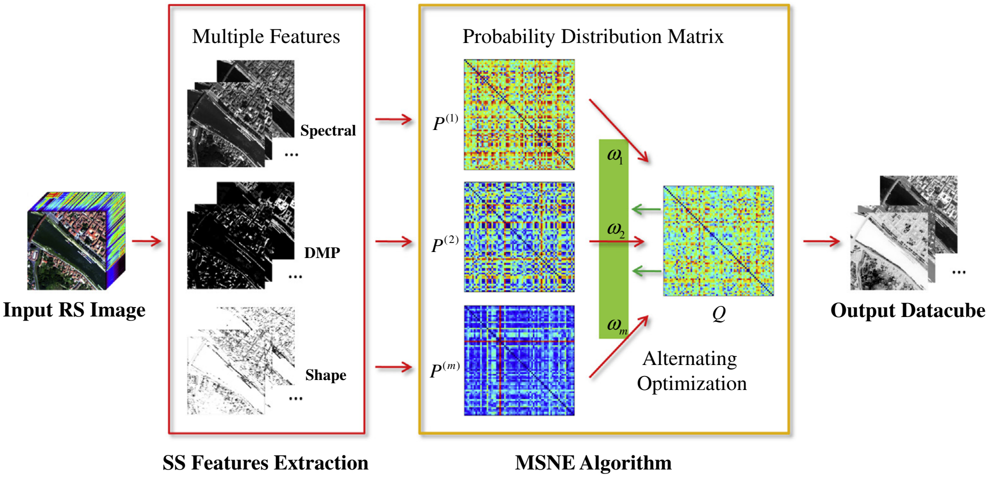


Fig. 1. Flowchart of the proposed multiple features dimension reduction strategy.

$$P^s_{ij} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{h\neq i}\exp(-\|x_i - x_h\|^2/2\sigma_i^2)} \in \mathrm{R}^{n\times n} \tag{1}$$

where $i, j, h \in [1, 2, \cdots, n]$ are index variables and σ_i^2 is the variance of the Gaussian distribution that is centered on data point x_i. Similarly, in the output low-dimensional feature space, suppose its feature dimensionality is d $(d < L)$, we define the probability distribution matrix Q^s of data samples $Y = \{y_1, \cdots, y_n\} \in \mathrm{R}^{d\times n}$ as follows (here we set the fixed variance to be 1/2 (Hinton and Roweis, 2003)):

$$Q^s_{ij} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{h\neq i}\exp(-\|y_i - y_h\|^2)} \in \mathrm{R}^{n\times n} \tag{2}$$

In Eqs. (1) and (2), we set $P^s_{ii} = 0$ and $Q^s_{ii} = 0$ since we are only interested in modeling the pairwise similarities.

The aim of SNE is to match these two distributions, P^s and Q^s, as well as possible, i.e., if all of the low-dimensional sample pairs (y_i, -y_j) exactly model the similarity between the high-dimensional sample pairs (x_i, x_j), the matrices P^s and Q^s will be equal. In SNE, this objective is achieved by minimizing the sum of the Kullback–Leibler divergences (Kullback and Leibler, 1951) between the two distributions over all the data points:

$$\min_Y KL(P^s, Q^s) = \min_Y \sum_i \sum_j P^s_{ij} \log \frac{P^s_{ij}}{Q^s_{ij}} \tag{3}$$

The above standard SNE is, however, always hampered by the optimization problem and a "crowding problem" (Maaten and Hinton, 2008); as a result, some variants of the SNE algorithm have been proposed (Cook et al., 2007; Yang et al., 2010). In this paper, we introduce t-SNE, which improves SNE in the following two aspects:

(a) In the high-dimensional feature space, a symmetric joint probability distribution P is defined, which leads to a simpler gradient computation in optimization:

$$P_{ij} = (P^s_{ij} + P^s_{ji})/2n \tag{4}$$

(b) In the low-dimensional feature space, a Student's t-distribution with one degree of freedom, rather than a Gaussian distribution, is used to compute the sample pairs' similarity Q, which can avoid the "crowding problem".

$$Q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{h\neq l}(1 + \|y_h - y_l\|^2)^{-1}} \tag{5}$$

where, again, we set P_{ii} and Q_{ii} to zero. Similar to Eq. (3), the objective function of t-SNE is given by:

$$\min_Y KL(P, Q) = \min_Y \sum_i \sum_j P_{ij} \log \frac{P_{ij}}{Q_{ij}} \tag{6}$$

Since Eq. (6) is not convex, gradient descent can be used to find a local solution. The gradient of the Kullback–Leibler divergence between P and Q in Eq. (6) is given by:

$$\partial KL(P, Q)/\partial y_i = 4\sum_j (P_{ij} - Q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1} \tag{7}$$

Based on the gradient (Eq. (7)), the basic gradient descent and some other improved strategies to optimize Eq. (6) are available in literature (Maaten and Hinton, 2008).

2.2. Multiple features t-SNE

In this subsection, we generalize t-SNE to handle the multiple features of samples and achieve a physically meaningful low-dimensional feature representation. According to t-SNE, for the kth feature matrix $F^{(k)} \in \mathrm{R}^{L_k\times n}$, we have its single feature based joint probability distribution $P^{(k)}$ using the definition in Eq. (4). When considering multiple features ($k = 1, \ldots, m$) simultaneously, we assume that the final probability distribution of the multiple features is a linear combination of all the single feature based joint probability distribution matrices, i.e.,

$$P = \sum_{k=1}^{m} \omega_k P^{(k)} \tag{8}$$

where ω_k is the nonnegative weight of each feature, with the strong constraints that $\omega_k > 0$ and $\sum_k \omega_k = 1$. It can be observed from Eq. (8) that the larger that ω_k is, the more important is the role of the kth feature in constructing the final probability distribution matrix P, which also indicates the contribution of this feature to the subsequent image classification. Thus, it is a key issue for MSNE to automatically optimize ω_k for each feature, according to its unique contribution. In this paper, we propose an alternating optimization to simultaneously optimize the MSNE objective function with respect to both the low-dimensional feature representation Y and the multiple features weight vector ω. By following the objective function of t-SNE (Eq. (6)), the final objective function of MSNE is given by:

$$\min_{Y,\omega} \sum_i \sum_j \left\{ \sum_{k=1}^{m} \omega_k P^{(k)}_{ij} \log \frac{\sum_{k=1}^{m}\omega_k P^{(k)}_{ij}}{Q_{ij}} \right\}, \quad \text{s.t.} \quad \omega_k > 0, \quad \sum_k \omega_k = 1 \tag{9}$$

in which matrix Q is derived from the output low-dimensional feature representation (Eq. (5)).

