axioms

MDPI

Article

Unsupervised Feature Selection with Latent Relationship Penalty Term

Ziping Ma [1,*], Yulei Huang [1], Huirong Li [2] and Jingyu Wang [1]

[1] School of Mathematics and Information Science, North Minzu University, Yinchuan 750030, China
[2] School of Mathematics and Computer Application, Shangluo University, Shangluo 726000, China
* Correspondence: 2006041@nmu.edu.cn; Tel.: +86-139-9528-6373

Abstract: With the exponential growth of high dimensional unlabeled data, unsupervised feature selection (UFS) has attracted considerable attention due to its excellent performance in machine learning. Existing UFS methods implicitly assigned the same attribute score to each sample, which disregarded the distinctiveness of features and weakened the clustering performance of UFS methods to some extent. To alleviate these issues, a novel UFS method is proposed, named unsupervised feature selection with latent relationship penalty term (LRPFS). Firstly, latent learning is innovatively designed by assigning explicitly an attribute score to each sample according to its unique importance in clustering results. With this strategy, the inevitable noise interference can be removed effectively while retaining the intrinsic structure of data samples. Secondly, an appropriate sparse model is incorporated into the penalty term to further optimize its roles as follows: (1) It imposes potential constraints on the feature matrix to guarantee the uniqueness of the solution. (2) The interconnection between data instances is established by a pairwise relationship situation. Extensive experiments on benchmark datasets demonstrate that the proposed method is superior to relevant state-of-the-art algorithms with an average improvement of 10.17% in terms of accuracy.

Keywords: unsupervised feature selection; latent relationship penalty term; attribute score; sparse model

MSC: 68Q99

check for updates

Citation: Ma, Z.; Huang, Y.; Li, H.; Wang, J. Unsupervised Feature Selection with Latent Relationship Penalty Term. *Axioms* **2024**, *13*, 6. https://doi.org/10.3390/axioms13010006

Academic Editor: Javier Fernandez

Received: 16 November 2023
Revised: 14 December 2023
Accepted: 18 December 2023
Published: 21 December 2023

Copyright: © 2023 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

With the explosive growth of data and information, dimensionality reduction techniques have become a crucial step in machine learning and data mining [1,2]. The primary dimensionality reduction techniques involve nonnegative matrix factorization (NMF) [3], principal component analysis (PCA) [4], locally linear embedding (LLE) [5], and feature selection (FS) [6]. These dimensionality reduction techniques are beneficial in accelerating the speed of the model's learning and enhancing clustering performance and prediction accuracy. Typically, feature selection is an effective strategy for dimension reduction due to its excellent property in removing unimportant or meaningless features, which will certainly enhance the interpretability of these models. Consequently, these approaches have been applied in various applied fields such as gene expression analysis [7], image processing [8], natural language processing [9], and other fields.

According to the availability of data labels [10], feature selection methods can be categorized as supervised feature selection methods (SFS) [11,12], semi-supervised feature selection methods (SSFS) [13,14], and unsupervised feature selection methods (UFS) [15,16]. Supervised feature selection methods and semi-supervised feature selection methods can identify discriminative features by effectively mining the latent information in labeled data. However, in practice acquiring label data is extremely time-consuming, especially in some cases there exists unreliability of labelled data. Unsupervised feature selection methods are

conducted to address this difficult challenge by identifying discriminative features without the availability of labeled data. As a result, it has garnered increasing attention in recent years. Generally, the strength of unsupervised feature selection methods lies in the ability to achieve satisficed results by assessing the importance of features based on appropriate criteria. However, there are several key factors that unsupervised feature selection methods need to emphasize.

Firstly, in unsupervised scenarios, it is crucial to design various strategies for extracting pseudo-label information to compensate for the absence of labeled data. Generally, the pseudo-label information is performed in UFS based on spectral regression [17,18]. In the unsupervised feature selection (UFS) algorithm, the essence of spectral regression is to construct a similarity measure that is employed to learn pseudo-label information such that each sample is more accurately guaranteed to be categorized into the ground truth class. For example, Zhao et al. [19] proposed a feature selection framework based on spectral regression using the spectrum of graphs to explore the intrinsic properties of features. Cai et al. [20] designed a two-step strategy by adopting spectral regression in the first step to retain the multi-cluster structure of the data, and sparse regression in the second step to simplify calculations. By performing this two-step strategy, the potential correlation between different features can be well investigated, thereby efficiently selecting more discriminative features. In contrast, Hou et al. [21] utilized a one-step strategy by embedding sparse regression and spectral regression into a joint learning framework, in which the clustering performance outperforms the above two-step strategy in literature [20]. Unfortunately, the pseudo-label information generated by spectral analysis in UFS often generally has the drawback of negative values and an inaccurate similarity matrix due to noise interference, which is a challenging issue in UFS.

Secondly, it is significant to explore the internal structure of the original data. It is noted that the sparsity of high-dimensional data implies that such data contains manifold information. That is, the feature subset obtained by UFS should preserve this manifold information. To meet this requirement, a large number of relevant methods have been presented to investigate the internal structure of data in UFS. For example, He et al. [22] exploited local manifold information to evaluate the importance of features by calculating the Laplacian score (LapScor). Shang et al. [23] introduced a graph regularization term into the objective function of UFS and constructed a feature graph to enable the feature vectors of the original data to be consistent with the vectors in the coefficient matrix. Thus, the feature subset preserves this manifold information by learning the manifold information through a similarity matrix. Liu et al. [24] constructed a loss term with ℓ_1-norm constraint to maintain the local geometric structure of the data through linear coefficients. To address the unreliability of constructing the similarity matrix in the above UFS methods, Nie et al. [25] constructed a structured graph optimization to learn the similarity matrix adaptively, in which the manifold information will be well preserved by constructing a more satisfactory similarity matrix. Based on the method in [25], Li et al. [26] and Chen et al. [27] innovatively extended structured graph optimization. The former incorporated maximum entropy with the generalized uncorrelated regression model into the method described in [25], in which local manifold information of the data can be retained so that uncorrelated yet distinctive features are effectively selected. The latter derived a flexible optimal graph by leveraging a flexible low-dimensional manifold embedding mechanism to compensate for the unreliability of the conventional UFS methods. However, the above methods ignore the dependency between the data instances (that is, mining interconnection information between the data instances is not well understood yet).

Thirdly, it is important to exploit interconnection information between data instances in that interconnection information inherently implies in data instances. Exploring this interconnection information will lead to effectively improving the clustering performance and reducing the impact of inevitable noises. For this reason, Tang et al. [28] embedded latent representation learning into UFS, named LRLMR, in which latent representation learning learned an affinity matrix with correlation information between data instances.

Based on LRLMR, Shang et al. [29] designed DSLRL, in which the affinity matrices regarding interconnection information are constructed to fully investigate the interconnection information. The distinction between LRLMR and DSLRL lies in the fact that the former only considers the interconnection information among data instances, whereas the latter takes into account the interconnection information existing between both data instances and features.

These methods mentioned above can address the above three issues respectively and effectively (i.e., designing successful strategies for extracting pseudo-label information to compensate for the absence of labeled data, exploring the internal structure of data instances, and exploiting interconnection information between data instances). However, the exploration of individual uniqueness in the existing literature remains an unresolved challenge. The reason may lie in that these methods generally assume the importance of each sample is no ranking so that the attribute score of each sample is assigned equally. Such an assumption is unsuitable since it neglects the diversity and particularity (i.e., the uniqueness of data instances). Specifically, three groups of face images illustrated in Figure 1a–c respectively chosen from three classes in the ORL dataset [30]. From a human visual perspective, each individual in Figure 1 is unique whether it belongs to the same class or a different class. Consequently, it is more reasonable to assume an attribute score is flexibly assigned to each individual according to their uniqueness. In other words, the scores of individuals within the same class tend to become more similar, while the scores of individuals from different classes tend to become more distinct. Therefore, emphasizing individual uniqueness can facilitate the effective distinction of individuals and eliminate the interference of redundant features to some extent. Although some researchers [31–33] have introduced the idea of LDA into UFS to minimize the within-class distance and maximize the between-class distance, they still ignore the learning of individual uniqueness.

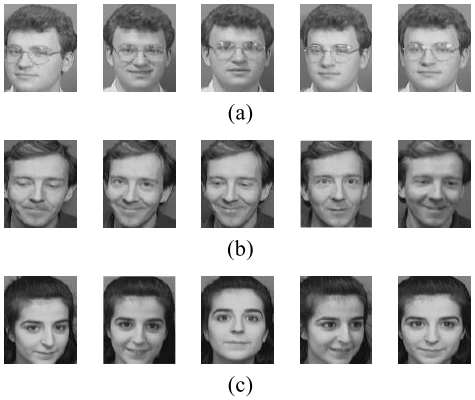


Figure 1. Samples from the ORL dataset. (**a**–**c**) are from different categories of the ORL dataset, respectively.

To alleviate the above issues outlined, in this paper a novel embedded unsupervised feature selection method, called unsupervised feature selection with latent relationship penalty term (LRPFS), is proposed. The original intention of LRPFS is to explore the data structure, pseudo-label information, interconnection information, and individual uniqueness. Specifically, (1) a novel method is developed to preserve the spatial data structure by quantifying sample distances in space via inner product relations; (2) we evaluate the uniqueness of samples based on their unique contributions to the whole, and determine the uniqueness of samples through pairwise relationships based on the principles of latent representation learning and symmetric nonnegative matrix factorization. Meanwhile, these pairwise relationships also create connections between samples, enabling the subspace matrix to provide pseudo-label information.

The main contributions of this paper are summarized as follows:

- A novel unsupervised feature selection with latent relationship penalty term is presented, which simultaneously performs an improved latent representation learning on the attribute scores of the subspace matrix and imposes a sparse term on the feature transformation matrix.
- The latent relationship penalty term, an improved latent representation learning, is proposed. By constructing a novel affinity measurement strategy based on pairwise relationships and attribute scores of samples, this penalty term can exploit the uniqueness of samples to reduce interference from noise and ensure that the spatial structure of both the original data and the subspace data remains consistent, which is different from other existing models.
- An optimum algorithm with convergence is designed and extensive experiments are conducted: (1) LRPFS has shown superior performance on publicly available datasets compared to other existing models in terms of clustering performance and more remarkable capability of discriminative feature selection. (2) Experiments verify that LRPFS has fast convergence, short computation time, and significant performance by explicitly evaluating an attribute score for each individual to eliminate redundant features.

The remaining sections of this article are structured as follows. Section 2 reviews the concepts of latent representation learning and inner product space about feature selection. In Section 3, the latent relationship penalty term is introduced, and the LRPFS method is presented. Section 4 demonstrates the superiority of the proposed method over other advanced algorithms through experimental design and analyses the properties of the algorithm itself. Finally, Section 5 provides the conclusion and discusses future work.

