Article

Hyperspectral Remote Sensing Image Feature Representation Method Based on CAE-H with Nuclear Norm Constraint

Xiaodong Yu ^{1,*}, Rui Ding ², Jingbo Shao ¹ and Xiaohui Li ³

¹ School of Computer Science and Information Engineering, Harbin Normal University, Harbin 150025, China; zro_bol@163.com

² College of Computer Science and Technology, Mudanjiang Normal University, Mudanjiang 157011, China; mdjdingrui@163.com

³ Department of Computer Science, Harbin Vocational and Technical College, Harbin 150081, China; lxhhrb@hrbeu.edu.cn

* Correspondence: yuxiaodong@hrbnu.edu.cn

Abstract: Due to the high dimensionality and high data redundancy of hyperspectral remote sensing images, it is difficult to maintain the nonlinear structural relationship in the dimensionality reduction representation of hyperspectral data. In this paper, a feature representation method based on high order contractive auto-encoder with nuclear norm constraint (CAE-HNC) is proposed. By introducing Jacobian matrix in the CAE of the nuclear norm constraint, the nuclear norm has better sparsity than the Frobenius norm and can better describe the local low dimension of the data manifold. At the same time, a second-order penalty term is added, which is the Frobenius norm of the Hessian matrix expressed in the hidden layer of the input, encouraging a smoother low-dimensional manifold geometry of the data. The experiment of hyperspectral remote sensing image shows that CAE-HNC proposed in this paper is a compact and robust feature representation method, which provides effective help for the ground object classification and target recognition of hyperspectral remote sensing image.

Keywords: hyperspectral remote sensing images; feature representation; nuclear norm; contractive auto-encoder

Citation: Yu, X.; Ding, R.; Shao, J.; Li, X. Hyperspectral Remote Sensing Image Feature Representation Method Based on CAE-H with Nuclear Norm Constraint. *Electronics* **2021**, *10*, 2667. <https://doi.org/10.3390/electronics10212667>

Academic Editor: Chiman Kwan

Received: 1 October 2021

Accepted: 28 October 2021

Published: 31 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Hyperspectral remote sensing images are rich in spatial, spectral and radiation information, including hundreds or even thousands of spectral bands, which can fully reflect the subtle features of the surface object spectrum and provide extremely rich information for the extraction of surface object information, which is beneficial to more detailed surface object classification [1–4]. In recent years, hyperspectral remote sensing images have attracted the attention of many scholars, and have been widely used in ecological monitoring [5,6], medical diagnosis [7,8], military reconnaissance [9] and other important fields. Due to the increase of spectral bands in hyperspectral remote sensing images, the problems of increased dimension and high data redundancy appear [10,11], resulting in the complexity of data processing [12]. In order to alleviate the above problems, it is usually necessary to reduce the dimension of hyperspectral data [13,14]. Hyperspectral data has a specific nonlinear structure in the high dimensional space, and this nonlinear structure is also the area where hyperspectral data are distributed and concentrated in high density [10,15]. The dimensional reduction representation of hyperspectral data can accurately describe the effective information in the data and keep the important information in the data only by maintaining the nonlinear structural relationship in the data [16,17]. Therefore, it is necessary to study the feature representation method that can keep the nonlinear structure relation in hyperspectral data.

Manifold learning algorithm can solve the problem of feature dimensionality reduction in hyperspectral remote sensing image. In many manifold learning algorithms, the local

structure of manifold is represented by the local basis of changing direction, that is, the tangent plane of any point on the manifold. In order to construct the global manifold structure or global density, various methods of splicing these local tangent planes are proposed by different algorithms. In the past decade, domestic and foreign scholars have proposed many improved manifold algorithms. For example, Wang et al. proposed an improved ISOMAP algorithm for the analysis of hyperspectral image features. The improved ISOMAP algorithm selects neighborhood according to spectral Angle, avoiding neighborhood instability in high-dimensional spectral space [18]. Wang et al. proposed a method combining UVE and LLE, which used LLE to reduce the dimension of the image composed of effective wavelength, and used partial least squares discriminant analysis to establish the classification model [19]. In order to solve the local tangent space alignment in the adaptability of higher order information loss problems in the manifold, such as Yang et al. proposed a local neighborhood information extraction of the new algorithm optimization [20], through the optimization of the extraction of tangent vector, which can improve higher dimensional nonuniform distribution manifold dimensionality reduction effect, the proposed algorithm can effectively reconstruct density curve of low dimensional coordinates, the low dimensional high-dimensional image has good adaptability. Huang proposed a sparse discriminant embedded retained projection (SDE) [21], which takes advantage of the advantages of sparsity and manifold structure. It not only preserves the sparse reconstruction relation, but also promotes the manifold structure of discriminating data. Pan et al. put forward a kind of based on local keep rule of two-dimensional local projections (2 DLPP) directly from the image matrix to extract the features [22], the algorithm can better simulate the image characteristics inherent in the manifold structure, improves the characteristics of robustness, and reduces the computational complexity and the characteristics of the final dimension, on the recognition accuracy and recognition speed have achieved good results. However, a potentially serious limitation of these manifold learning algorithms is that they are based on local generalization, mainly using training points near points of interest to infer these local tangent planes. In order to overcome the disadvantage of local generalization, Reference [23] suggested neural network algorithm of Contractive Auto-encoder (CAE), the method by changing the main singular vectors of the Jacobian matrix, captured every input point around the local manifold structure, the corresponding singular value specified in related to the corresponding singular vectors direction how many local changes are credible, and keep in high density area of the input space. In order to maintain a smoother nonlinear manifold structure with data in a low-dimensional space, Rifai et al. proposed a method of Higher Order Contractive auto-encoders (CAE-H) based on CAE [24]. Yu et al. proposed a stacked contractive auto-encoder (SCAE) to improve the robustness of feature extraction through unsupervised training and learning [25]. Aamir et al. proposed an improved variant of CAE based on layered architecture following feed forward mechanism named as deep CAE. By encoding and decoding in each layer of CAE, reconstruction errors were reduced, so as to obtain more robust information features [26]. Ng et al. proposed a denoising-contractive auto-encoder (DCAE), which can learn robust feature representations from noisy and sparse feature vectors [27]. Zhang et al. proposed a new ensemble deep contractive auto-encoder (EDCAE), which automatically learns invariant feature representations by designing a variety of different DCAE models [28]. Due to the Jacobian penalty term in DCAE and different characteristics, these models can deal with various noisy data effectively. Finally, an effective EDCAE model is designed using the combination strategy. The above CAE models can effectively solve the invariance of characteristic jitter to varying degrees. However, due to the poor sparsity of the Frobenius norm of CAE model, the low-dimensional manifold structure cannot be effectively described, which reduces the expression ability of effective information in the data.