The optimization of the MSNE algorithm can be locally minimized as follows. To start the alternating optimization of Eq. (9), we set the initial value of the multiple features weight vector to $\omega_k = 1/m$, which means that each feature has the same weighting factor at initialization. Then, in every round of iteration, we first fix ω to optimize Y, then fix Y to optimize ω. The details of these two steps are given below:

(a) Fix ω to optimize Y. Since ω is fixed, we simply compute P using Eq. (8), and then the objective function (Eq. (9)) will be reduced to exactly the same as t-SNE:

$$\min_Y \sum_i \sum_j P_{ij} \log \frac{P_{ij}}{Q_{ij}} \tag{10}$$

The local minimum of Eq. (10) can be reached by the strategies introduced in Section 2.1.

(b) Fix Y to optimize ω. The objective function (Eq. (9)) reduces to:

$$\min_{\omega} \sum_i \sum_j \left\{ \sum_{k=1}^{m} \omega_k P^{(k)}_{ij} \log \frac{\sum_{k=1}^{m}\omega_k P^{(k)}_{ij}}{Q_{ij}} \right\}, \quad \text{s.t.} \quad \omega_k > 0, \quad \sum_k \omega_k = 1 \tag{11}$$

which is an entropy maximization (Boyd and Vandenberghe, 2004). As the solution of this optimization, the optimized vector ω must be the only one of ω_k equal to 1, and the others must be equal to zeros (at the vertex of the variable feasible region), which means that only one feature works, while the contributions of the other features vanish in the output low-dimensional feature representation (Xie et al., 2011). To avoid this problem, we add an l_2 norm regularization term into the current objective function:

$$\min_{\omega}\sum_i\sum_j\left\{\sum_{k=1}^{m}\omega_k P_{ij}^{(k)}\log\frac{\sum_{k=1}^{m}\omega_k P_{ij}^{(k)}}{Q_{ij}}\right\}+r\|\omega\|^2,\quad \text{s.t.}\quad \omega_k > 0,\quad \sum_k \omega_k = 1 \qquad (12)$$

It is worth noting that the regularization parameter r in (12) plays an important role in the optimization. This parameter actually controls the weighting of each feature for the reduced feature representation. The detailed analysis of the effect of this parameter to the algorithm performance would be discussed later.

It is known that the Kullback–Leibler divergence and the l_2 norm are both convex functions with respect to ω (Boyd and Vandenberghe, 2004); therefore, the minimization (12) is convex with respect to ω. In fact, this optimization can be globally minimized by the use of Nesterov's accelerated first-order method (Nesterov, 2005).

2.3. Linearization of MSNE

The introduced MSNE finds an optimal feature embedding for the original multiple features in the high-dimensional feature spaces. It should be emphasized that this feature mapping from $F=\{F^{(k)}\in \mathrm{R}^{L_k\times n}\}_{k=1}^{m}$ to $Y=\{y_1,\cdots,y_n\}\in \mathrm{R}^{d\times n}$ is always nonlinear and implicit. In fact, we usually have to process hundreds of thousands of pixels in remote sensing image classification, i.e., $n = 10^5$ in image classification. However, it is not possible to use MSNE to find the low-dimensional subspace using all the pixels as input, because the size of the joint probability distribution matrices $P^{(k)}$ scales with the number of input samples; therefore, the suggested MSNE suffers from the out-of-sample problem. To address this problem, just as with some linear versions of the manifold learning DR algorithms (He et al., 2005), only a subset of pixels in the image is used as the input data of MSNE. These samples can be generated by uniform selection or random selection from the full dataset. An explicit linear projection matrix learned by MSNE is then applied to approximately construct the low-dimensional representation. This linear transformation for MSNE feature mapping is solved by linear regression:

$$U^T = YF^T(FF^T)^{-1} \qquad (13)$$

Finally, for each pixel in the RS image with multiple features $f=\{f^{(k)}\in \mathrm{R}^{L_k}\}_{k=1}^{m}\in \mathrm{R}^{\sum_m L_k}$, the corresponding low-dimensional feature representation can be computed by:

$$y = U^T f \qquad (14)$$

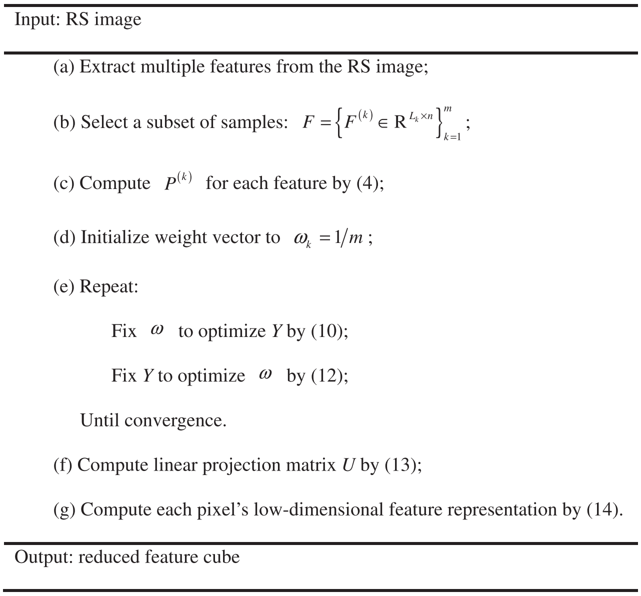
As a summary of this section, the detailed procedure of MSNE for RS image multiple features DR is shown in Fig. 2.

3. Experiments and analysis

In this section, we provide the experimental results and analysis from a hyperspectral remote sensing image acquired by the Reflective Optics System Imaging Spectrometer (ROSIS). Firstly, we give a brief introduction into the adopted study dataset. We then investigate the extracted multiple features of the remote sensing image and show the image classification results using the proposed multiple features DR algorithm, compared to some other DR technologies. Finally, we discuss the parameter analysis of MSNE.

3.1. Study area and dataset

In this section, the experiments are conducted on a publicly available airborne hyperspectral remote sensing dataset. The studied RS image was acquired by the ROSIS-03 (Gege and Mooshuber, 1997) optical sensor on July 8, 2002, at the urban test area of Pavia, northern Italy (45.11 N, 9.09E). The whole dataset size is 1400 × 512 pixels, and we use a 400 × 400 subset in this study. The spatial resolution of this RS image is 1.3 m per pixel. The number of bands in the acquired image is 115, with a spectral coverage ranging from 0.43 to 0.86 μm, while 13 noisy bands have been removed by the dataset provider; therefore, the spectral dimension of this image is 102. This dataset was provided by the IEEE GRSS Data Fusion Technical Committee (Licciardi et al., 2009).



```
Input: RS image
(a) Extract multiple features from the RS image;
(b) Select a subset of samples: $F=\left\{F^{(k)}\in \mathrm{R}^{L_k\times n}\right\}_{k=1}^{m}$;
(c) Compute $P^{(k)}$ for each feature by (4);
(d) Initialize weight vector to $\omega_k = 1/m$;
(e) Repeat:
        Fix $\omega$ to optimize $Y$ by (10);
        Fix $Y$ to optimize $\omega$ by (12);
    Until convergence.
(f) Compute linear projection matrix $U$ by (13);
(g) Compute each pixel's low-dimensional feature representation by (14).
Output: reduced feature cube
```

Fig. 2. Procedure of MSNE for RS image multiple features DR.