2. Related Works

In this section, latent representation learning in feature selection is briefly reviewed. In addition, the inner product space and some notations are introduced.

2.1. Notations

The following notations are illustrated in this paper. For example, an arbitrary $A = [a_1, a_2, \ldots, a_n]^T \in \mathcal{R}^{n \times d}$ is donated as a matrix, a_i indicates that the i-th row of matrix A is a vector, a_{ij} represents the elements of the i-th row and j-th column of the matrix A, the F-norm of the matrix A is defined as $\|A\|_F = (\sum_{i=1}^{n} \sum_{j=1}^{d} |a_{ij}|^2)^{1/2}$, and the $\ell_{2,1}$-norm of the matrix A is defined as $\|A\|_{2,1} = \sum_{i=1}^{n} \left(\sum_{j=1}^{d} a_{ij}{}^2 \right)^{1/2}$.

2.2. Review of Feature Selection and Latent Representation Learning

Generally, feature selection methods are classified into three categories according to the strategy of feature evaluation: filter [22,34], wrapper [35,36], and embedded [37,38]. The filter feature selection aims to evaluate each feature directly with a specific ranking criterion, i.e., variance, Laplacian score, feature similarity, and trace ratio, to select a feature subspace [39]. However, the filter feature selection only considers the feature itself but ignores the interdependence between features [40]. In contrast, the wrapper feature selection constructs a quasi-optimal subset of features by emphasizing feature combinations and correlations between features [41]. In most approaches based on wrapper feature selection, the algorithm complexity tends to increase with the dimension of data space, which may lead to computationally expensive. In addition, to achieve an effective feature subset, these methods based on embedded feature selection integrate feature selection with model learning by adjusting feature priority in the learning iteration process. Compared to filter feature selection, embedded feature selection has gained more attention due to its superior performance in that it can reduce feature redundancy more effectively except that it has greater robustness, whereas, compared to wrapper feature selection, embedded feature selection effectively reduces training time and cost, and speeds up computation. Nevertheless, these approaches based on the above three kinds have drawbacks such as premature convergence and obtaining local optimum.

Traditional UFS methods assume that data are independently and identically distributed, whereas, in the real world, samples are correlated with each other. With the development of embedded feature selection, embedding latent representation learning into UFS has been widely applied to investigate the interconnection information between samples such that it can reveal the latent structure in the data [28,42,43]. Latent representation learning employs a model of symmetric nonnegative matrix factorization [44] to learn the interconnection information between samples. The specific representation is as follows:

$$R = \|T - VV^T\|_F^2 = \sum_{i,j=1}^{n} |t_{ij} - v_i v_j^T|^2, \tag{1}$$

where $T \in \mathcal{R}^{n \times n}$ is the affinity matrix containing interconnection information between samples [28], $V \in \mathcal{R}^{n \times f}$ represents the latent representation matrix, n is the number of samples, and f is the number of latent variables. The t_{ij} is defined as follows [29]:

$$t_{ij} = exp(\frac{\left\|x_i - x_j\right\|^2}{-2{\sigma_1}^2}), \tag{2}$$

where $i, j = 1, 2, \ldots, n$, σ_1 is a bandwidth parameter. The latent representation learning is integrated into the UFS, and the objective function is expressed as follows:

$$\begin{gathered} O = \|XW - V\|_F^2 + \alpha\|T - VV^T\|_F^2, \\ s.t.\ V \geq 0, \end{gathered} \tag{3}$$

where α is the balance parameter, $X \in \mathcal{R}^{n \times d}$ is the data matrix, and $W \in \mathcal{R}^{d \times f}$ is the feature transformation matrix. The first term of the objective function is the loss function, which enables matrix X to approximate matrix V under the influence of matrix W. Consequently, matrix V can be considered as a subspace of X. Moreover, under the influence of latent representation learning, matrix V can be used as a pseudo-label matrix to guide feature selection. Through the application of latent representation learning, the performance and efficiency of these models can be enhanced in that there exists a degree of reduction in noise and redundant information.

2.3. Inner Product Space

The essence of feature selection is to find a suitable feature subset to represent the original data, which is the mapping of data from a high-dimensional space to a low-dimensional space. Therefore, exploring the spatial structure has become an important issue in feature selection [1,45,46]. First, the objective is to aggregate several elements into a cohesive set with an establishment of the "relationship" or "structure" among these elements in a set to construct a space. However, there are various spaces in mathematics, such as metric space, vector space, normed linear space, and inner product space. Among these spaces, the inner product space adds a "structure" [47], i.e., inner product, in which the angles and lengths of vectors are discussed and it can possess the properties of non-negativity such as non-degeneracy, conjugate symmetry, first-variable linearity and second-variable conjugate linearity. Therefore, in this paper, our objective is to explore samples in the inner product space and a framework that can potentially preserve the structure of the data space by integrating the structure of different elements in the higher-dimensional space into the lower-dimensional space via inner product operations.

3. Methodology

In this section, a novel regularization term, known as the latent relationship penalty term, is presented and incorporated into the derivation of the LRPFS mechanism. Additionally, an optimization algorithm, convergence analysis, and computational complexity analysis for LRPFS are provided.

3.1. Latent Relationship Penalty Term

In a practical application, each individual exhibits distinct characteristics (i.e., uniqueness) that contribute to the whole model, which assumes specific roles in data representation and establishing interconnections with the entire dataset. However, it is important to note that this individual uniqueness is not solely dependent on irrelevant and redundant features, but rather on the presence of significant and informative features. The characteristics of these features play a crucial role in capturing the essence of the data and enabling meaningful representations. Therefore, our objective is to acquire a spatial subset that can accurately identify the significant features while preserving individual uniqueness. According to this premise, a novel latent relationship penalty term based on Equation (1) is developed to exploit individual uniqueness while maintaining the inherent structural relationships within the data through pairwise associations. Subsequently, the methodology and principles of the latent relationship penalty term are presented to illustrate the proposed method, which involves two main steps in constructing the latent relation penalty term as follows.

3.1.1. Preservation of Data Structures

The dataset $X \in \mathcal{R}^{n\times d}$ can be considered as a distribution of n individuals $x_i (i = 1, 2, \ldots, n)$ within a d dimensional linear space, and $\langle x_i, x_j \rangle$ reflects inherent (inner product) relationship between two vectors. The subspace matrix $V \in \mathcal{R}^{n\times f}$ is designated to represent the underlying structure of the dataset X. Based on this, our objective is to preserve the approximate data structure between individuals in both the original space and the subspace by leveraging the inner product space metric distance. This is accomplished through the pairwise relationship of vector multiplication by ensuring that $\langle v_i, v_j \rangle \approx \langle x_i, x_j \rangle$. As a result, this pairing establishes a direct correspondence between the high-dimensional inner product space and the low-dimensional inner product space, as depicted in the following Equation (4):

$$R_1 = \sum_{i,j=1}^{n} |\langle v_i, v_j \rangle - \lambda \langle x_i, x_j \rangle|^2 = \sum_{i,j=1}^{n} |v_i v_j^T - \lambda x_i x_j^T|^2 = \|VV^T - \lambda XX^T\|_F^2, \quad (4)$$

where λ is a scale parameter to regulate the scale relationship between the original data matrix X and the subspace matrix V. Accounting for the presence of inherent noise in the original dataset, which frequently undermines the integrity of the data structure, our approach focuses on mitigating the influence of irrelevant and redundant features within each sample. To achieve this, the uniqueness of each sample should be exploited, thereby enhancing the preservation of the underlying data structure.

3.1.2. Exploring the Uniqueness of Individuals

To assess the uniqueness of individuals, the concept of an attribute score, denoted as q, is introduced. This score is referred to as the contribution of each individual to the overall dataset by taking into account their interrelationship with the entire sample. Specifically, the attribute score q_{ii} is defined as $\|s_i\|_1$, where q_{ii} represents the score of the i-th sample x_i, and s_i is the i-th vector in the similarity matrix S. To construct the similarity matrix, $S \in \mathcal{R}^{n\times n}$, a k-neighbourhood graph, denoted as N_k, is employed. The value of k is set to 0 or 5, where a value of 0 corresponds to a complete graph. The similarity matrix is defined as follows:

$$[S]_{ij} = \begin{cases} exp(\frac{-\|x_i - x_j\|^2}{2\sigma^2}), & if\ x_j \in N_k(x_i) \\ 0, & otherwise \end{cases} \quad i, j = 1, 2, 3, \ldots, n, \quad (5)$$

where σ is the width parameter [23], the obtained score q is introduced into Equation (4) such that $v_i \approx q_{ii} x_i$, resulting in the final expression for the latent relationship penalty term as follows.

$$R_2 = \sum_{i,j=1}^{n} \left| v_i v_j^T - \lambda q_{ii} q_{jj} x_i x_j^T \right|^2. \quad (6)$$

Classical latent representation learning in UFS typically employs Gaussian functions to measure interconnection information between samples, as demonstrated in Equations (1) and (2). In contrast, Equation (6) introduces a novel measurement approach by defining t_{ij} as $\lambda q_{ii} q_{jj} x_i x_j^T$. The affinity matrix constructed using this method not only leverages the uniqueness of individuals as a prior condition to mitigate noise interference but also regulates the structural approximation consistency between the original space and the subspace by capturing pairwise relationships among samples.

3.2. Objective Function

Aiming to incorporate Equation (6) into UFS such that the latent relational penalty term has a potentially constraining function on the feature transformation matrix $W \in \mathcal{R}^{d \times f}$, the objective function is summarized as follows.

$$\begin{gathered}(W^*, V^*) = arg\ min \sum_{i=1}^{n} \|x_i W - v_i\|_2^2 + \sum_{i,j=1}^{n} |v_i v_j^T - \lambda q_{ii} q_{jj} x_i x_j^T|^2, \\ s.t. W > 0, V > 0.\end{gathered} \tag{7}$$

To prevent the occurrence of trivial solutions, a potentially constraining function is integrated by the latent relationship penalty term. The detailed process is presented in Theorem 1. Furthermore, by imposing the latent relationship penalty term, the subspace matrix V can act as a pseudo-label matrix within the UFS framework to guide the feature selection process. Our goal is to acquire a sparse feature transformation matrix W to improve the efficiency of feature selection. To achieve this purpose, the $\ell_{2,1}$-norm regularization term is introduced, which results in the final expression of the objective function as follows:

$$\begin{gathered}\left(W^*, V^*\right) = arg\ min \sum_{i=1}^{n} \|x_i W - v_i\|_2^2 + \sum_{i,j=1}^{n} |v_i v_j^T - \lambda q_{ii} q_{jj} x_i x_j^T|^2 + \alpha \|W\|_{2,1}, \\ s.t. W > 0, V > 0,\end{gathered} \tag{8}$$

where α is a sparsity constraint parameter to adjust the sparsity of W. The feature transformation matrix W is obtained by the optimization of the objective function with the score of each feature calculated using $\|w_i\|_2$. The higher the score, the more important the features are. The top l features are selected to generate a new data matrix X_{new} by ranking the feature scores in descending order.