In view of the problem of high dimensionality and high data redundancy in hyperspectral remote sensing images, the dimensionality reduction representation of existing hyperspectral data cannot effectively maintain the nonlinear structural relationship in

the data. Starting from the effective characterization of the nonlinear manifold structure of hyperspectral image data in low-dimensional space, this paper analyzes the Jacobian matrix Frobenius norm approximation of CAE, the geometric interpretation of CAE and the Hessian matrix Frobenius norm approximation of CAE-H, and proposes a feature representation method of Higher Order Contractive Auto-Encoders With Nuclear Norm Constraint (CAE-HNC). By introducing Jacobian matrix in the CAE of the nuclear norm constraint, the nuclear norm has better sparsity than the Frobenius norm and can describe the local low dimension of the data manifold better. At the same time, a second-order penalty term is added, which is the Frobenius norm of the Hessian matrix expressed in the hidden layer of the input, encouraging a smoother low-dimensional manifold geometry of the data.

2. Contractive Auto-Encoders

From the point of view of manifold learning, the high-dimensional training data is located on a low-dimensional manifold [23]. Changes in the data correspond to local changes in the manifold (along the direction of the tangent plane), while changes in the data correspond to directions that are orthogonal to the manifold. Therefore, as long as we learn the changes and invariable directions in the data, the manifold structure of the high-dimensional data is also characterized. The goal of Contractive auto-encoders is to learn the manifold structure of high-dimensional data in low-dimensional space. The two driving forces for CAE learning are the contraction penalty term to keep the learned feature constant in all directions (shrinking in all directions), and the reconstruction error term to be able to reconstruct the learned feature back into the input. Therefore, during the learning process, the force of the contraction penalty term makes the change direction in the data (that is, the direction of the manifold tangent plane) able to resist the contraction, which is reflected in the large singular value in its corresponding Jacobian. The directions that do not resist the contractive ability correspond to the invariant directions in the data (orthogonal to the manifold's direction), and the gradient in Jacobian becomes very small. It can be seen that the shrink autoencoder can effectively describe the low-dimensional manifold structure of the data so as to obtain a more compact data representation.

2.1. CAE Model

Contractive auto-encoders is a kind of regular auto-encoder, whose model structure is consistent with the traditional self-encoder. The main purpose of CAE is to suppress the disturbance of the training sample data in all directions and to achieve the effect of local space contraction by adding a penalty term on the target function of the traditional auto-encoder. The penalty term is the Frobenius norm of Jacobian matrix expressed in the hidden layer of input, whose purpose is to shrink the mapping of feature space near the training data, specifically expressed as follows:

$$\|J_f(x)\|_F^2 = \sum_{ij} \left(\frac{\partial h_j(x)}{\partial x_i} \right)^2, \quad (1)$$

where, $J_f(x)$ is Jacobian matrix of $f(x)$, represented as the sum of the partial derivatives of $h_j(x)$ with respect to x_i . $\|\cdot\|_F$ is the Frobenius norm of the matrix. $h_j(x)$ contains j function groups and x_i contains i variables. CAE adds Equation (1) to the loss function as a penalty term to reduce the sensitivity of the model to small changes in input, so as to achieve the purpose of a good system. This formula is used as a penalty term because it has the following characteristics: when the penalty term has a relatively small first-order derivative, it indicates that the hidden layer expression corresponding to the input signal is relatively smooth. Then when the input changes to a certain extent, the hidden layer expression will not change much, which achieves the purpose of being insensitive to the input changes. Therefore, the loss function of CAE can be expressed as follows:

$$J_{CAE}(W, b) = \sum_{i=1}^N \|y_i - x_i\|_2^2 + \frac{\lambda}{N} \sum_{i=1}^N \|J_f(x_i)\|_F^2, \quad (2)$$

where W is the weight vector and b is the bias of the loss function, N is the number of training samples, λ is a super parameter to control the strength of the penalty term, and you can choose any value between 0 and 1. The former term of the loss function is to make the reconstruction error as small as possible, so that CAE can obtain all the information of the input signal as far as possible, and the latter term can be regarded as the information that CAE is discarding as much as possible. Therefore, CAE will finally obtain the disturbance information on the training data, making the model invariable to the disturbance. According to Equation (2), the optimization problem can be described as:

$$\min_{W, b} \left\{ J_{CAE}(W, b) = \sum_{i=1}^N \|y_i - x_i\|_2^2 + \frac{\lambda}{N} \sum_{i=1}^N \|J_f(x_i)\|_F^2 \right\}. \quad (3)$$