3.2. Multiple features of the RS image

In our experiments, three kinds of features are employed as a case study, i.e., the spectral feature, the morphological feature, and the shape feature.

(1) The spectral feature**:** the spectral feature of a pixel in a remote sensing image is obtained by arranging its observed surface reflectance in all of the l bands:

$$\text{Spectral} = [v_1, v_2, \cdots, v_l]^T \qquad (15)$$

in which v_i denotes the DN in band i.

(2) The morphological feature: the differential morphological profile (DMP) (Benediktsson et al., 2005), which can record image structural information, is based on two commonly used morphological operators, i.e., opening and closing.

Let γ_s and ϕ_s be the morphological opening and closing operators by reconstruction with scale element $s \in [0, S]$, and $\Pi\gamma_s$ and $\Pi\phi_s$ are the opening and closing profiles of image I with a single scale s. Therefore, the multiscale opening and closing profiles of image I are defined as vectors:

$$\Pi\gamma = \{\Pi\gamma_s : \Pi\gamma_s = \gamma_s(I),\quad s\in[0,S]\} \qquad (16)$$

$$\Pi\varphi = \{\Pi\varphi_s : \Pi\varphi_s = \varphi_s(I),\quad s\in[0,S]\} \qquad (17)$$

in Eqs. (16) and (17), we define $\Pi\gamma_0 = \Pi\phi_0 = I$ for scale $s = 0$. The DMP is then defined as a vector where the measure of the slope of the opening-closing profile is stored for every step of an increasing scale series (Huang and Zhang, 2009):

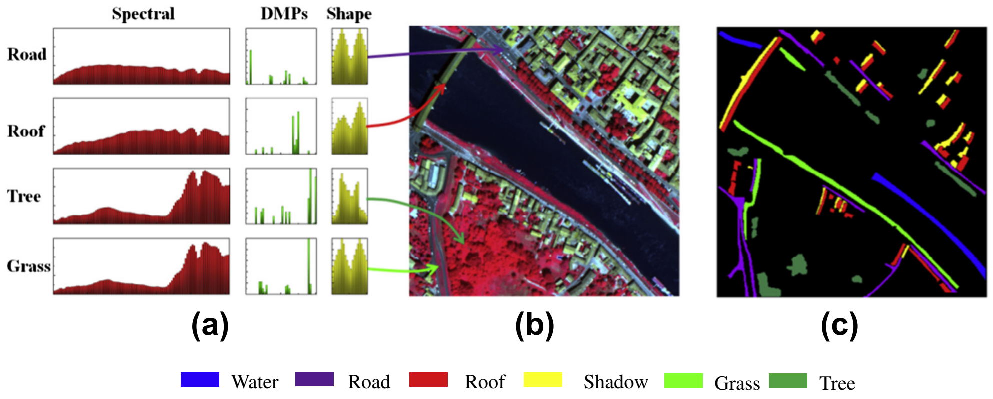


Fig. 3. (a) Multiple features of four pixels with different land-cover classes in the ROSIS dataset, (b) false color image, and (c) the reference ground truth.

$$\mathrm{DMP}_{\Omega} = \{\mathrm{DMP}_{\Omega_s} : \mathrm{DMP}_{\Omega_s} = |\Pi_{\Omega_s} - \Pi_{\Omega_{s-1}}|, \quad \Omega \in [\gamma, \varphi], \quad s \in [1, S]\} \tag{18}$$

It is worth noting that the above DMP definition is proposed for a gray-level image, i.e., image I in Eqs. (16) and (17). However, for hyperspectral images (HSI), the hundreds of spectral bands always lead to a high-dimensional feature space. As a result, according to the previous literature, DMP should be implemented on several principal component images for HSI morphological feature extraction.

(3) The shape feature: the pixel shape index (PSI) (Zhang et al., 2006) based method is adopted to describe the shape feature in a local area. For a certain pixel in an image, the PSI shape feature extraction consists of three steps: (i) extension of the direction lines based on gray-level similarity, i.e., pixel homogeneity PH_i; (ii) measurement of the length of each direction line d_i, based on the direction line; and (iii), finally, the shape feature can be represented as:

$$\mathrm{Shape} = [d_1, d_2, \cdots, d_p]^{\mathrm{T}} \tag{19}$$

in which d_i is the length of the ith direction line, and p is the number of directions.

In the experiments, the aforementioned multiple features are extracted by the following detailed parameter settings: in the spectral domain, we use the 102-D spectral feature for each pixel. In the spatial domain, we extract the DMP feature using the top four principal component images of HSI, with scale elements of s = 0, 1, 2, 3, and 4 for the opening and closing profiles, respectively, which results in the 40-D DMP feature. For the shape feature, by setting the number of directions to 20, we obtain the 20-D shape feature. In order to compare these features more clearly, each feature is linearly stretched to a range of [0, 1], according to its statistical maximal and minimal values.

Fig. 3a shows the spectral, DMP and shape features for different pixels in the ROSIS image. The pixels correspond to the varieties of land-cover classes, e.g., road, roof, grass, and tree, respectively. Usually, the spectral signature is the most discriminative feature in remote sensing image classification, especially in hyperspectral image classification. However, as this RS image also has a high spatial resolution, which can provide a large amount of detailed spatial information, due to the complex spectral attributes within each land-cover class and between different classes, single spectral feature based image classification suffers from an increasing of the intra-class variance and a decreasing of the inter-class variance. This leads to a decrease in the discriminability of features in the spectral domain, particularly for the spectrally similar classes. This phenomenon can be observed in Fig. 3a, in that the pixel pairs (road, roof) and (grass, tree) have very similar spectral signatures, which inspired us to integrate the spatial features (i.e., DMP and shape), as well as the spectral feature, to enhance the discriminability between the pixels of different land-cover classes.

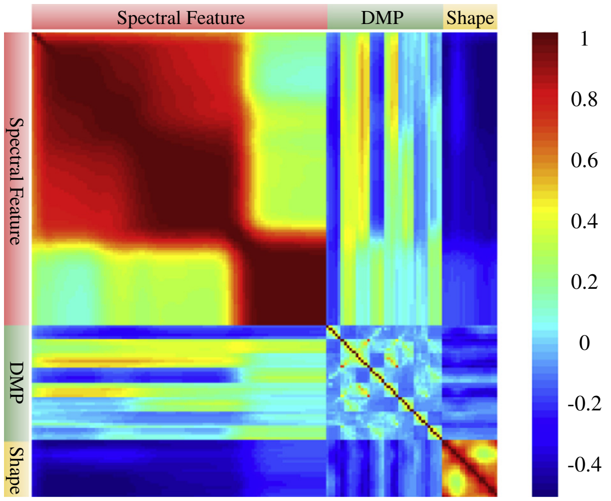


Fig. 4. Correlation coefficient matrix of multiple features in the ROSIS dataset.