3.3. Optimization

To simplify the operation, all variables in the objective function of LRPFS are represented as matrices, and Equation (8) can be substituted as follows.

$$\begin{gathered}(W^*, V^*) = arg\ min \|XW - V\|_F^2 + \|VV^T - \lambda Q X X^T Q^T\|_F^2 + \alpha \|W\|_{2,1}, \\ s.t. W > 0, V > 0,\end{gathered} \tag{9}$$

where $Q \in \mathcal{R}^{n \times n}$ is a diagonal matrix and q_{ii} $(i = 1, 2, \ldots, n)$ is the i-th diagonal element in the matrix Q. The model (9) is a nonconvex problem concerning W and V, so it is not practical to find the global optimal solution at the same time. Nevertheless, the model is convex concerning the other variable when one variable is fixed, therefore this model can be solved by alternately optimizing W and V, respectively. The Lagrange function is constructed as follows.

$$L = \|XW - V\|_F^2 + \|VV^T - \lambda Q X X^T Q^T\|_F^2 + \alpha Tr(W^T U W) + Tr(\varphi W) + Tr(\varnothing V), \tag{10}$$

where, $U \in \mathcal{R}^{d \times d}$ is a diagonal matrix. The calculation of the i-th diagonal element u_{ii} of U is performed as follows:

$$u_{ii} = \frac{1}{2\|W_i\|_2}. \tag{11}$$

To avoid overflow, a sufficiently small constant ε is introduced, leading to the following rewriting of Equation (10):

$$u_{ii} = \frac{1}{2\max(\|W_i\|_2, \varepsilon)}. \tag{12}$$

(1) Fix V and update W:

The partial derivative of the Lagrange Function (11) with respect to W is computed, and further results in the following expression:

$$\frac{\partial L}{\partial W} = 2X^T XW - 2X^T V + 2\alpha UW + \varphi. \tag{13}$$

According to the Karush-Kuhn-Tucker (KKT) condition [48], the following iterative update formula for W can be derived.

$$w_{ij} \leftarrow w_{ij} \frac{[X^T V]_{ij}}{[X^T XW + \alpha UW]_{ij}}. \tag{14}$$

(2) Fix W and update V:

Similar to the optimization variable W, the partial derivative of the Lagrange Function (10) for V is taken, yielding the following result:

$$\frac{\partial L}{\partial V} = -2XW + 2V + 4VV^T V - 4\lambda QXX^T Q^T V + \varnothing. \tag{15}$$

Following the Karush-Kuhn-Tucker (KKT) condition, the following iterative update formula for V is obtained.

$$v_{ij} \leftarrow v_{ij} \frac{[XW + 2\lambda QXX^T Q^T V]_{ij}}{[V + 2VV^T V]_{ij}}. \tag{16}$$

With the above analysis, Algorithm 1 summarizes the procedure of LRPFS. The pipeline of LRPFS is visualized in Figure 2.

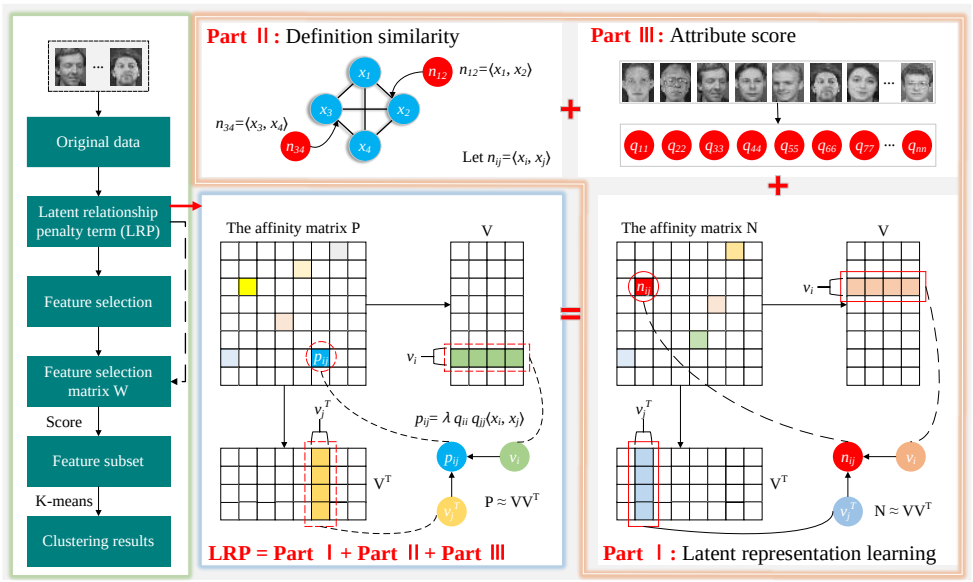


Figure 2. The framework of LRPFS method.

Algorithm 1: LRPFS algorithm steps

1: **Input:** Data matrix $X \in \mathcal{R}^{n \times d}$; Parameters α, λ, f, and k; The maximum number of iterations $maxIter$;
2: **Initialization:** The iteration time $t = 0$; $W = \text{rand}(d, f)$; $V = \text{rand}(n, f)$; $U = \text{eye}(m)$; Construct the attribute score matrix Q;
3: **while not converged do**
4: Update W using $w_{ij} \leftarrow w_{ij} \frac{[X^T V]_{ij}}{[X^T XW + \alpha UW]_{ij}}$;
5: Update V using $v_{ij} \leftarrow v_{ij} \frac{[XW + 2\lambda QXX^T Q^T V]_{ij}}{[V + 2VV^T V]_{ij}}$;
6: Update U using $u_{ii} = \frac{1}{2\max(\|W_i\|_2, \varepsilon)}$;
7: Update t by: $t = t + 1$, $t \leq maxIter$;
8: **end while**
9: **Output:** The feature transformation matrix W and the subspace matrix V.
10: **Feature selection:** Calculate the scores of the d features according to $\|w_i\|_2$ and select the first l features with high scores.

The potential constraint of error function embedding latent relationship penalty term in LRPFS can be explained by the following Theorem 1:

Theorem 1. *Let $X \in \mathcal{R}^{n \times d}$, $V \in \mathcal{R}^{n \times c}$, and $W \in \mathcal{R}^{d \times c}$, Assume that there exists a X with a left inverse, such that $XW \approx V$ and $VV^T \approx \lambda QXX^T Q^T$, then the term $WW^T \approx \lambda X_L QXX^T Q^T X_R$ (λ is a scalar, X_L is the left inverse of X, and X_R is the right inverse of X).*

Proof. Properties of one side inverse matrices [49]: If the matrix $M \in \mathcal{R}^{n \times d}$ has rank $\rho(M) = d$, then there exists a left inverse matrix $M_L \in \mathcal{R}^{d \times n}$ such that $M_L M = I_d$. Similarly, the matrix $M^T \in \mathcal{R}^{d \times n}$ exists a right inverse matrix $M_R \in \mathcal{R}^{n \times d}$ such that $M^T M_R = I_d$, where $I_d \in \mathcal{R}^{d \times d}$ is an identity matrix.

According to the properties above, if X is one side inverse matrice and $\rho(X) = d$, then X has a left inverse X_L and X^T has a right inverse X_R, and the proof is as follows:

$$\begin{aligned} VV^T &\approx \lambda QXX^T Q^T, \text{ (Since } XW \approx V) \\ &\Leftrightarrow XWW^T X^T \approx \lambda QXX^T Q^T, \\ &\Leftrightarrow X_L XWW^T X^T \approx \lambda X_L QXX^T Q^T, \\ &\Leftrightarrow WW^T X^T \approx \lambda X_L QXX^T Q^T, \\ &\Leftrightarrow WW^T X^T X_R \approx \lambda X_L QXX^T Q^T X_R, \\ &\Leftrightarrow WW^T \approx \lambda X_L QXX^T Q^T X_R. \end{aligned} \tag{17}$$

The constraint on WW^T can be derived by $WW^T \approx \lambda X_L QXX^T Q^T X_R$, indicating a latent relationship by embedding constraint on W. □

In Theorem 1, the scenario where X has a left inverse is presented. However, in practical applications, it's possible that the matrix X does not possess a left inverse, resulting in an approximation of $XWW^T X^T \approx \lambda QXX^T Q^T$. The final expressions in both situations express the relationship between the matrix W and the known matrices Q and X. The latent relationship penalty term, through this potential constraint, makes the generated W more accurate and avoids trivial solutions.

3.4. Convergence Analysis

This subsection proves the convergence of LRPFS by demonstrating that under the update rules (14) and (16), the objective Function (8) is monotonically decreasing.

First, the introduction of a theorem [45] is necessary, which provides theoretical assurance for the convergence of LRPFS.

Definition 1. *If there is a function* $G(x,\ x')$ *such that* $F(x)$ *satisfies:*

$$G(x, x') \geq F(x), G(x, x) = F(x). \tag{18}$$

Thus $F(x)$ *is nonincreasing function under the following updating formula:*

$$x^{(t+1)} = \arg\min_{x} G(x, x^{(t)}), \tag{19}$$

where $G(x,\ x')$ *is an auxiliary function for* $F(x)$.

Proof.

$$F(x^{(t+1)}) \leq G(x^{(t+1)}, x^{(t)}) \leq G(x^{(t)}, x^{(t)}) = F(x^{(t)}). \tag{20}$$

If the objective function is proved to be monotonic, the objective function is retained to contain the W term, and get the following equations:

$$F(W) = \|XW - V\|_F^2 + \alpha\|W\|_{2,1}. \tag{21}$$

Through the computation of first-order and second-order partial derivatives of $F(W)$ with respect to W, the following expressions can be derived:

$$F'_{ij} = [2X^T XW - 2X^T V + 2\alpha UW]_{ij}, \tag{22}$$

$$F''_{ij} = [2X^T X + 2\alpha U]_{ii}. \tag{23}$$

□

Lemma 1.

$$G(W_{ij}, W_{ij}^{(t)}) = F_{ij}(W_{ij}^{(t)}) + F'_{ij}(W_{ij}^{(t)})(W_{ij} - W_{ij}^{(t)}) + \frac{[X^T XW + \alpha UW]_{ij}}{W_{ij}^{(t)}}(W_{ij} - W_{ij}^{(t)})^2, \tag{24}$$

where $G(W_{ij},\ W_{ij}^{(t)})$ *is the auxiliary function of* F_{ij}. *When* $W_{ij} = W_{ij}^{(t)}$, $G(W_{ij}^{(t)}, W_{ij}^{(t)}) = F_{ij}(W_{ij}^{(t)})$.