Learning parameters are:

$$W, b = \arg\min_{W, b} \left\{ J_{CAE}(W, b) = \sum_{i=1}^N \|y_i - x_i\|_2^2 + \frac{\lambda}{N} \sum_{i=1}^N \|J_f(x_i)\|_F^2 \right\}, \quad (4)$$

$J_f(x)$ is the Jacobian matrix of the hidden layer relative to the input x , $J_f(x) = SW$, where S is the diagonalization of the first derivative function of the function, that is, $S = \text{diag } s'(y)$. When the stochastic gradient descent algorithm is used to solve the parameters, it is necessary to calculate the gradient of each parameter corresponding to the loss function.

CAE can describe the local complex manifold structure around each data point by the singular value decomposition (SVD) of input Jacobian matrix. The corresponding singular value specifies how much local variation is trusted in the direction associated with the corresponding singular vector, while remaining in the dense region of the input space. The penalty term of the CAE loss function excites the insensitivity of $h(x)$ in all input space directions. This pressure is balanced by the need for accurate reconstruction, resulting in $h(x)$ being essentially sensitive to only a few input directions and requiring training samples to distinguish closed input directions. When $J(x)$ contains all the information to calculate the sensitivity of $h = f(x)$ to the motion in any input direction, performing SVD produces a more direct orthonormal basis for the direction of information, ranking from the most sensitive to the least sensitive. The subset of the most sensitive directions in this orthogonal basis can be interpreted as a manifold generated from the tangent space at point x .

2.2. Jacobian's Frobenius Norm Approximation

In order to solve the Frobenius norm of Jacobian expressed by the hidden layer of the input data in the penalty term in CAE, while calculating $\|\nabla_x f(x)\|_2^2$, this paper uses the method of ref [24] of the following:

$$\|\nabla_x f(x)\|_2^2 = \lim_{\sigma \rightarrow 0} \frac{1}{\sigma^2} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 I)} \left[\|J_f(x) - J_f(x + \epsilon)\|_F^2 \right], \quad (5)$$

Where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is an isotropic gaussian distribution with variance of σ^2 . Theoretically, the smaller σ is the more accurate the random approximation. In practice, however, the larger σ is used because it actually allows regularization to explore relatively distant observation points from the data manifold. In fact, the above approximation use n random sampling approximations, that is:

$$\|\nabla_x J(x)\|_F^2 \approx \frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma^2} \|J_f(x) - J_f(x + \epsilon_i)\|_F^2. \quad (6)$$

With parameter of Jacobian Frobenius norm of the gradient, about parameter Θ regular function is:

$$R(\Theta) = \frac{1}{2} \|\nabla_x J(x)\|_F^2 = \lim_{\sigma \rightarrow 0} E \frac{1}{2\sigma^2} [\|J_f(x; \Theta) - J_f(x + \varepsilon; \Theta)\|_F^2]. \quad (7)$$

It can be approximated as:

$$R(\Theta) \approx \frac{1}{2n} \sum_{i=1}^n \frac{1}{\sigma^2} \|J_f(x; \Theta) - J_f(x + \varepsilon_i; \Theta)\|_F^2. \quad (8)$$

Purpose is to calculate gradient on Θ . Function $R(\Theta)$ about Θ differential for:

$$dR(\Theta) \approx \frac{1}{n\sigma^2} \sum_{i=1}^n \left(J_f(x; \Theta) - J_f(x + \varepsilon; \Theta) \right)^T \left(dJ_f(x; \Theta) - dJ_f(x + \varepsilon; \Theta) \right). \quad (9)$$

2.3. Geometric Interpretation of CAE

The regular term in CAE encourages the hidden layer to encode $h(x)$ in an input space that is insensitive in all directions (i.e., the activation function is saturated), which means that the points in the training set in the input space are not different (the so-called sensitivity is the difference). However, the reconstruction task in CAE needs to be able to identify different points in the training set, and this balance makes $h(x)$ sensitive to only a few directions in the input space, so as to be able to distinguish different points near these phases in the training sample set. The geometric interpretation is that these sensitive directions span the locally tangent plane of the manifold. The tangent bundle of a smooth manifold consists of a set of tangent planes along all the sample points on the manifold. Each tangent plane corresponds to a Euclidean coordinate system or chart. In topology, atlas is a collection of such charts. Although the charts collection can form a non-Ou manifold, and each chart is Ou manifold.