Fig. 4 shows the correlation coefficient matrix of the multiple features in the ROSIS dataset. The size of this matrix is 162 × 162, in which 1–102, 103–142, and 143–162 are the spectral, DMP, and shape features, respectively. It is obvious that the correlation matrix shown in Fig. 4 contains two red blocks in the spectral domain, which means that the corresponding features are highly correlated. The same phenomenon can also be observed in the shape domain. However, the correlations between these multiple features show much lower values (close to zero in many places, as shown in Fig. 4), which indicates that there are some complementary properties of the above spectral and spatial features. This can also be validated in Fig. 3a, in that although the pixel pair (road, roof) have a very similar spectral signature, we can still distinguish them according to the DMP and shape features. These complementary properties of the multiple features in a RS image provide information that could potentially improve the image classification accuracy.

3.3. Classification result of the ROSIS dataset

For the MSNE algorithm, we uniformly select n = 1200 samples (about 0.75% of all the pixels in this RS image), as per step (b) of the

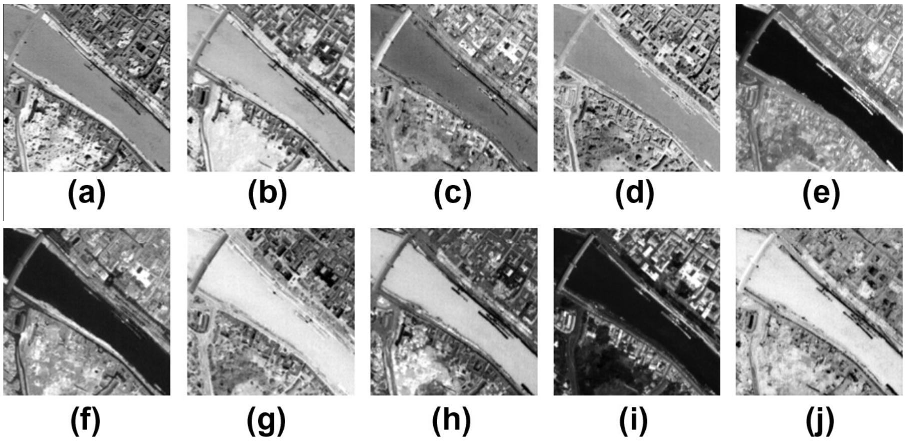


Fig. 5. (a)–(j) Top ten reduced features of all the pixels in the ROSIS dataset, as obtained by MSNE.

whole procedure. Fig. 5a–j shows the top ten reduced feature components of all the pixels in the ROSIS dataset obtained by MSNE. In particular, we highlight the output features of six pixels with different land-cover classes, as shown in Fig. 6a–f. The subsequent image classification is then performed using the extracted feature cube.

This subsection gives the image classification results of the proposed DR approach, as well as some other DR technologies, including PCA (Jolliffe, 2002), neighborhood preserving embedding (NPE) (He et al., 2005), locality preserving projections (LPP) (He and Niyogi, 2004), nonparametric weighted feature extraction (NWFE) (Kuo and Landgrebe, 2004), and *t*-distributed stochastic neighbor embedding (*t*-SNE) (Maaten and Hinton, 2008). Furthermore, two classification results are also considered as baselines, they being the spectral feature (SF, 102-D) based and the vector stacking feature (VS, 162-D) based image classification. Apart from MSNE, all the other DR algorithms directly adopt the VS 162-D feature as input and then turn out to be a low-dimensional feature representation for the subsequent image classification. We first employ the *k*-NN classifier (Cover and Hart, 1967) with the setting of $k = 1$ to achieve the supervised image classification, with the training samples being randomly selected from the reference data shown in Fig. 3c. The remaining reference data are applied as test samples for accuracy assessments of the resulting classification maps. The numbers of all the reference data as well as the training and test samples are listed in Table 1. Note that NWFE is a supervised DR algorithm, which needs discriminative information from training samples in the objective function; therefore, we use the training samples, which were the same as in the classification step, to guarantee the fairness of the comparison.

Eight different feature-based classification maps are compared in Fig. 7a–h. Among them, for the six DR approaches, the size of the reduced low-dimensional feature space is fixed at 30. In Fig. 7, the proposed MSNE-based image classification achieves

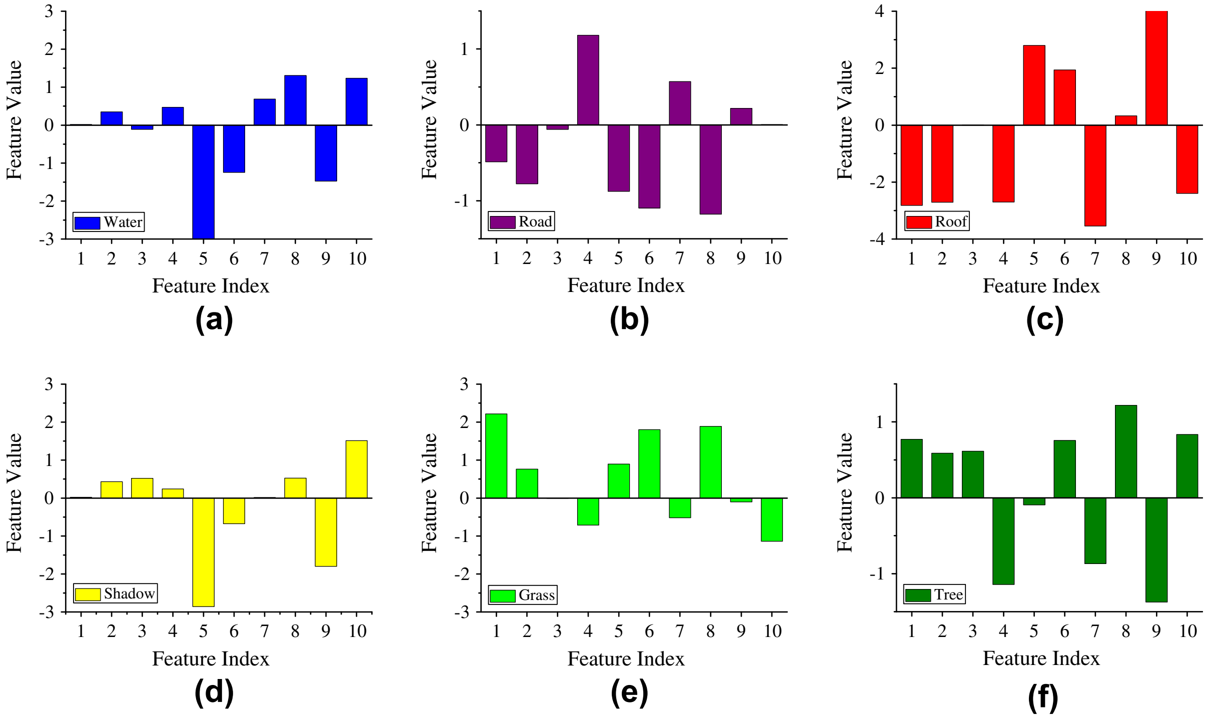


Fig. 6. Graphical representation of the reduced features for different classes. (a) Water, (b) road, (c) roof, (d) shadow, (e) grass, and (f) tree.