Proof. The Taylor series expansion of $F_{ij}(W_{ij})$ is:

$$F_{ij}(W_{ij}) = F_{ij}(W_{ij}^{(t)}) + F'_{ij}(W_{ij}^{(t)})(W_{ij} - W_{ij}^{(t)}) + [X^T X + \alpha U]_{ii}(W_{ij} - W_{ij}^{(t)})^2. \tag{25}$$

$G(W_{ij}, W_{ij}^{(t)}) \geq F_{ij}(W_{ij})$ is equivalent to:

$$\frac{[X^T XW + \alpha UW]_{ij}}{W_{ij}^{(t)}} \geq [X^T X + \alpha U]_{ii}. \tag{26}$$

Since:

$$[X^T XW + \alpha UW]_{ij} = \sum_k [X^T X + \alpha U]_{ik} W_{kj}^{(t)} \geq [X^T X + \alpha U]_{ii} W_{ij}^{(t)}. \tag{27}$$

The inequality $G(W_{ij}, W_{ij}^{(t)}) \geq F_{ij}(W_{ij})$ holds. □

Next, it is demonstrated that, in accordance with the iterative update rule (14), F_{ij} exhibits a monotonically decreasing.

Proof. Substituting Equation (24) with $x^{(t+1)} = \arg\min_x G(x, x^{(t)})$.

$$W_{ij}^{(t+1)} = W_{ij}^{(t)} - W_{ij}^{(t)} \frac{F'_{ij}(W_{ij}^{(t)})}{2[X^T XW + \alpha UW]_{ij}} = W_{ij}^{(t)} \frac{[X^T V]_{ij}}{[X^T XW + \alpha UW]_{ij}}. \quad (28)$$

It can be seen from the updating rules of W that F_{ij} monotonically decreases under updating (14). The proof of the updating rules of V is similar to that of W such that the acquisition of updating rule (16) will be generated. Therefore, the conclusion can be drawn that F_{ij} exhibits a monotonically decreasing trend, and the objective function of LRPFS converges. □

3.5. Computational Time Complexity

The time complexity of LRPFS consists of two main parts. The first part involves constructing the feature score matrix Q, which has a time complexity of $O(dn^2)$. The second part involves iteratively optimizing the feature transformation matrix W and the subspace matrix V, with a calculation complexity of $O(nd^2)$ per iteration. Therefore, the total time complexity is $O(dn^2 + tnd^2)$, where t is the number of iterations.

4. Experiments

In this section, the superiority of LRPFS is demonstrated through a series of experiments conducted on benchmark datasets. These experiments consist of two main parts: comparative experiments (Section 4.5) and LRPFS analysis experiments (Section 4.6). All of the experimental results are implemented with MATLAB R2018b on a Windows machine with 3.10-GHZ i5-11300H, 16-GB main memory. The code of our proposed LRPFS is available at https://github.com/huangyulei1/LRPFS accessed on 5 December 2023.

4.1. Datasets

The benchmark datasets include COIL20, Colon, Isolet, JAFFE, Yale64, PIE [29], nci9, PCMAC, Lung_dis, and TOX_171, downloaded at https://jundongl.github.io/scikit-feature/datasets.html, accessed on 3 August 2022, and https://www.face-rec.org/databases/, accessed on 3 August 2022, and Table 1 illustrates the details of these datasets.

Table 1. Details of nine datasets.

No.	Datasets	Samples	Features	Class	Type
1	COIL20	1440	1024	20	Object image
2	Colon	62	2000	2	Biological
3	Isolet	1560	617	26	Speech Signal
4	JAFFE	213	676	10	Face image
5	Lung_dis	73	325	7	Biological
6	nci9	60	9712	9	Biological
7	PCMAC	1943	3289	2	Text
8	PIE	2856	1024	68	Face image
9	TOX_171	171	5748	4	Biological
10	Yale64	165	4096	11	Face image

4.2. Comparison Methods

Since LRPFS belongs to UFS, the comparison experiments are performed under unsupervised conditions, and the selected 10 state-of-the-art UFS methods are briefly described as follows.

Baseline: The method utilizes the original dataset as a feature subset for clustering.

LapScor [22]: A classical filter FS method to evaluate features by local preservation ability.

SPEC [19]: It selects feature subsets by utilizing spectral regression.

MCFS [20]: A two-step framework for selecting features is constructed by combining spectral regression and sparse regression.

UDFS [40]: Based on $\ell_{2,1}$-norm minimization and discriminant analysis ideas, a joint framework is imposed to guide the UFS.

SOGFS [25]: The method adaptively learns the local manifold structure and constructs a more accurate similarity matrix to select more discriminative features.

SRCFS [46]: The idea of collaboration and randomization of multiple subspaces under high-dimensional space is introduced to select more discriminable features by exploring the ability of various subspaces.

RNE [24]: A method to preserve local geometric structure is constructed through a novel robust objective function.

inf-FS_U [39]: It assigns a score for each feature through graph theory to obtain a feature subset.

S^2DFS [38]: This method constructs a parameter-free UFS method based on the trace ratio criterion with $\ell_{2,0}$-norm constraint to maintain more feature discrimination power.

4.3. Evaluation Metrics

In our experiments, the performance of the proposed LRPFS is evaluated by using two evaluation metrics: clustering accuracy (ACC) and normalized mutual information (NMI) [50]. The values of these evaluation metrics range from 0 to 1, and the higher the value is, the better the performance of the algorithm is.

4.4. Experimental Settings

Concerning the parameter settings, the number of k-neighbours is set to 0 or 5 for these methods that require the construction of a similarity matrix, and the parameter σ is fixed at 10 [23]. For the inf-FS_U method, the parameter α is tuned in the range of $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$. For the RNE method, the parameter range is set according to ref. [24]. For our LRPFS method, both the scaling parameter λ and the sparse parameter α are searched from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$, and the number of latent variables f is default to the number of classes of the dataset. For the remaining methods, the parameters are tuned in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$. In the experimental process, the maximum number of iterations $maxIter$ is set to 30 and the iteration will be terminated early when the objective function value (Obj) satisfies $|Obj(t) - Obj(t-1)|/Obj(t-1) < 10^{-6}$, where $Obj(t)$ denotes the objective function value for the t-th iteration. The number of feature subsets l varies in {20, 30, 40, 50, 60, 70, 80, 90, 100}. Since the result of k-means depends on the initialization, we repeat the clustering 20 times independently and take the means and standard deviations as the final results.

4.5. Comparison Experiment

The performance of LRPFS is compared with 10 state-of-the-art FS methods on 9 datasets, i.e., COIL20, Colon, Isolet, JAFFE, Lung_dis, nci9, PCMAC, PIE and TOX_171. Firstly, feature selection is performed to achieve feature subsets from the datasets. Secondly, k-means is applied to the feature subsets to derive clustering results. Finally, the results of these algorithms are evaluated by two metrics (i.e., ACC and NMI).

According to the above experimental settings, the clustering results (i.e., ACC, NMI) of LRPFS and the comparison methods on nine datasets are shown in Tables 2 and 3, where the best results for each dataset are bolded, the second-best results are underlined and labelled with the number of selected features. Table 4 illustrates the running time of all algorithms on various datasets. According to these Tables, it can be seen that LRPFS outperforms all other comparison methods in terms of ACC, while in most cases, the NMI values of LRPFS are higher than other algorithms, which fully demonstrates the effectiveness of selecting discriminative features. The specific summary is as follows.

(1) Overall, most of the UFS methods outperform baseline across a majority of datasets. This performance differential highlights the substantial superiority of these UFS methods in effectively eliminating irrelevant and redundant features.

(2) The results presented in Tables 2 and 3 indicate that our proposed method, LRPFS, achieves significant performance compared to other state-of-the-art techniques. Specially, LRPFS showcases a substantial increase in Accuracy (ACC) of 32.26%, 30.65%, 30.57%, 33.87%, 24.84%, 25.81%, 29.04%, 24.2%, 14.76%, and 1.54%, respectively, as compared to baseline, LapScor, SPEC, MCFS, UDFS, SOGFS, SRCFS, RNE, inf-FS_U, and S^2DFS. The main reason for this phenomenon is that LRPFS excels in extracting the inherent information within the data structure and assigning unique attribute scores to individual samples. These distinct attributes significantly contribute to its outstanding performance, especially on the Colon dataset.

(3) The insights revealed by the results in Table 4 serve to emphasize LRPFS's remarkable competitiveness in terms of computation time against a significant proportion of the algorithms under comparison. While the performance of LRPFS might be slightly inferior to baseline, LapScor, and MCFS, it still exhibits excellent computational efficiency, coupled with the highest clustering accuracy. This is particularly evident when comparing LRPFS with UDFS, SOGFS, RNE, inf-FS_U, and S^2DFS. Compared with baseline, the running time of LRPFS on some datasets with few samples, such as Colon, nci9, and TOX_171, is slower than baseline. The main reason is that the process of selecting discriminant features takes a certain amount of time. However, when dealing with datasets with large samples such as PIE and Isolet, the running time of LRPFS is superior to baseline and the clustering accuracy is also significantly improved, which verifies the dimension reduction capability of LRPFS and provides a theoretical basis for the implementation of practical problems.

(4) MCFS performs better than SPEC on Isolet, JAFFE, Lung_dis, PCMAC, and PIE since MCFS takes sparse regression into account in the FS model, which can improve the learning ability of the model. Specially, S^2DFS is slightly better than UDFS even if there exists the same idea of discriminant analysis in S^2DFS and UDFS. The reason may lie in that in S^2DFS a trace ratio criterion framework with $\ell_{2,0}$-norm constraint plays a positive role.

(5) RNE, SOGFS, and LRPFS exhibit commendable performance, affirming the importance of capturing the underlying manifold structure inherent in the data. Notably, SOGFS outperforms RNE across some datasets, especially on JAFFE, Lung_dis, nci9, and PIE. This distinction can be attributed to SOGFS's incorporation of an adaptive graph mechanism, thereby engendering more precise similarity matrices. Unlike the aforementioned two techniques, LRPFS introduces the refinement of attribute scores to mitigate the harmful impact of noise while preserving the inherent data structure. This distinctive attribute underscores the superiority of LRPFS to a certain degree.

Table 2. ACC (MEAN ± STD % (The number of selected features)) of different algorithms on real-world datasets, where the best results for each dataset are bolded, the second-best results are underlined.