Given data set D , $h(x)$, $x \in D$ satisfies the necessary condition of local injectivity, and consider how to define the local chart around x by the property of h . Since h must be sensitive to the change of a sample x_i from itself to one of its adjacent points x_j , but not to other changes, we expect this sensitivity to be reflected in the Jacobian matrix $J(x) = \partial h(x) / \partial x$ spectrum where training is concentrated at each x . Assuming that the rank of $J(x)$ is k , $h(x + \varepsilon v)$ and $h(x)$ are, ideally, only different if v is the span of a singular vector corresponding to a nonzero singular value of $J(x)$. That is to say, the sensitive direction of the singular vector corresponding to the nonzero singular value of $J(x)$, is that is that when x becomes x' , $h(x)$ becomes $h(x')$. In fact, $J(x)$ has many smaller eigenvalues. Therefore, SVD decomposition of $J(x)$ is used:

$$J^T(x) = U(x)S(x)V^T(x). \quad (10)$$

Define the partial chart of x . The tangent plane of sample x is defined as:

$$H_x = \{x + v \mid v \in \text{span}(\mathcal{B}_x)\}. \quad (11)$$

where $\mathcal{B}_x = \{U_k(x) \mid S_{kk}(x) > \epsilon\}$, this is a larger k eigenvalues of the singular vector (columns vectors). The left singular value of Jacobian matrix transpose (gradient matrix) SVD decomposition spans the tangent plane. The coordinates of vector $v \in \text{span}(\mathcal{B}_x)$ on $\mathcal{B}_x$ are the coordinates of vector $v \in \text{span}(\mathcal{B}_x)$ on the tangent plane H_x . Based on the local linear approximation, the atlas described by the encoder function h is defined as:

$$\mathcal{A} = \{(M_x, \mathcal{O}_x) \mid x \in D, \mathcal{O}_x(\bar{x}) = B_x(\bar{x} - x)\}, \quad (12)$$

Given training set sample $x \neq x'$, sensitivity refers to $h(x) \neq h(x')$, and insensitivity refers to $h(x) = h(x')$.

2.4. CAE-H Model and Its Norm Approximation

In order to improve the robustness of the input with small changes, CAE-H improves the objective function on the basis of CAE and adds a second-order penalty term, which

is the Frobenius norm of the Hessian matrix expressed in the hidden layer of the input, as follows:

$$\|H_f(x)\|_F^2 = \left\| \frac{\partial J_f(x)}{\partial x} \right\|_F^2, \quad (13)$$

where, $J_f(x)$ is Jacobian matrix. The Frobenius norm constraint of Hessian matrix is used in CAE-H to punish curvature and encourage smoother manifold structure. From the above equation, the objective function of the CAE-H can be obtained as follows:

$$J_{CAE-H}(W, b) = \sum_{i=1}^N \|y_i - x_i\|_2^2 + \frac{\lambda}{N} \sum_{i=1}^N \|J_f(x_i)\|_F^2 + \frac{\gamma}{N} \sum_{i=1}^N \|H_f(x_i)\|_F^2. \quad (14)$$

Because the second derivative is added, the complexity of the model is greatly increased. The second derivative is converted into the first derivative to reduce the computational complexity. The Hessian Frobenius norm $\|H_f(x)\|_F^2$ can be approximated as:

$$\|H_f(x)\|_F^2 = \lim_{\sigma \rightarrow 0} \frac{1}{\sigma^2} E \left[\|J_f(x) - J_f(x + \varepsilon)\|_F^2 \right]. \quad (15)$$

Therefore, the final objective function of CAE-H is as follows:

$$J_{CAE-H}(W, b) = \sum_{i=1}^N \|y_i - x_i\|_2^2 + \frac{\lambda}{N} \sum_{i=1}^N \|J_f(x_i)\|_F^2 + \gamma E \left[\|J_f(x) - J_f(x + \varepsilon)\|_F^2 \right]. \quad (16)$$

3. CAE-H Based on Nuclear Norm Constraint

Part 2 introduces the shortcomings of the Frobenius norm constrained optimization model in CAE. The advantage of the nuclear norm is that it has a sparse low-dimensional manifold direction, and the nuclear norm is similar to the L_1 norm of singular value vectors. L_1 norm has good sparsity, the model is easy to be interpreted, and the local low dimension of the characterizing manifold is relatively good, which can completely retain the geometric characteristics of the original data, so that the information of the original data is not lost, and the resistance to noise is strong. Therefore, this section proposes a CAE-HNC.

3.1. Definition of Nuclear Norm and Its Jacobian Approximation

The nuclear norm function F , also known as trace norm, is defined as:

$$F(A) = \|A\|_* = \text{Tr}(\sqrt{A^T A}) = \text{Tr}(\sqrt{B}), \quad (17)$$

where $B = A^T A$. The nuclear norm is a convex function which can be optimized effectively and is the best convex approximation of the rank function on the matrix identity sphere with the norm less than 1. When matrix variables are symmetric and positive semidefinite, this heuristic is equivalent to the tracking heuristic often used in control systems. The nuclear norm heuristic has been observed in practice to produce very low-rank solutions, but the theoretical representation of when it produces the minimum rank solution has not been obtained before.

Theorem 1. The Jacobian matrix of the nuclear norm $F(A)$ defined in Equation (17) with respect to A can be approximated as:

$$J_F(A) = D^T [\sqrt{B}; B] A^T + D [\sqrt{B}; B] A^T, \quad (18)$$

where $D [\sqrt{B}; B] = (\sqrt{B}^T \oplus \sqrt{B})^{-1}$, $\oplus$ is Kronecker sum. If $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{m \times m}$, Kronecker sum is defined as:

$$A \oplus B = A \otimes I_m + I_n \otimes B, \quad (19)$$

And

$$\sqrt{B} = VD^{\frac{1}{2}}V^T. \quad (20)$$

Proof of Theorem 1. Since the differential of the nuclear norm function can be expressed as:

$$\begin{aligned} dF &= \text{Tr}\left(D\left[\sqrt{B}; B\right]dB\right) = \text{Tr}\left(D\left[\sqrt{B}; B\right](dA^T A + A^T dA)\right) \\ &= \text{Tr}\left(D\left[\sqrt{B}; B\right]dA^T A\right) + \text{Tr}\left(D\left[\sqrt{B}; B\right]A^T dA\right) \\ &= \text{Tr}\left(dA^T AD\left[\sqrt{B}; B\right]\right) + \text{Tr}\left(D\left[\sqrt{B}; B\right]A^T dA\right) \\ &= \text{Tr}\left(D^T\left[\sqrt{B}; B\right]A^T dA\right) + \text{Tr}\left(D\left[\sqrt{B}; B\right]A^T dA\right) \end{aligned}$$