Table 1
Numbers of all the reference data and training and test samples for the ROSIS image.

	Water	Road	Roof	Shadow	Grass	Tree	All
Reference	2224	3696	4187	2662	3400	3216	19,385
Training	30	30	30	30	30	30	180
Test	2194	3666	4157	2632	3370	3186	19,205

Table 2
Classification accuracies of the various features for the ROSIS image.

	Water	Road	Roof	Shadow	Grass	Tree	OA	Kappa
SF	99.95	87.01	91.94	92.02	83.38	88.32	89.83	0.8769
VS	98.08	70.32	80.37	89.36	82.10	83.67	82.56	0.7890
PCA	98.08	70.26	80.29	89.39	82.10	83.80	82.56	0.7890
NPE	100	84.39	88.21	91.26	94.36	93.78	91.25	0.8945
LPP	99.95	83.90	85.30	90.19	90.32	90.55	89.13	0.8691
NWFE	99.90	94.10	89.43	91.41	88.93	93.15	92.32	0.9071
t-SNE	99.86	90.09	86.62	90.95	90.77	93.18	91.21	0.8938
MSNE	99.95	94.68	95.83	92.47	92.99	92.31	94.54	0.9340

the best performance. By a detailed comparison with the other classification maps in Fig. 7a–g, the proposed MSNE shows a good classification result, especially at the following places in the image: (1) the roof and its shadow across the river, (2) the large number of roof pixels in the north-east of the image, and (3) the continuous road pixels in the south-west of the image. In order to thoroughly evaluate the discriminability of the different features, the class-specific accuracies and overall accuracies (OA), as well as the kappa coefficients of Fig. 7a–h, are also reported in Table 2. From this table, the improvements can be highlighted in that MSNE obtains the highest classification rate in all three classes (road, roof, and shadow), and achieves the top OA and kappa coefficient values.

For a more detailed and comprehensive comparison of the dimension reduction algorithms (PCA, NPE, LPP, NWFE, *t*-SNE, and MSNE), the feature DR and classification operations are conducted using these algorithms combined with two different classifiers, i.e., *k*-NN and maximum likelihood, with an increase in subspace dimension *d*. Fig. 8a shows the classification OAs of the different algorithms using the *k*-NN classifier and, again, we use the SF and VS classification results as baselines. As shown in Fig. 8a, MSNE performs better than the other algorithms when $d > 10$ and achieves the best classification rate around $d = 20$. When the subspace dimension is increased to a larger value, the classification OA of MSNE stabilizes at the highest OA. Similarly, Fig. 8b plots the classification OAs of the different algorithms using the maximum likelihood classifier. Note that for such a classifier, the covariance matrix of any class will become singular if the amount of ground truth data in this class is less than the input feature dimensionality; as a result of this, in our experiment, the maximal feature dimension should be set to less than 30 (the number of training samples for each class is given in Table 1). The curves in Fig. 8b show the significant effect of the Hughes phenomenon: when the subspace dimension is greater than 20, the classification performance of all the DR algorithms falls rapidly. However, when the subspace dimension increases from $d = 5$ to $d = 15$, the MSNE algorithm achieves the highest image classification OA among all the reference DR technologies.

3.4. Parameter analysis of MSNE

In the optimization of the weighting factor (Eq. (11)), an l_2 norm regularization is introduced as a relaxation with parameter *r* to avoid the unexpected solution given by Eq. (10). According to the theoretical analysis in Section 2.2, this regularization actually ensures each feature has a unique weight for the reduced feature representation, adapted to its contribution to the image classification. Here, we investigate the effect of this parameter *r* in the alternating optimization step. Fig. 9a–h describe the relationship of the regularization parameter *r* and the weighting factors of the extracted multiple features. From these figures we can see that the spectral feature is the most discriminative feature for the image classification, because the weighting factor of the spectral feature is the largest in all the figures. It can also be observed that if *r* is close to 0, as shown in Fig. 9a, the weighting factors are very sparse; therefore, the most discriminative feature will be assigned a large coefficient, and vice versa. In particular, when $r = 0$, the weighting factor of the spectral feature will be assigned to 1, while the weighting factors of the other features will be zero. This point can also be theoretically justified by convex optimization, as mentioned before. With the increase in the *r* value, the weighting

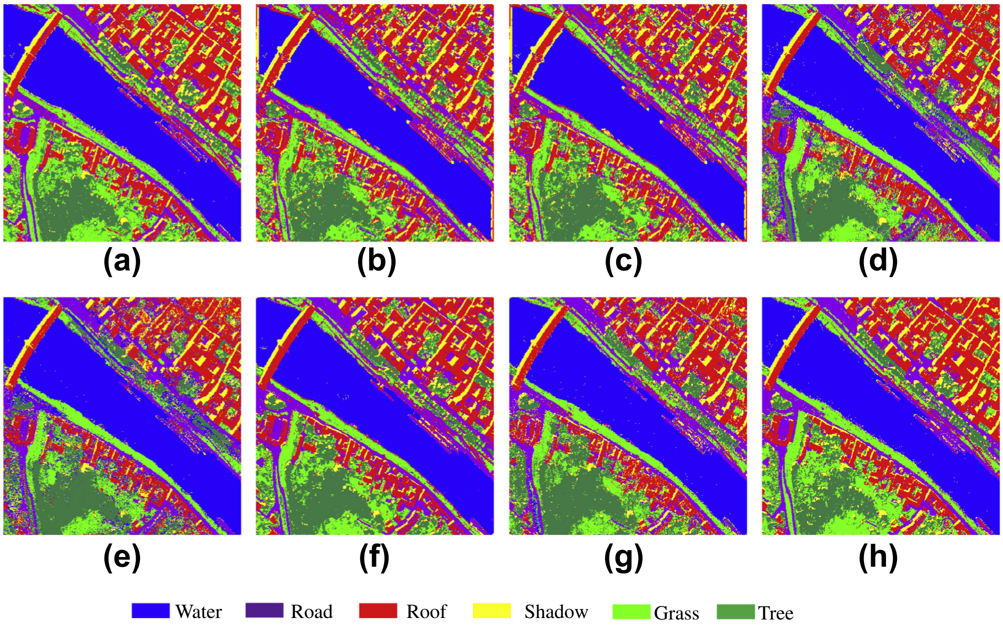


Fig. 7. Classification maps of all the methods for the ROSIS image. (a) SF, (b) VS, (c) PCA, (d) NPE, (e) LPP, (f) NWFE, (g) *t*-SNE and (h) MSNE.