Methods	COIL20	Colon	Isolet	JAFFE	Lung_dis	nci9	PCMAC	PIE	TOX_171
Baseline	65.75 ±4.16 (all)	54.84 ±0.00 (all)	61.73 ±2.77 (all)	82.04 ±5.59 (all)	73.63 ±5.26 (all)	40.75 ±5.26 (all)	50.49 ±0.00 (all)	24.68 ±1.09 (all)	44.77 ±3.93 (all)
LapScor	60.41 ±2.11 (100)	56.45 ±0.00 (40)	55.83 ±2.14 (100)	76.74 ±4.58 (50)	70.41 ±7.34 (90)	37.58 ±3.08 (80)	50.23 ±0.00 (50)	39.00 ±1.05 (70)	52.81 ±0.27 (20)
SPEC	64.74 ±3.47 (90)	56.53 ±0.36 (30)	46.84 ±1.89 (100)	80.94 ±5.35 (100)	71.03 ±5.38 (100)	46.33 ±4.17 (50)	50.08 ±0.00 (20)	17.88 ±0.89 (100)	50.32 ±1.31 (90)

Table 2. *Cont.*

Methods	COIL20	Colon	Isolet	JAFFE	Lung_dis	nci9	PCMAC	PIE	TOX_171
MCFS	64.22 ±3.37 (50)	53.23 ±0.00 (60)	56.81 ±2.20 (90)	85.40 ±4.41 (100)	81.92 ±4.80 (80)	45.58 ±3.26 (40)	50.13 ±0.00 (20)	27.87 ±1.48 (70)	43.63 ±1.87 (20)
UDFS	58.71 ±2.14 (100)	62.26 ±1.22 (30)	42.65 ±1.72 (100)	84.55 ±4.10 (100)	77.60 ±6.66 (90)	35.25 ±2.25 (60)	51.02 ±0.39 (30)	20.74 ±0.69 (100)	45.06 ±4.18 (70)
SOGFS	57.83 ±2.78 (100)	61.29 ±0.00 (30)	45.44 ±1.88 (100)	86.03 ±4.78 (70)	76.71 ±5.50 (90)	42.75 ±4.43 (50)	52.50 ±0.41 (80)	36.60 ±1.01 (30)	49.80 ±2.21 (80)
SRCFS	57.14 ±2.68 (100)	58.06 ±0.00 (100)	55.91 ±2.04 (100)	76.17 ±4.65 (100)	70.68 ±5.36 (80)	39.33 ±2.98 (90)	50.49 ±0.00 (100)	39.77 ±1.05 (90)	47.46 ±0.29 (100)
RNE	61.52 ±1.91 (70)	62.90 ±0.00 (70)	49.44 ±1.51 (70)	73.66 ±4.52 (90)	73.29 ±5.32 (90)	36.08 ±3.56 (30)	53.86 ±4.82 (20)	27.82 ±0.83 (40)	54.35 ±3.64 (80)
inf-FS_U	58.32 ±2.69 (100)	72.34 ±0.79 (50)	56.75 ±1.35 (100)	66.01 ±3.83 (100)	78.90 ±6.21 (80)	32.92 ±2.85 (100)	50.75 ±0.00 (40)	40.32 ±1.11 (100)	39.94 ±1.02 (30)
S^2DFS	67.10 ±3.18 (60)	85.56 ±0.36 (30)	63.64 ±2.09 (100)	81.46 ±7.67 (100)	78.70 ±4.76 (80)	46.50 ±4.25 (100)	50.08 ±0.00 (30)	27.79 ±0.81 (60)	46.55 ±2.43 (50)
LRPFS	**69.31 ±3.17 (90)**	**87.10 ±1.17 (70)**	**66.41 ±1.74 (100)**	**86.31 ±3.63 (60)**	**82.23 ±4.53 (60)**	**47.25 ±4.69 (30)**	**57.84 ±0.97 (60)**	**41.62 ±1.43 (70)**	**54.44 ±0.87 (50)**

Table 3. NMI (MEAN ± STD % (The number of selected features)) of different algorithms on real-world datasets, where the best results for each dataset are bolded, the second-best results are underlined.

Methods	COIL20	Colon	Isolet	JAFFE	Lung_dis	nci9	PCMAC	PIE	TOX_171
Baseline	**76.69 ±1.99 (all)**	0.60 ±0.00 (all)	**76.06 ±1.26 (all)**	83.61 ±3.37 (all)	69.27 ±4.21 (all)	37.96 ±5.92 (all)	0.01 ±0.00 (all)	48.84 ±0.62 (all)	24.17 ±3.73 (all)
LapScor	69.67 ±1.18 (100)	0.97 ±0.00 (40)	69.45 ±0.91 (100)	83.45 ±2.28 (50)	64.86 ±5.71 (90)	36.49 ±1.66 (40)	0.58 ±0.00 (50)	64.51 ±0.65 (70)	34.93 ±0.36 (20)
SPEC	73.52 ±1.49 (100)	1.75 ±0.02 (30)	59.89 ±1.38 (100)	83.92 ±3.73 (80)	67.09 ±3.55 (100)	45.41 ±3.64 (40)	0.45 ±0.00 (20)	42.57 ±0.43 (20)	25.64 ±1.49 (90)
MCFS	74.14 ±1.90 (60)	0.10 ±0.00 (20)	69.80 ±0.66 (100)	85.82 ±2.20 (100)	74.01 ±4.11 (80)	45.76 ±3.27 (40)	0.41 ±0.00 (20)	50.89 ±0.69 (70)	19.00 ±3.78 (60)
UDFS	69.43 ±0.99 (100)	3.12 ±0.63 (30)	57.41 ±0.87 (100)	84.95 ±2.69 (90)	69.80 ±4.65 (90)	34.99 ±2.95 (20)	0.11 ±0.05 (30)	44.13 ±0.38 (100)	15.78 ±5.16 (100)
SOGFS	70.79 ±1.72 (100)	7.79 ±0.00 (30)	61.35 ±0.36 (100)	**88.08 ±2.97 (70)**	70.60 ±2.57 (100)	42.42 ±3.77 (70)	**2.16 ±0.52 (80)**	58.55 ±0.53 (30)	28.18 ±4.16 (100)

Table 3. *Cont.*

Methods	COIL20	Colon	Isolet	JAFFE	Lung_dis	nci9	PCMAC	PIE	TOX_171
SRCFS	69.20 ±0.98 (100)	1.83 ±0.00 (100)	68.18 ±0.98 (100)	81.18 ±3.89 (100)	65.45 ±4.25 (100)	37.65 ±3.32 (80)	0.63 ±0.00 (20)	64.95 ±0.86 (90)	31.40 ±0.30 (40)
RNE	72.01 ±1.56 (100)	3.95 ±0.00 (70)	62.85 ±1.01 (100)	78.76 ±3.70 (90)	67.31 ±4.40 (90)	33.66 ±3.73 (30)	1.26 ±1.24 (20)	54.42 ±0.55 (40)	28.26 ±4.07 (80)
inf-FS_U	68.65 ±1.16 (100)	16.80 ±0.77 (50)	71.52 ±0.58 (100)	69.26 ±2.28 (100)	74.60 ±5.81 (80)	27.53 ±3.30 (100)	0.21 ±0.00 (40)	**65.66 ±0.73 (100)**	13.11 ±1.17 (60)
S^2DFS	76.33 ±1.43 (80)	40.70 ±0.78 (30)	74.75 ±0.83 (100)	83.85 ±4.07 (100)	74.43 ±3.25 (80)	46.55 ±4.06 (100)	0.29 ±0.00 (20)	52.64 ±0.64 (60)	30.18 ±2.36 (100)
LRPFS	**76.55 ±1.03 (100)**	**41.80 ±3.10 (70)**	**75.96 ±0.61 (100)**	86.71 ±2.12 (90)	**76.63 ±4.38 (60)**	**47.12 ±4.56 (70)**	2.00 ±0.30 (90)	64.97 ±0.48 (80)	**35.80 ±1.00 (80)**

Table 4. Computation time (seconds) of different methods on real-world datasets.

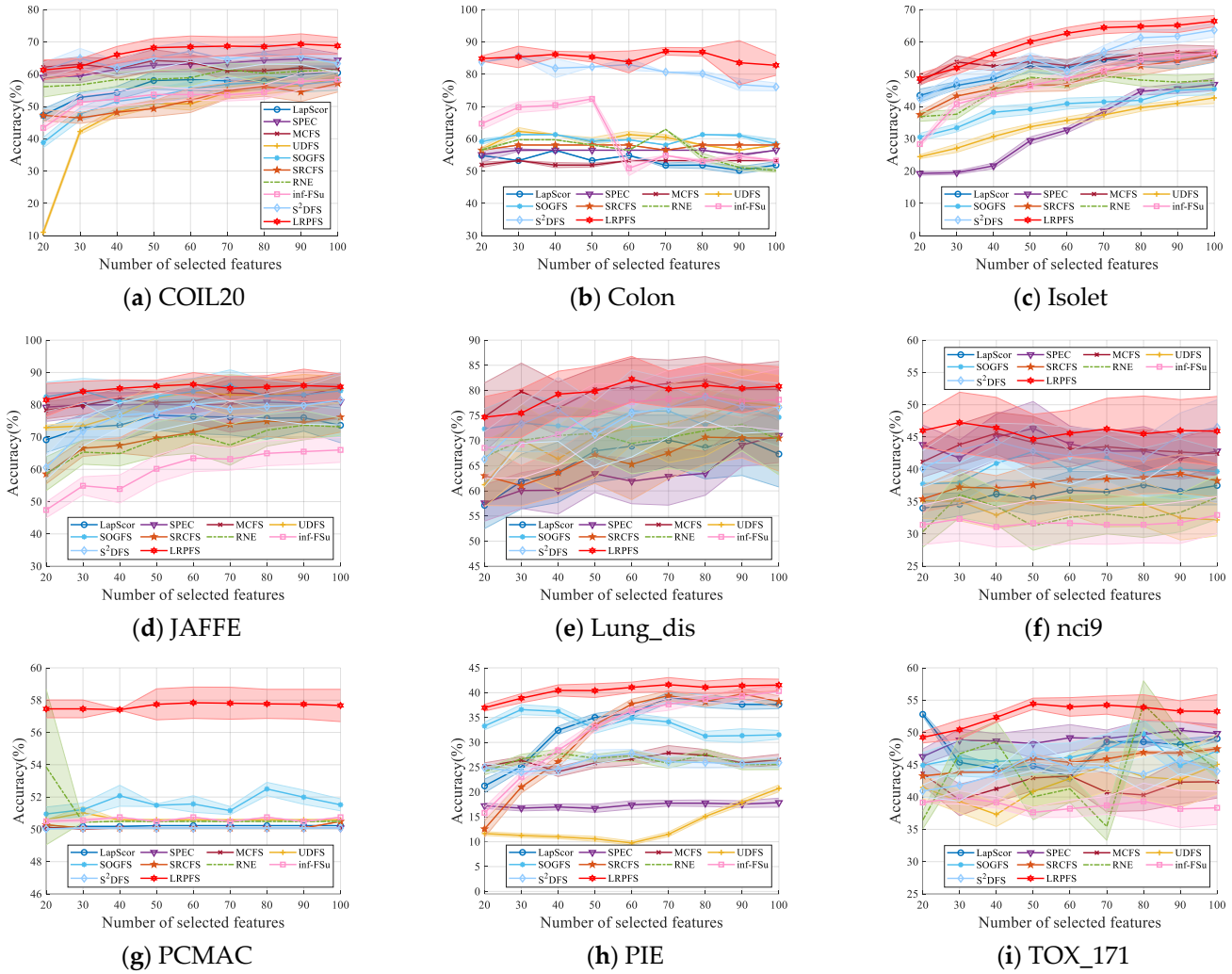
Methods	COIL20	Colon	Isolet	JAFFE	Lung_dis	nci9	PCMAC	PIE	TOX_171
Baseline	24.91	0.62	19.13	1.64	0.43	2.56	3.04	99.57	7.84
LapScor	5.83	0.38	7.96	0.95	0.35	0.47	1.75	33.24	0.94
SPEC	10.58	0.31	12.94	0.96	0.41	0.54	16.52	53.50	1.08
MCFS	5.97	1.01	8.01	1.24	0.65	2.14	2.83	30.88	1.39
UDFS	13.94	12.43	13.41	1.48	0.52	1198.87	67.97	55.85	314.23
SOGFS	96.53	3.09	23.73	1.98	0.86	12,137.22	608.57	58.83	929.49
SRCFS	10.53	0.53	12.91	1.22	0.55	0.62	12.59	51.55	1.42
RNE	12.71	29.53	12.67	5.42	1.36	476.33	50.15	33.79	179.91
inf-FS_U	10.75	2.88	9.31	1.90	0.61	55.31	50.91	47.07	28.03
S^2DFS	7.21	12.80	7.86	1.77	0.65	1379.36	60.78	31.01	291.46
LRPFS	9.27	1.24	12.83	1.19	0.41	23.24	29.56	52.40	8.66