Therefore, the Jacobian approximation of the nuclear norm function is:

$$J_F(A) = D^T\left[\sqrt{B}; B\right]A^T + D\left[\sqrt{B}; B\right]A^T$$

where $D\left[\sqrt{B}; B\right] = \left(\sqrt{B}^T \oplus \sqrt{B}\right)^{-1}$ and $\oplus$ is Kronecker sum. If $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{m \times m}$, then Kronecker sum is defined as:

$$A \oplus B = A \otimes I_m + I_n \otimes B$$

And

$$\sqrt{B} = VD^{\frac{1}{2}}V^T$$

□

3.2. The Robust CAE-H with Nuclear Norm Constraints

Nuclear norm with sparse is the advantage of low dimensional manifold directions, and nuclear norm is similar to the singular value vector of L_1 norm, and L_1 norm has good sparse, the model is easy to explain, and characterization of the manifold local low dimensional relatively well, which can keep complete geometric characteristics of original data, the original data information is not lost, the stronger noise resistance. CAE-HNC is still a regular auto-encoder, whose main purpose is to suppress the disturbance of the training sample data in all directions, and to better achieve the local space contraction effect by adding a Jacobian nuclear norm constraint penalty term on the objective function, as follows:

$$\|J_f(x)\|_* = \text{Tr}\left(\sqrt{A^T A}\right) = \text{Tr}\left(\sqrt{B}\right). \quad (21)$$

The Frobenius norm of Hessian can be used to describe the geometric structure of the manifold with smoother data. Combining Equation (21), the objective function of the high-order contraction auto-encoder with nuclear norm constraint can be obtained as follows:

$$J_{CAE-HNC}(W, b) = \sum_{i=1}^N \|y_i - x_i\|_2^2 + \frac{\lambda}{N} \sum_{i=1}^N \|J_f(x_i)\|_* + \frac{\gamma}{N} \sum_{i=1}^N \|H_f(x_i)\|_F^2. \quad (22)$$

The high-order contraction auto-encoder with nuclear norm constraint is transformed into the following optimization problem:

$$\min_{W, b} \left\{ J_{CAE-HNC}(W, b) = \sum_{i=1}^N \|y_i - x_i\|_2^2 + \frac{\lambda}{N} \sum_{i=1}^N \|J_f(x_i)\|_* + \frac{\gamma}{N} \sum_{i=1}^N \|H_f(x_i)\|_F^2 \right\}. \quad (23)$$

The learning parameters of solution are:

$$\begin{aligned}
W, b &= \underset{W, b}{\operatorname{argmin}} \{ J_{\text{CAE-HNC}}(W, b) \\
&= \sum_{i=1}^N \|y_i - x_i\|_2^2 + \frac{\lambda}{N} \sum_{i=1}^N \|J_f(x_i)\|_* \\
&\quad + \frac{\gamma}{N} \sum_{i=1}^N \|H_f(x_i)\|_F^2 \}.
\end{aligned} \tag{24}$$

3.3. Solution Algorithm

Aiming at the specific problem of feature representation of wetland hyperspectral remote sensing image, this part proposes a feature representation method of CAE-HNC. The regular terms of the Jacobian matrix with nuclear norm constraint are designed, and the sparse representation of the Jacobian matrix is realized, the model is easier to understand, and the local low dimension of manifolds is easier to describe. The parameter updating process in the solution algorithm of method CAE-HNC is as follows Algorithm 1:

Algorithm 1: High order contractive auto-encoder with nuclear norm constrain

Input: X ;

Output: $\Theta^K = \{W^K, b_h^K, b_r^K, U^K\}$.

1: Initialization: $\Theta = \{W^0, b_h^0, b_r^0, U^0\}$;

2: For $t = 0, 1, 2, \dots, K$;

3: Update W $W^{t+1} = W^t - \eta \nabla_W \varphi(\{W^t, b_h^t, b_r^t, U^t\})$;

4: Update b_h $b_h^{t+1} = b_h^t - \eta \nabla_{b_h} \varphi(\{W^{t+1}, b_h^t, b_r^t, U^t\})$;

5: Update b_r $b_r^{t+1} = b_r^t - \eta \nabla_{b_r} \varphi(\{W^{t+1}, b_h^{t+1}, b_r^t, U^t\})$;

6: Update U $U^{t+1} = P_{\psi^*}^\beta \left(\beta J_f^T(X; W^{t+1}, b_h^{t+1}, b_r^{t+1}) + U^t \right)$;

7: End

4. Experimental Results and Analysis

Three groups of hyperspectral remote sensing images were selected in this paper. In order to verify the effectiveness of the proposed algorithm, experimental simulation analysis was carried out on three groups of hyperspectral images, and feature extraction and representation of different algorithms were performed using hyperspectral data such as CAE, CAE-H, SCAE, DCAE, and CAE-HNC, then classification comparison was conducted to verify the robustness of the proposed method. The experiment was run on a Windows (Inter(R) Celeron(R) G4900 cpu@3.10ghz 3.10 GHz 8GB RAM) 64-bit operating system, using Matlab2016 and ENVI5.3 for simulation verification.