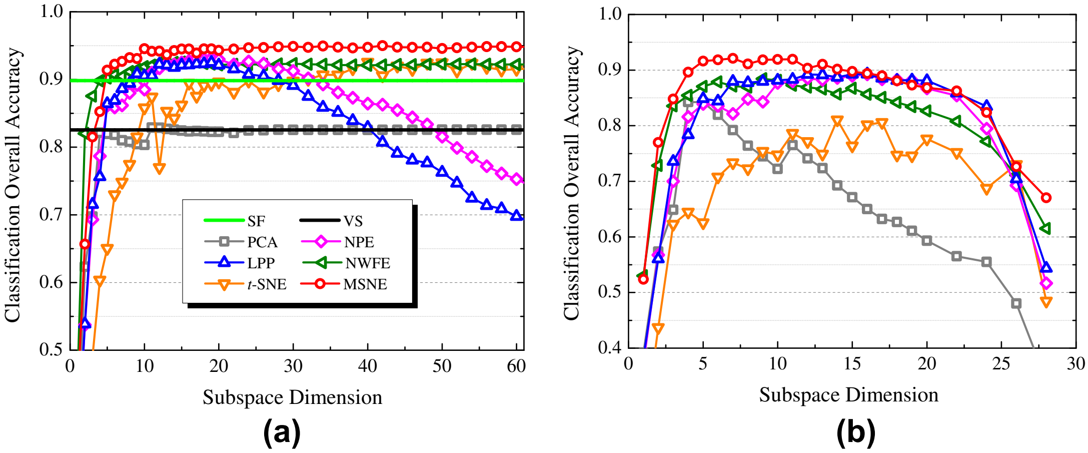


Fig. 8. Classification OAs with respect to reduced feature dimensionality, with the ROSIS dataset. (a) *k*-NN classifier and (b) maximum likelihood classifier.

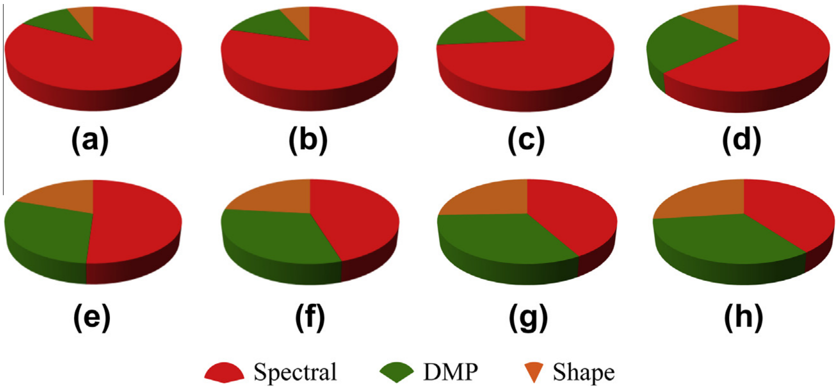


Fig. 9. The effect of parameter r on the weighting factors of each feature: (a) $r = 0.25$, (b) $r = 0.5$, (c) $r = 1$, (d) $r = 2$, (e) $r = 4$, (f) $r = 6$, (g) $r = 8$, and (h) $r = 10$.

factors become close to each other, which indicates that the multiple features will share similar weights in the low-dimensional feature representation. On the other hand, if r is increased to infinity, the weighting factors of the multiple features will be equal.

Fig. 10 gives a specific comparison of the MSNE classification OAs with various values of parameter r, using the k-NN classifier. In this comparison, we use the best performance of the t-SNE algorithm in Fig. 8a ($d = 60$ for t-SNE) as the baseline. It is clear that the proposed MSNE outperforms t-SNE over all the r range, with the peak at $r = 6$. Before r reaches this location, the classification OA increases with respect to this parameter; when r is larger than this value, the classification OA decreases slowly and shows a stable tendency. By a comprehensive analysis of the phenomenon appearing in Figs. 9 and 10, we can conclude that the selection of the regularization parameter r should be based on the complementary properties of the input multiple features. If the available features are complementary to each other, a larger r is preferred to guarantee that all the input features properly contribute to the low-dimensional feature representation for the subsequent image classification; otherwise, we should choose a small r.

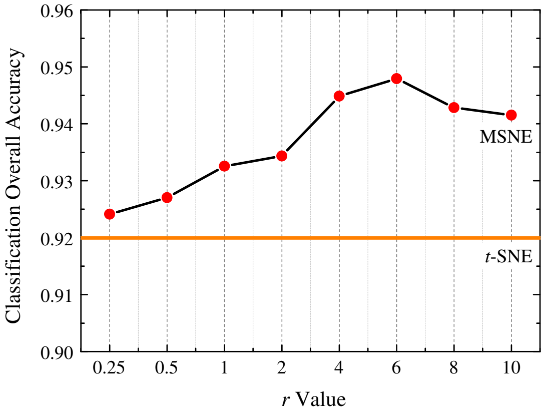


Fig. 10. The effect of parameter r on the MSNE classification OA.

To validate the convergence rate of the adopted alternating optimization in the ROSIS dataset, Fig. 11a–h shows the weighting factors of the multiple features at the end of each iteration in the alternating optimization procedure. The initial weighting factors are set to $\omega_k = 1/3$, ($k = 1, 2, 3$), as declared in Fig. 2. It is clear that the weighting factors often converge at a stable value in about five iterations; therefore, the low-dimensional feature representation Y should also reach convergence, correspondingly. In numbers of experiments on other RS images, the same trend in convergence has also been clearly observed. These experimental results suggest that we should fix the number of iterations to five to guarantee that the MSNE algorithm converges.