To further investigate the effect of the number of selected features on LRPFS, the clustering performance of various methods is illustrated upon different numbers of features as shown in Figures 3 and 4, where the horizontal coordinate indicates the number of features selected according to the FS methods, the vertical coordinate denotes the clustering performance and the shaded section represents the error range of ACC and NMI. It can be explicitly observed that the curves of LRPFS are mostly uppermost, especially on Colon, PCMAC, and PIE, which achieves a satisfactory performance and demonstrates the superiority of LRPFS over other compared methods.

To verify the noise reduction ability of LRPFS, noise tests are conducted on the COIL20 dataset with random noise of 8×8, 12×12 and 16×16 sizes added to each sample (32×32) respectively to generate three synthetic datasets as shown in Figure 5b–d, and the clustering results are shown in Table 5. It can be seen that LRPFS is superior to other comparison methods under the influence of various noises and still achieves excellent performance, especially in Figure 5b, it is extremely difficult to select significant features since most features of the pictures are blocked according to the excessive size of the noise. However, LRPFS with latent relationship penalty term still achieves satisfactory results, for example, the ACC of LRPFS is 9.06% higher than that of RNE on the 16×16 noised COIL20 datasets. Consequently, LRPFS has the strong learning capability of identifying discriminative features and diminishing noise.

Table 5. ACC and NMI (MEAN ± STD %) on the noised COIL20 datasets, where the best results for each dataset are bolded, the second-best results are underlined.

Methods	Accuracy (%)			Normalized Mutual Information (%)		
	8 × 8 Noise	12 × 12 Noise	16 × 16 Noise	8 × 8 Noise	12 × 12 Noise	16 × 16 Noise
LapScor	56.03 ± 3.43	58.72 ± 2.96	58.32 ± 3.56	67.53 ± 1.06	69.83 ± 0.84	68.82 ± 1.28
SPEC	63.90 ± 2.78	63.78 ± 2.34	56.46 ± 2.29	72.18 ± 1.23	73.35 ± 1.34	67.75 ± 0.92
MCFS	64.01 ± 3.58	63.94 ± 2.02	61.96 ± 2.45	<u>73.69 ± 1.84</u>	73.37 ± 1.08	<u>70.77 ± 1.06</u>
UDFS	59.28 ± 2.68	63.45 ± 2.49	60.85 ± 2.31	69.15 ± 1.33	73.12 ± 1.17	68.27 ± 1.48
SOGFS	57.19 ± 2.39	57.60 ± 2.59	57.43 ± 2.88	70.41 ± 1.10	70.47 ± 0.77	69.85 ± 1.60
SRCFS	57.70 ± 3.19	58.20 ± 2.00	57.86 ± 2.42	69.81 ± 1.50	69.87 ± 1.28	68.13 ± 1.00
RNE	62.05 ± 4.12	61.63 ± 2.98	53.52 ± 2.35	72.44 ± 1.62	71.74 ± 1.44	65.76 ± 0.88
inf-FS_U	57.33 ± 2.52	58.01 ± 2.27	59.23 ± 2.19	69.06 ± 1.12	68.07 ± 1.20	68.21 ± 1.42
S^2DFS	<u>65.39 ± 3.19</u>	<u>65.75 ± 3.94</u>	<u>62.39 ± 3.27</u>	73.27 ± 1.81	<u>73.40 ± 1.88</u>	70.76 ± 1.55
LRPFS	**65.95 ± 2.79**	**66.85 ± 2.50**	**62.58 ± 2.21**	**73.71 ± 1.33**	**74.14 ± 1.15**	**70.94 ± 1.28**



(**a**) COIL20 (**b**) Colon (**c**) Isolet

(**d**) JAFFE (**e**) Lung_dis (**f**) nci9

(**g**) PCMAC (**h**) PIE (**i**) TOX_171

Figure 3. The ACC of all of the algorithms for selecting different numbers of features on the nine datasets.

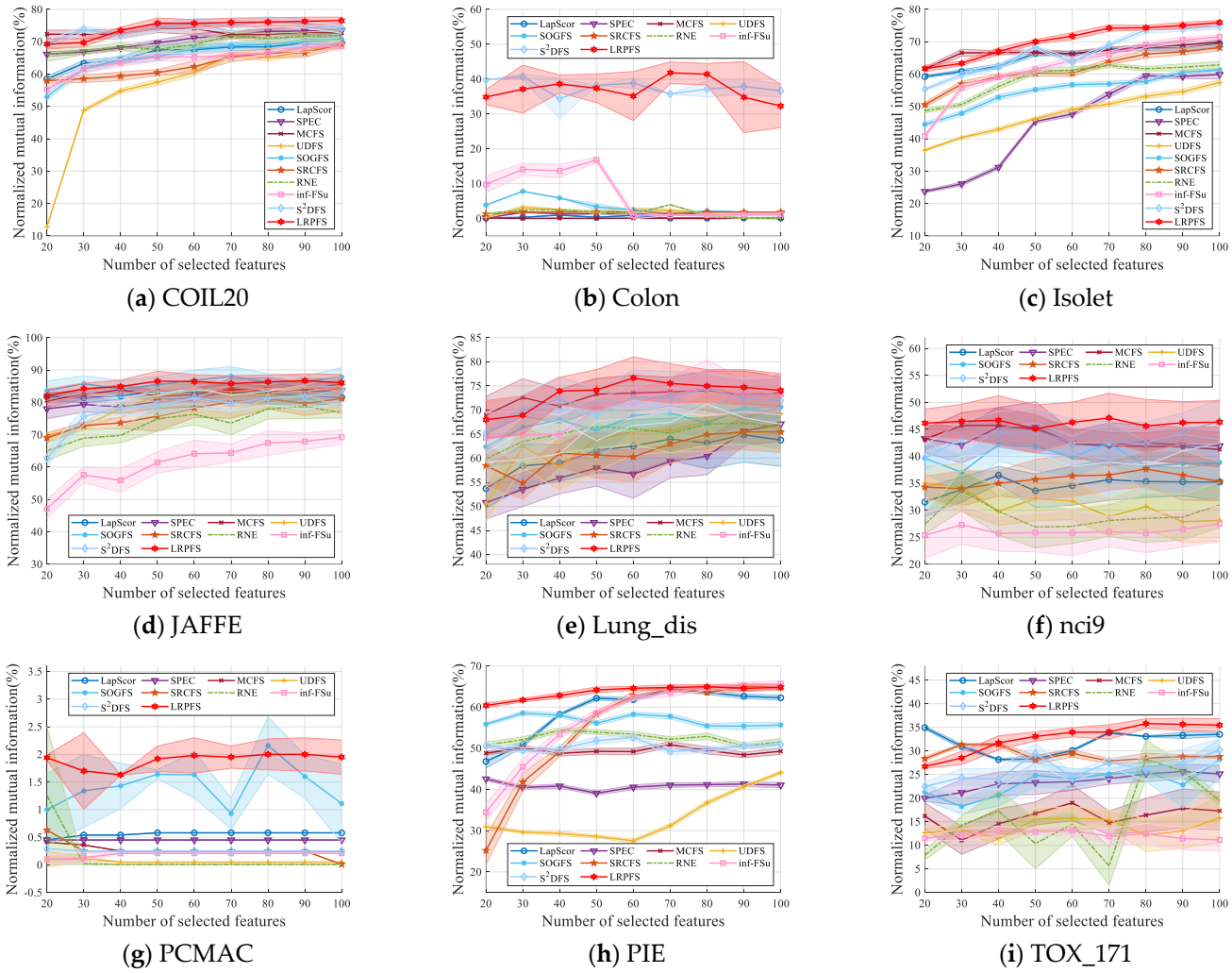


Figure 4. The NMI of all of the algorithms for selecting different numbers of features on the nine datasets.