4.1. The Robust CAE-H

In this paper, simulation experiments are carried out on three groups of data, all of which are obtained by spatial and spectral degradation simulation using hyperspectral images, as shown in Figure 1.

The first set of data was collected by an AVIRIS sensor over the Indian Pine Proving Ground in northwestern Indiana. The image size was 145×145 , containing 21,025 pixels. Out of the original 224 spectral bands, 24 water absorption bands (104–108, 150–163 and 220) were removed, and a total of 200 bands participated in the experiment. With a wavelength range of 0.4 to 2.5 μm and a spatial resolution of 20 m, most areas of the image represent fields with a variety of crops, while the rest represent forests and dense vegetation. The map contains 16 different types of crops, including corn, grass, soybean, forest and so on.

The second set of data is a hyperspectral image acquired by ROSIS sensor over the University of Pavia, Italy. The image size is 610×340 and contains 220,740 pixels. There are 103 spectral bands with a wavelength range of 0.4–0.9 μm and a spatial resolution of 1.3 m after removing 12 bands affected by noise. The corresponding map of real features contains 9 types of features, including trees, Asphalt roads, Bricks, Meadows, etc.

The third set of data, also obtained by AVIRIS sensor, is located in the Salinas Valley region of California. Image size is 512×217 , contains 111,104 pixels, removing water

absorption band to participate in a total of 204 regions of the spectrum experiment, the wavelength range of 0.4–2.5 microns, the spatial resolution of 3.7 m, the corresponding real line map contains 16 class features, including Fallow, Celery, etc.

The experimental data have been radiometric correction and geometric registration. The number of available samples for each group of data is shown in Table 1.

Figure 1. Experimental data sets of hyperspectral images. (a) Hyperspectral and ground truth images of Indian Pine; (b) Hyperspectral and ground truth images over University of Pavia; (c) Hyperspectral and ground truth images of Salinas.

Table 1. Number of available samples for each group of data.

Indian Pine		University of Pavia		Salinas	
Type	Number of samples	Type	Number of samples	Type	Number of samples
Alfalfa	46	Asphalt	6631	Brocoli_green_weeds_1	2009
Corn-notill	1428	Meadows	18,649	Brocoli_green_weeds_2	3726
Corn-min	830	Gravel	2099	Fallow	1976
Corn	237	Trees	3064	Fallow_rough_plow	1394
Grass/Pasture	483	Painted metal sheets	1345	Fallow_smooth	2678

Table 1. Cont.

Indian Pine		University of Pavia		Salinas	
Type	Number of samples	Type	Number of samples	Type	Number of samples
Grass/Trees	730	Bare soil	5029	Stubble	3959
Grass/Pasture-mowed	28	Bitumen	1330	Celery	3579
Hay-windrowed	478	Self-Blocking Bricks	3682	Grapes_untrained	11,271
Oats	20	Shadows	947	Soil_vinyard_develop	6203
Soybeans-notill	972			Corn_senesced_green_weeds	3278
Soybeans-min	2455			Lettuce_romaine_4wk	1068
Soybeans-clean	693			Lettuce_romaine_5wk	1927
Wheat	205			Lettuce_romaine_6wk	916
Woods	1265			Lettuce_romaine_7wk	1070
Bldg-Grass-Tree-Drives	386			Vinyard_untrained	7268
Stone-steel towers	93			Vinyard_vertical_trellis	1807

4.2. Classification Results of Hyperspectral Image Features

In order to verify the effectiveness of CAE-HNC feature representation method based on the kernel norm constraint proposed in this paper, this paper uses the CAE-HNC algorithm to extract the feature representation information of hyperspectral images, and uses the SVM classifier to realize the feature classification. The kernel function of SVM adopts gaussian RBF kernel function, the kernel function parameter = 0.3, the penalty factor $C = 50$. In order to test the robustness of the proposed algorithm, 5% and 10% of the hyperspectral data were selected as training samples, and CAE-HNC method was used for data feature extraction. Then 40% of the extracted feature data were used as training samples and 60% as test samples for ground object classification in SVM.

The proposed CAE-HNC method is compared with different methods, including Contractive Auto-encoder (CAE), High-order Contractive auto-encoder (CAE-H), Stacked Contractive Auto-encoder (SCAE) and Denoising Contractive Auto-encoder (DCAE). The experimental results were measured by the Overall Accuracies (OA) A and Kappa coefficient (R) of ground object classification in hyperspectral images. For each group of experiments, the experiment was repeated 10 times, and then the average value of the 10 results was selected for comparison. The experimental results are shown in Table 2, where the results shown in bold are the relatively highest accuracy and Kappa coefficient.

Table 2. Number of available samples in each experimental data set.