4. Conclusion

In this paper, we introduce a multiple features dimension reduction algorithm under a probabilistic framework which can consider the features from both the spectral and spatial domains

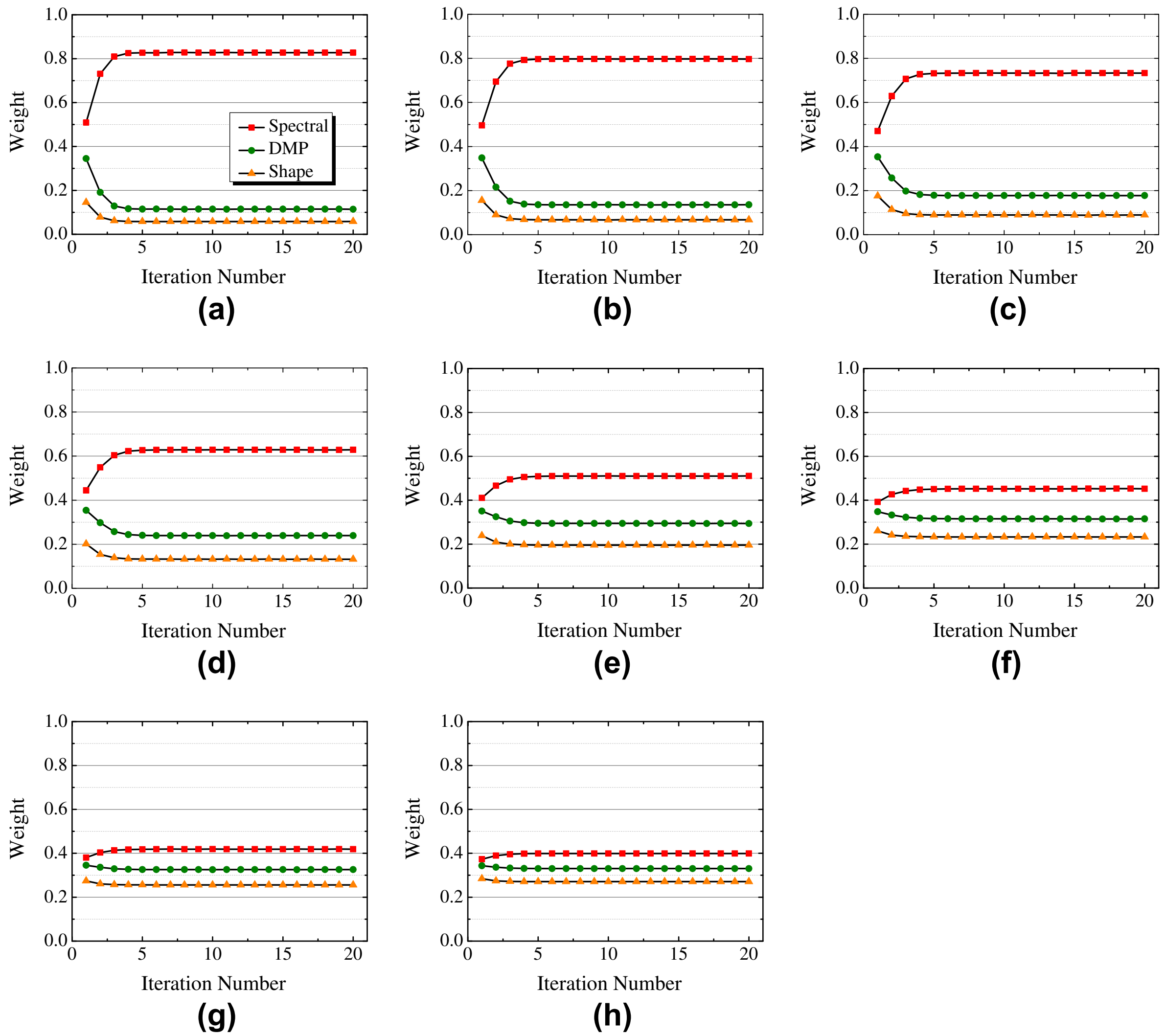


Fig. 11. Convergence of the alternating optimization: (a) $r = 0.25$, (b) $r = 0.5$, (c) $r = 1$, (d) $r = 2$, (e) $r = 4$, (f) $r = 6$, (g) $r = 8$, and (h) $r = 10$.

of a pixel to achieve a physically meaningful low-dimensional feature representation for an effective and accurate remote sensing image classification. For each input feature, a probability distribution is constructed based on *t*-SNE, and we then alternately solve *t*-SNE and learn the optimal weighting coefficients for different features in the MSNE optimization. The linear transformation for MSNE feature mapping is achieved by linear regression in order to deal with the out-of-sample problem in remote sensing image classification. Experiments on the classification of ROSIS hyperspectral remote sensing image demonstrate that the proposed approach can explore the complementary properties of different features and find an optimal low-dimensional feature representation for the subsequent classification. The effect of the weighting factors of each feature on the image classification OA is also investigated. Our future work will explore how to automatically select the optimal regularization parameter *r* in the MSNE alternating optimization, according to an analysis of the complementary properties of the input multiple features, which will help the proposed procedure to achieve a better classification rate for remote sensing images.

Acknowledgments

The authors would like to thank the handling editor Dr. Dar Roberts, and anonymous reviewers for their support and constructive comments on this paper. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2011CB707105, by the National Natural Science Foundation of China under Grants 40930532, 41101336, and by the Program for New Century Excellent Talents in University of China under Grant NCET-11-0396.

References

Acqua, F.D., Gamba, P., Ferrari, A., Palmason, J.A., Benediktsson, J.A., Arnason, K., 2004. Exploiting spectral and spatial information in hyperspectral urban data with high resolution. IEEE Geoscience and Remote Sensing Letters 1 (4), 322–326.

Bachmann, C.M., Ainsworth, T.L., Fusina, R.A., 2005. Exploiting manifold geometry in hyperspectral imagery. IEEE Transactions on Geoscience and Remote Sensing 43 (3), 441–454.

Baraldi, A., Parmiggiani, F., 1995. A Neural network for unsupervised categorization of multivalued input patterns: an application to satellite image clustering. IEEE Transactions on Geoscience and Remote Sensing 33 (2), 305–316.

Benediktsson, J.A., Palmason, J.A., 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., Roux, N.L., Ouimet, M., 2004. Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering. In: Advances in Neural Information Processing Systems, Vancouver, Canada, 13–18 December, pp. 177–184.

Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge Univ. Press, Cambridge, UK.

Campbell, J.B., 2000. Introduction to Remote Sensing. London, UK, Taylor & Francis.

Chang, C.-I., 2003. Hyperspectral Imaging: Techniques for Spectral Detection and Classification. Kluwer Academic/Plenum Publishers, New York.

Conese, C., Maselli, F., 1993. Selection of optimum bands from tm scenes through mutual information analysis. ISPRS Journal of Photogrammetry and Remote Sensing 48 (3), 2–11.

Cook, J., Sutskever, I., Mnih, A., Hinton, G., 2007. Visualizing similarity data with a mixture of maps. In: International Conference on Artificial Intelligence and Statistics, San Juan, Puerto Rico, 21–24 March, pp. 67–74.

Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13 (1), 21–27.