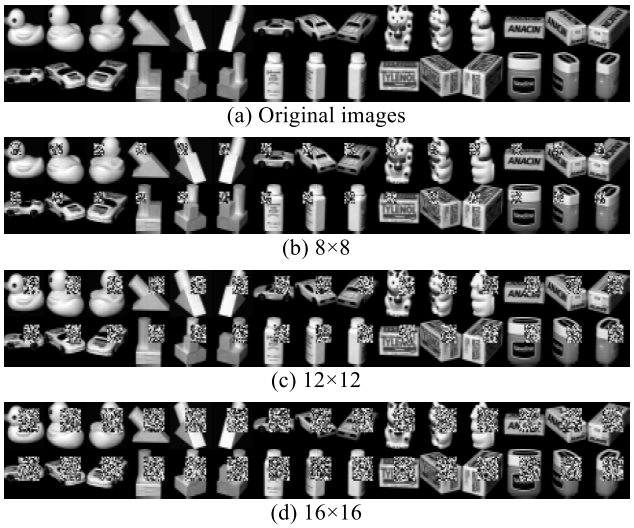
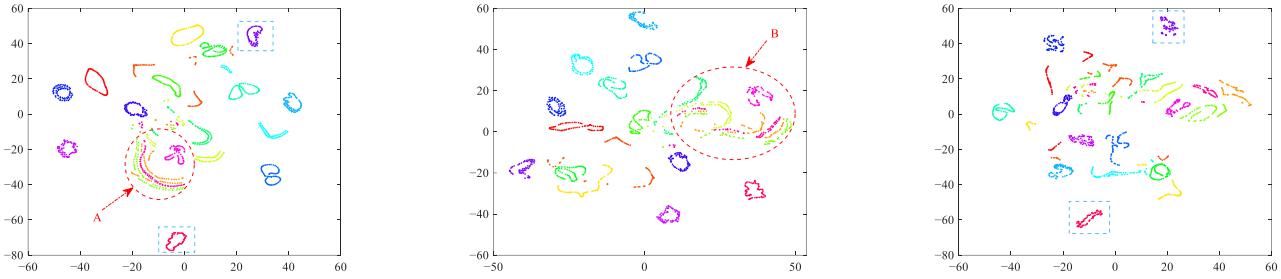


Figure 5. Samples from COIL20 dataset with noise of different sizes.

The feature subset obtained from the feature selection method on the COIL20 dataset is visualized using t-SNE. In our experiments, a comparative experiment is conducted

with the Baseline, S^{2}DFS, and LRPFS methods. For the baseline method, all features are selected as feature subsets to represent the original dataset. For both S^{2}DFS and LRPFS, the top 100 features are selected as feature subsets. The experimental results correspond to Figure 6a–c, respectively. It is obvious that in Figure 6a,b, the inter-class distance in regions A and B is very small, which means that baseline and S^{2}DFS fail to distinguish different classes clearly, whereas, in Figure 6c, our LRPFS succeeds in enlarging the distance of different classes. Especially, when the coordinate scales are the same as in Figure 6c, the overall spatial structure of LRPFS remains consistent compared to S^{2}DFS, which further verifies that the potential relationship penalty term can explore the uniqueness of the samples to maximize the inter-class distance while preserving the spatial structure of the data and selecting more discriminative features.



(**a**) Baseline (**b**) S^{2}DFS (**c**) LRPFS

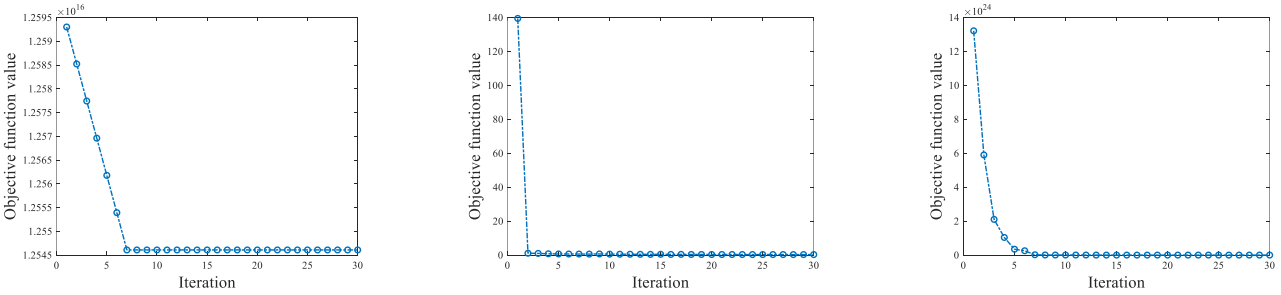
Figure 6. The 2-D demonstration of the COIL20 benchmark dataset.

4.6. LRPFS Experimental Performance

In this subsection, to assess the efficiency of LRPFS, convergence, and parameter sensitivity experiments are conducted on nine benchmark datasets (i.e., COIL20, Colon, Isolet, JAFFE, Lung_dis, nci9, PCMAC, PIE, and TOX_171). Additionally, the feature selection performance of LRPFS is evaluated on the Yale64 dataset.

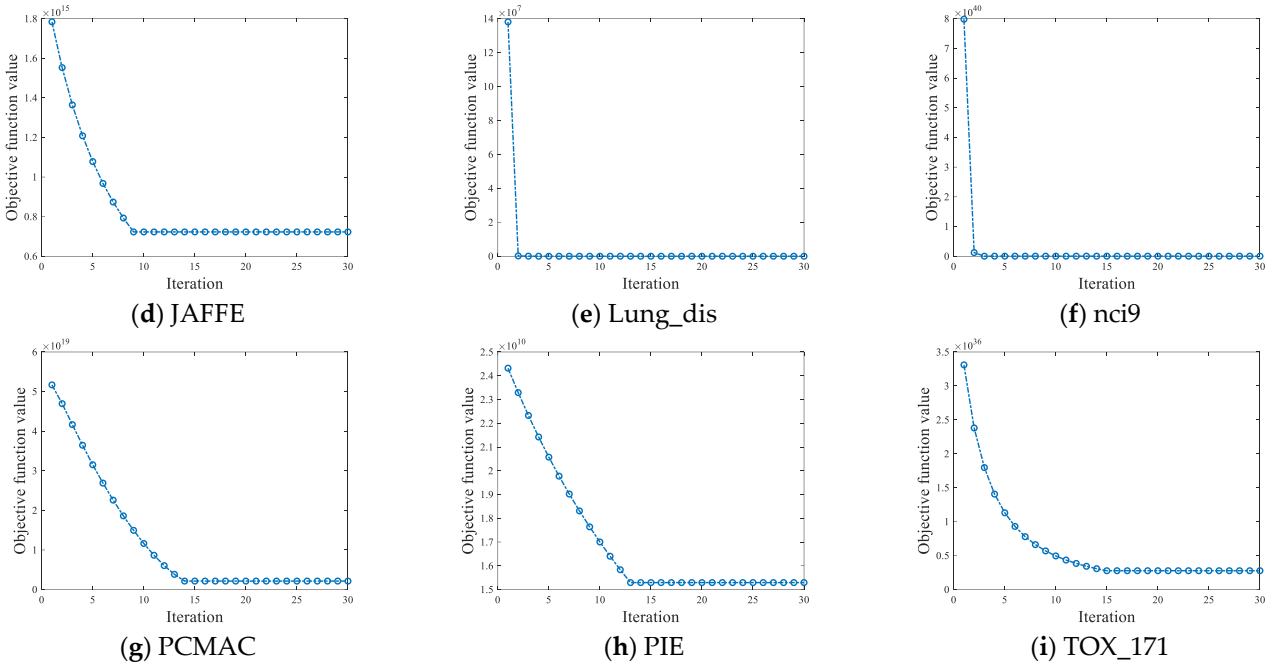
4.6.1. Convergence Analysis

To empirically demonstrate the convergence of LRPFS, convergence curves for nine datasets are depicted in Figure 7, where the horizontal axis represents the number of iterations and the vertical axis denotes the objective function values. From these plots, it is observed that the curves of the objective function exhibit significant and rapid variations, particularly in the Colon, Isolet, Lung_dis, and nci9 datasets, and convergence can be achieved within 15 iterations on all datasets. This observation serves as evidence that LRPFS achieves effective and stable convergence across all datasets, which further validates the correctness of the theoretical convergence proof.



(**a**) COIL20 (**b**) Colon (**c**) Isolet

Figure 7. *Cont.*



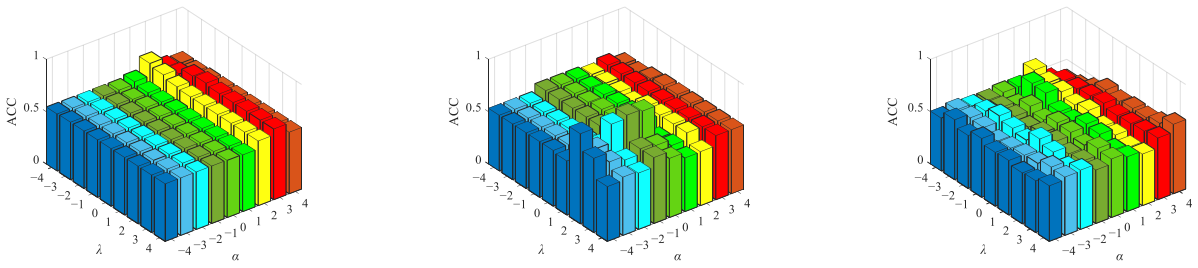
(**d**) JAFFE (**e**) Lung_dis (**f**) nci9

(**g**) PCMAC (**h**) PIE (**i**) TOX_171

Figure 7. Convergence curves of LRPFS on nine datasets.

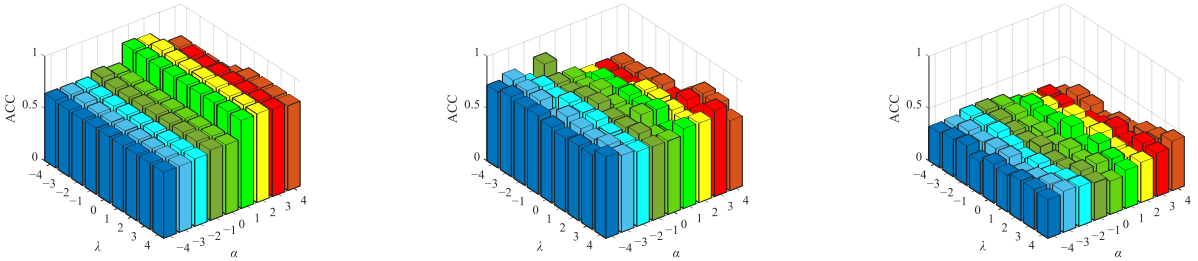
4.6.2. Parameter Sensitivity Experiment

LRPFS involves parameters k, σ, λ, and α. Among them, parameters k and σ are associated with constructing the sample attribute scores, which have an indirect and slight influence on the algorithm. Therefore, we mainly conduct parameter sensitivity experiments on parameters λ and α. We fix $k = 0$, $\sigma = 10$, and adjust the range of parameters λ and α in $\{10^{-4}, 10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}, 10^{4}\}$. The 3-D grids of ACC and NMI are displayed under different parameter values on test datasets as shown in Figures 8 and 9. As can be observed on most of the datasets, the values of ACC and NMI are positively correlated, and the clustering results are comparatively stable with the varying parameters. In particular, in these datasets where the number of samples is larger than the number of features. For example, COIL20 and PIE, λ has a minor impact on the clustering performance of LRPFS than α. However, on Colon, nci9, and PCMAC where the number of features is larger than the number of samples, LRPFS is more sensitive to λ. The reason may lie in that latent relationship penalty mining plays a more significant role in guiding feature selection when the number of sample features is larger. In a word, the parameters λ and α both play an indispensable role in LRPFS in that the feature selection mechanism of LRPFS can perform efficiently under the combined influence of latent relationship penalty and sparse constraints. Meanwhile, the experiment verifies that the parameters λ and α are suitable for 12 benchmark datasets in the range of $\{10^{-4}, 10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}, 10^{4}\}$, which provides a suitable parameter reference range for LRPFS in practical application.