Data Set Name	Training Sample	Evaluation	CAE [23]	CAE-H [24]	SCAE [25]	DCAE [27]	Proposed CAE-HNC
Indian Pine (IP)	5%	OA (%)	74.39	80.39	89.06	91.83	93.39
		Kappa	0.7250	0.7831	0.8650	0.9014	0.9278
	10%	OA (%)	75.51	81.16	90.33	91.88	93.87
		Kappa	0.7403	0.7707	0.8403	0.8911	0.9283
University of Pavia (UP)	5%	OA (%)	83.96	89.17	96.06	96.81	98.04
		Kappa	0.8256	0.8807	0.9505	0.9599	0.9701
	10%	OA (%)	84.92	90.35	96.79	97.44	98.13
		Kappa	0.8388	0.8905	0.9515	0.9634	0.9703
Salinas (SV)	5%	OA (%)	76.33	82.77	92.75	93.78	95.36
		Kappa	0.7433	0.8123	0.9143	0.9234	0.9406
	10%	OA (%)	76.99	83.14	94.09	94.95	95.88
		Kappa	0.7527	0.8212	0.9302	0.9401	0.9472

Table 2 shows that the proposed CAE-HNC method can well describe the low-dimensional manifold structure of local features of hyperspectral images, with good robustness and satisfactory classification results.

Figure 2 is the local region magnification of several different feature representation methods on the University of Pavia (UP) data set after SVM classification results, in which the training sample is 5%. As can be seen from Figure 2, SCAE method (Figure 2d), DCAE method (Figure 2e) and CAE-HNC feature representation method (Figure 2f) are superior to CAE (Figure 2b) and CAE-H method (Figure 2c) in overall classification effect. Figure 2d–f can better eliminate salt and pepper noise in classification mapping, so that the image becomes clearer and smoother. In fact, this is the result of effectively improving the phenomena of “different objects with the same spectrum” and “the same object with different spectra”. For ground objects with similar spectrum, such as grass and trees, buildings composed of gravel and bricks, there is a certain degree of misclassification. The CAE-HNC method proposed in this paper has more accurate classification results in the red area of Figure 2. The contours of trees and asphalt buildings are clearer and more in line with the real ground features.

Figure 2. Local region enlarge image of Classification results by different algorithms.

5. Conclusions

Hyperspectral data has a special nonlinear structure in high dimensional space, and this nonlinear structure is also the area where hyperspectral data are distributed and concentrated in high density. This kind of nonlinear structure relation is not effectively described in the dimensionless representation of hyperspectral data, resulting in poor robustness of the feature representation of hyperspectral image. Therefore, this article focus on the study of the characteristics of wetland hyperspectral image said learning method, aimed at effectively depict hyperspectral data in low dimensional space nonlinear manifold structure, analyzes the Jacobian matrix of CAE Frobenius norm approximation, CAE geometric interpretation and CAE-H Frobenius norm Hession matrix approximation,

is proposed based on nuclear norm constraint of high order contractive auto-encoder (CAE-HNC) feature representation. By introducing Jacobian matrix in the CAE of the nuclear norm constraint, the sparse feature is enhanced, the low-dimensional manifold structure of the data is characterized effectively, the Frobenius norm second-order penalty term of the Hessian matrix is added, and the smoother low-dimensional manifold geometry is encouraged. Experiments on hyperspectral images show that CAE-HNC is a compact and robust feature representation method, which provides effective help for the ground object classification and target recognition of hyperspectral images.