Gege, P., Mooshuber, W., 1997. Electronic performance of the imaging spectrometer ROSIS-03. In: Proceedings of the Workshop of ISPRS, Working Groups I/1, I/3 and IV/4: Sensors and Mapping From Space, Hannover, Germany, October, pp. 49–67.

Harsanyi, J.C., Chang, C.-I., 1994. Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach. IEEE Transactions on Geoscience and Remote Sensing 32 (4), 779–785.

He, X., Niyogi, P., 2004. Locality preserving projections. In: Advances in Neural Information Processing Systems, Vancouver, Canada, 13–18 December, pp. 153–160.

He, X., Cai, D., Yan, S., Zhang, H.-J., 2005. Neighborhood preserving embedding. In: IEEE International Conference on Computer Vision, Beijing, China, 17–20 October, pp. 1208–1213.

Hinton, G., Roweis, S., 2003. Stochastic neighbor embedding. In: Advances in Neural Information Processing Systems, Vancouver, British Columbia, Canada, 8–13 December, pp. 857–864.

Huang, X., Zhang, L., 2009. A Comparative study of spatial approaches for urban mapping using hyperspectral ROSIS images over Pavia city, Northern Italy. International Journal of Remote Sensing 30 (12), 3205–3221.

Huang, X., Zhang, L., Gong, W., 2011. Information fusion of aerial images and LIDAR data in urban areas: vector-stacking, re-classification and post-processing approaches. International Journal of Remote Sensing 32 (1), 69–84.

Hughes, G.F., 1968. On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory 14 (1), 55–63.

Jiao, L., Liu, Y., Li, H., 2012. Characterizing land-use classes in remote sensing imagery by shape metrics. ISPRS Journal of Photogrammetry and Remote Sensing 72, 46–55.

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

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. The Annals of Mathematical Statistics 22 (1), 79–86.

Kuo, B.-C., Landgrebe, D.A., 2004. Nonparametric weighted feature extraction for classification. IEEE Transactions on Geoscience and Remote Sensing 42 (5), 1096–1105.

Landgrebe, D.A., 1980. The development of a spectral-spatial classifier for earth observational data. Pattern Recognition 12 (3), 165–175.

Landgrebe, D., 2002. Hyperspectral image data analysis. IEEE Signal Processing Magazine 19 (1), 17–28.

Li, W., Prasad, S., Fowler, J.E., Bruce, L.M., 2012. Locality-preserving dimensionality reduction and classification for hyperspectral image analysis. IEEE Transactions on Geoscience and Remote Sensing 50 (4), 1185–1198.

Licciardi, G., Pacifici, F., Tuia, D., Prasad, S., West, T., Giacco, F., Thiel, C., Inglada, J., Christophe, E., Chanussot, J., Gamba, P., 2009. Decision fusion for the classification of hyperspectral data: outcome of the 2008 GRS-S data fusion contest. IEEE Transactions on Geoscience and Remote Sensing 47 (11), 3857–3865.

Liu, K., Shi, W., Zhang, H., 2011. A fuzzy topology-based maximum likelihood classification. ISPRS Journal of Photogrammetry and Remote Sensing 66 (1), 103–114.

Lunga, D., Ersoy, O., 2013. Spherical stochastic neighbor embedding of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing 51 (2), 857–871.

Ma, L., Crawford, M.M., Tian, J., 2010. Generalised supervised local tangent space alignment for hyperspectral image classification. Electronics Letters 46 (7), 497–498.

Maaten, L.v.d., Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9 (November), 2579–2605.

McLachlan, G.J., 1992. Discriminant Analysis and Statistical Pattern Recognition. Wiley-Interscience, New York.

Nesterov, Y., 2005. Smooth minimization of non-smooth functions. Mathematical Programming 103 (1), 127–152.

Phillips, R.D., Watson, L.T., Wynne, R.H., Blinn, C.E., 2009. Feature reduction using a singular value decomposition for the iterative guided spectral class rejection hybrid classifier. ISPRS Journal of Photogrammetry and Remote Sensing 64 (1), 107–116.

Puissant, A., Hirscha, J., 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.

Segl, K., Roessner, S., Heiden, U., 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 (1–2), 99–112.

Shao, Y., Lunetta, R.S., 2012. Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. ISPRS Journal of Photogrammetry and Remote Sensing 70, 78–87.

Stavrakoudis, D.G., Theocharis, J.B., Zalidis, G.C., 2011. A Boosted genetic fuzzy classifier for land cover classification of remote sensing imagery. ISPRS Journal of Photogrammetry and Remote Sensing 66 (4), 529–544.

Vaiphasa, C., 2006. Consideration of smoothing techniques for hyperspectral remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 60 (2), 91–99.

Walter, V., Luo, F., 2011. Automatic interpretation of digital maps. ISPRS Journal of Photogrammetry and Remote Sensing 66 (4), 519–528.

Xia, T., Tao, D., Mei, T., Zhang, Y., 2010. Multiview spectral embedding. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 40 (6), 1438–1446.

Xie, B., Mu, Y., Tao, D., Huang, K., 2011. M-SNE: Multiview stochastic neighbor embedding. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 41 (4), 1088–1096.

Yang, J., Wang, Y., 2012. Classification of 10m-resolution SPOT data using a combined bayesian network classifier-shape adaptive neighborhood method. ISPRS Journal of Photogrammetry and Remote Sensing 72, 36–45.

Yang, Z., King, I., Xu, Z., Oja, E., 2010. Heavy-tailed symmetric stochastic neighbor embedding. In: Advances in Neural Information Processing Systems, Vancouver, British Columbia, Canada, 6–9 December, pp. 2169–2177.

Zhang, L., Huang, X., Huang, B., 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, T., Tao, D., Li, X., Yang, J., 2009. Patch alignment for dimensionality reduction. IEEE Transactions on Knowledge and Data Engineering 21 (9), 1299–1313.

Zhang, L., Zhang, L., Tao, D., Huang, X., 2012. On combining multiple features for hyperspectral remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing 50 (3), 879–893.

Zhang, L., Zhang, L., Tao, D., Huang, X., 2013. Tensor discriminative locality alignment for hyperspectral image spectral-spatial feature extraction. IEEE Transactions on Geoscience and Remote Sensing 51 (1), 242–256.

Zhao, G., Maclea, A.L., 2000. A comparison of canonical discriminant analysis and principal component analysis for spectral transformation. Photogrammetric Engineering and Remote Sensing 66 (7), 841–847.

Zhong, Y., Zhang, L., 2012. An adaptive artificial immune network for supervised classification of multi-/hyperspectral remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing 50 (3), 894–909.

Zhu, G., Blumberg, D.G., 2002. Classification using ASTER data and svm algorithms: the case study of Beer Sheva, Israel. Remote Sensing of Environment 80 (2), 233–240.