Author Contributions: Conceptualization, R.D.; Data curation, X.L.; Methodology, X.Y.; Validation, X.L. and J.S.; Writing—Original draft, X.Y.; Writing—Review and editing, X.Y. and R.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Basic Scientific Research Operating Expenses Project of Provincial Universities in Heilongjiang Province (No.2020-KYYWF-358) and Doctor Initiation Fund Project of Harbin Normal University (No.XKB202113).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, C.J.; Li, G.D.; Lei, R.M.; Du, S.H.; Wu, Z.F. Deep feature aggregation network for hyperspectral remote sensing image classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5314–5325. [\[CrossRef\]](#)
2. Sui, C.H.; Tian, Y.; Xu, Y.P.; Xie, Y. Weighted Spectral-Spatial Classification of Hyperspectral Images via Class-Specific Band Contribution. *IEEE Trans. Geosci. Remote* **2017**, *55*, 7003–7017. [\[CrossRef\]](#)
3. Zhang, B.; Zhao, L.; Zhang, X.L. Three-dimensional convolutional neural network model for tree species classification using airborne hyperspectral images. *Remote Sens. Environ.* **2020**, *247*, 111938. [\[CrossRef\]](#)
4. Zhang, S.Z.; Li, S.T. Spectral-spatial classification of hyperspectral images via multiscale superpixels based sparse representation. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium, Beijing, China, 10–15 July 2016; p. 7729625.
5. Li, Q.T.; Wang, C.Z.; Zhang, B.; Lu, L.L. Object-Based Crop Classification with Landsat-MODIS Enhanced Time-Series Data. *Remote Sens.* **2015**, *7*, 16091–16107. [\[CrossRef\]](#)
6. Calders, K.; Adams, J.S.; Armston, J.; Bartholomeus, H.; Verbeeck, H. Terrestrial laser scanning in forest ecology: Expanding the horizon. *Remote Sens. Environ.* **2020**, *251*, 112102. [\[CrossRef\]](#)
7. Khan, S.; Hussain, S.; Yang, S.K. Contrast Enhancement of Low-Contrast Medical Images Using Modified Contrast Limited Adaptive Histogram Equalization. *J. Med. Imaging Health Inform.* **2020**, *10*, 1795–1803. [\[CrossRef\]](#)
8. Lefkimmatis, S.; Bourquard, A.; Unser, M. Hessian-based Norm Regularization for Image Restoration with Biomedical Applications. *IEEE Trans. Image Process.* **2011**, *21*, 983–995. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Prashnani, M.; Chekuri, R.S. Identification of military vehicles in hyper spectral imagery through spatio-spectral filtering. In Proceedings of the 2013 IEEE Second International Conference on Image Information Processing, Shimla, India, 9–11 December 2013; pp. 527–532.
10. Bioucas-Dias, J.M.; Plaza, A.; Camps-Valls, G.; Scheunders, P.; Nasrabadi, N.M.; Chanussot, J. Hyperspectral Remote Sensing Data Analysis and Future Challenges. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–36. [\[CrossRef\]](#)
11. Li, J.; Marpu, P.R.; Plaza, A.; Bioucas-Dias, J.M.; Benediktsson, J.A. Generalized Composite Kernel Framework for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote* **2013**, *51*, 4816–4829. [\[CrossRef\]](#)
12. Zhou, Y.C.; Peng, J.T.; Chen, C.L.P. Dimension Reduction Using Spatial and Spectral Regularized Local Discriminant Embedding for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote* **2015**, *53*, 1082–1095. [\[CrossRef\]](#)
13. Garcia-Salgado, B.P.; Ponomaryov, V.I.; Sadovnychiy, S.; Reyes-Reyes, R. Efficient dimension reduction of hyperspectral images for big data remote sensing applications. *J. Appl. Remote Sens.* **2020**, *14*, 032611. [\[CrossRef\]](#)
14. Deng, Y.J.; Li, H.C.; Pan, L.; Shao, L.Y.; Du, Q.; Emery, W. Modified Tensor Locality Preserving Projection for Dimensionality Reduction of Hyperspectral Images. *IEEE Geosci. Remote Sens.* **2018**, *15*, 277–281. [\[CrossRef\]](#)
15. Gu, Y.F.; Liu, T.Z.; Jia, X.P.; Benediktsson, J.A.; Chanussot, J. Nonlinear Multiple Kernel Learning with Multiple-Structure-Element Extended Morphological Profiles for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote* **2016**, *54*, 3235–3247. [\[CrossRef\]](#)
16. Zhang, P.; He, H.X.; Gao, L.R. A Nonlinear and Explicit Framework of Supervised Manifold-Feature Extraction for Hyperspectral Image Classification. *Neurocomputing* **2019**, *337*, 315–324. [\[CrossRef\]](#)
17. Wen, J.H.; Yan, W.D.; Lin, W. Supervised linear manifold learning feature extraction for hyperspectral image classification. In Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014; pp. 3710–3713.
18. Wang, L.L.; Li, Z.Y.; Sun, J.X. Improved ISOMAP Algorithm for Anomaly Detection in Hyperspectral Images. In Proceedings of the Fourth International Conference on Machine Vision, Singapore, 9–10 December 2011; p. 834902.

19. Wang, B.J.; Huang, M.; Zhu, Q.B.; Wang, S. UVE-LLE Classification of Apple Mealiness Based on Hyperspectral Scattering Image. *Acta Photonica Sin.* **2011**, *40*, 1132–1136. [[CrossRef](#)]
20. Yang, Q.W.; Li, Y.; Sun, F.C. Independent Component Analysis: Embedded LTSA. In *Foundations and Applications of Intelligent Systems*; Springer: Berlin, Germany, 2014; Volume 213, pp. 711–722.
21. Huang, H. Classification of Hyperspectral Remote-sensing Images Based on Sparse Manifold Learning. *J. Appl. Remote Sens.* **2013**, *7*, 464–477. [[CrossRef](#)]
22. Pan, X.; Ruan, Q.Q. Palmprint Recognition with Improved Two-dimensional Locality Preserving Projections. *Image Vision Comput.* **2008**, *26*, 1261–1268. [[CrossRef](#)]
23. Rifai, S.; Vincent, P.; Muller, X.; Glorot, X.; Bengio, Y. Contractive Auto-encoders: Explicit Invariance During Feature Extraction. In Proceedings of the 28th International Conference on International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011; pp. 836–844.
24. Rifai, S.; Mesnil, G.; Vincent, P.; Muller, X.; Glorot, X. Higher Order Contractive Auto-Encoder. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Athens, Greece, 5–9 September 2011; pp. 645–660.
25. Yu, Y.Z.; Hui, J. A Study on Text Classification based on Stacked Contractive Auto-Encoder. In Proceedings of the International Conference on Electronics Instrumentation and Information Systems, Harbin, China, 3–5 June 2017; pp. 1–6.
26. Aamir, M.; Nawari, N.M.; Wahid, F.; Mahdin, H. A Deep Contractive Autoencoder for Solving Multiclass Classification Problems. *Evol. Intell.* **2020**, *14*, 1619–1633. [[CrossRef](#)]
27. Ng, P.C.; She, J. Denoising-Contractive Autoencoder for Robust Device-free Occupancy Detection. *IEEE Internet Things* **2019**, *6*, 9572–9582. [[CrossRef](#)]
28. Zhang, Y.Y.; Li, X.Y.; Gao, L.; Chen, W.; Li, P.G. Ensemble Deep Contractive Auto-encoders for Intelligent Fault Diagnosis of Machines under Noisy Environment. *Knowl.-Based Syst.* **2020**, *196*, 105764. [[CrossRef](#)]