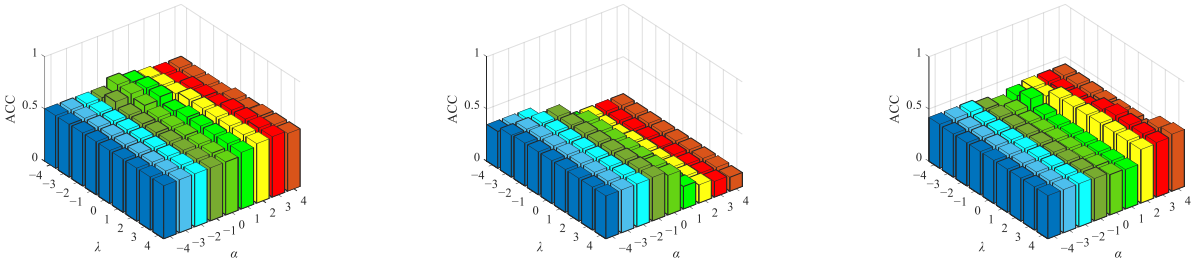


(**a**) COIL20 (**b**) Colon (**c**) Isolet

Figure 8. *Cont.*

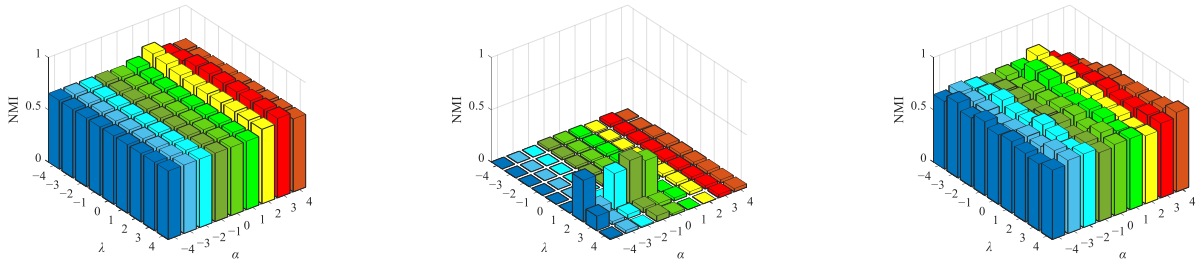


(**d**) JAFFE (**e**) Lung_dis (**f**) nci9

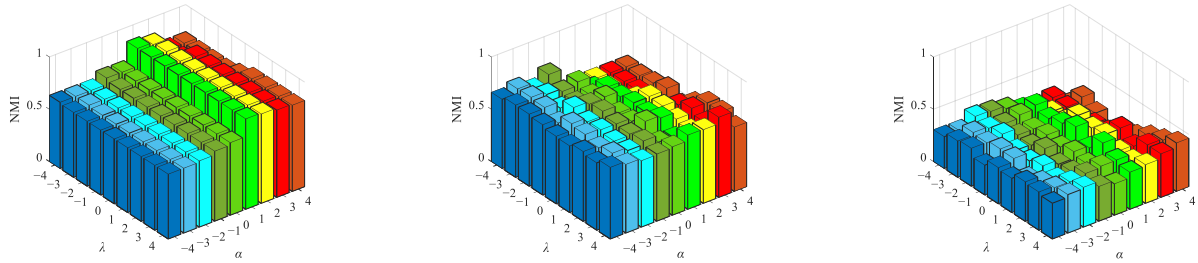


(**g**) PCMAC (**h**) PIE (**i**) TOX_171

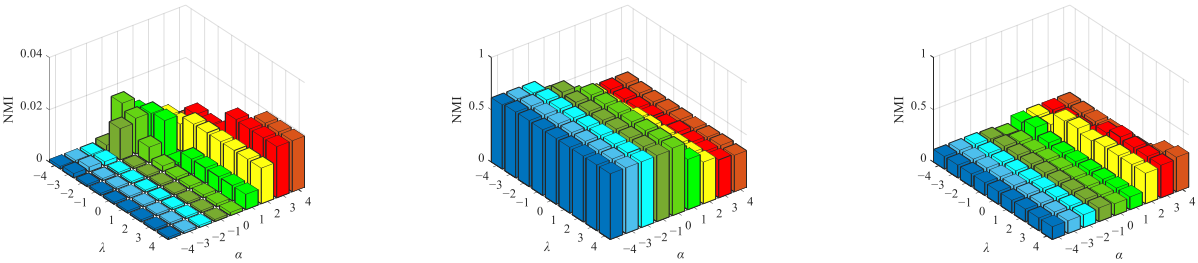
Figure 8. The ACC results of LRPFS on the nine datasets under different parameters varying in $\log_{10}$.



(**a**) COIL20 (**b**) Colon (**c**) Isolet



(**d**) JAFFE (**e**) Lung_dis (**f**) nci9



(**g**) PCMAC (**h**) PIE (**i**) TOX_171

Figure 9. The NMI results of LRPFS on the nine datasets under different parameters varying in $\log_{10}$.

4.6.3. The Effectiveness Evaluation of Feature Selection

On the Yale64 dataset, the selected features of LRPFS are visualized. In our experiments, two samples are randomly selected from the Yale64 dataset and select {0, 50, 100, 200, 500, 800, 1000, 2000} features from the selected samples under LRPFS, and the selected features are displayed in white pixels, which correspond to the images from left to right in the illustration in Figure 10, sequentially. It can be observed that when 50 features are selected, the selected features are mainly concentrated in hair, eyes, and nose, whereas the selected features appear in the mouth when the selected features are increased to 100 so that the features of hair, eyes, and nose are more discriminative than the mouth in Yale64 dataset. As the selected features gradually increase, the selected features are mainly divided into hair, eyes, glasses, nose, mouth, beard, etc., which is consistent with the perception of selected features for face recognition. Thus, it is demonstrated that LRPFS can effectively identify discriminative features and reasonably evaluate feature scores.

Figure 10. Results of two Yale64 samples with different numbers of selected features.

5. Conclusions

In this paper, a novel unsupervised feature selection with latent relationship penalty term, named LRPFS, is proposed, which takes into account the uniqueness of the samples and sufficiently exploits the attributes of preserving the data structure. LRPFS incorporates latent relationship penalty term into UFS, which provides a latent constraint on the feature transformation matrix and generates a pseudo-label matrix for feature selection. Additionally, the $\ell_{2,1}$-norm sparsity constraint is applied to the feature transformation matrix to enhance the computational efficiency of the algorithm significantly.

Comparative experiments are conducted between the LRPFS and 10 UFS methods. The comparison experiments covered various aspects, including clustering tasks, running speed, and noise experiments, utilizing datasets from different domains, such as images, text, Speech Signal, and biological data. The experimental results demonstrate that, compared to the comparison methods, LRPFS can effectively select discriminative features and reduce the interference of noise, especially on Colon dataset the ACC value of LRPFS is increased by 32.26% over the baseline, which further confirms the effectiveness of the LRPFS mechanism. The reason is that LRPFS can preserve pairwise relationships on the uniqueness score of the sample to explore the interconnection between individuals. At the same time, our proposed learning framework is beneficial for providing a theoretical basis for the realization of practical problems.

On the other hand, one limitation of LRPFS is that it requires the tuning of two parameters, which can be time-consuming. In Section 4, extensive experiments demonstrate the importance of a well-tuned set of parameters. Hence, our future work aims to develop a new mechanism that eliminates the need for parameter tuning or to design a novel optimization mechanism capable of simultaneously optimizing all variables. We also plan to apply this method to other fields such as remote sensing images and gene expression analysis in the future.

Author Contributions: Z.M., conceptualization, methodology, writing—review and editing, and validation; Y.H., methodology, software, data curation, and writing—original draft preparation; H.L., visualization and investigation; J.W., supervision and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Natural Science Foundation of Ningxia (Nos. 2020AAC03215 and 2022AAC03268), National Natural Science Foundation of China (No. 61462002), and Basic Scientific Research in Central Universities of North Minzu University (Nos. 2021KJCX09 and FWNX21).

Data Availability Statement: The data and code that support the findings of this study are available from the corresponding author (Z.M.) upon reasonable request.

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

References

1. Jain, A.; Zongker, D. Feature selection: Evaluation, application, and small sample performance. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 153–158. [CrossRef]
2. Nie, F.; Wang, Z.; Wang, R.; Li, X. Submanifold-preserving discriminant analysis with an auto-optimized graph. *IEEE Trans. Cybern.* **2020**, *50*, 3682–3695. [CrossRef] [PubMed]
3. Lee, D.; Seung, H. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [CrossRef] [PubMed]
4. Lipovetsky, S. PCA and SVD with nonnegative loadings. *Pattern Recognit.* **2009**, *42*, 68–76. [CrossRef]
5. Roweis, S.; Saul, L. Nonlinear dimensionality reduction by locally linear embedding. *Science* **2000**, *290*, 2323–2326. [CrossRef] [PubMed]
6. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 1–45. [CrossRef]
7. Saberi-Movahed, F.; Rostami, M.; Berahmand, K.; Karami, S.; Tiwari, P.; Oussalah, M.; Band, S. Dual regularized unsupervised feature selection based on matrix factorization and minimum redundancy with application in gene selection. *Knowl. Based Syst.* **2022**, *256*, 109884. [CrossRef]
8. Wang, Y.; Wang, J.; Tao, D. Neurodynamics-driven supervised feature selection. *Pattern Recogn.* **2023**, *136*, 109254. [CrossRef]
9. Plaza-del-Arco, F.; Molina-González, M.; Ureña-López, L.; Martín-Valdivia, M. Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. *Knowl. Based Syst.* **2022**, *258*, 109965. [CrossRef]
10. Ang, J.; Mirzal, A.; Haron, H.; Hamed, H. Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2016**, *13*, 971–989. [CrossRef]
11. Bhadra, T.; Bandyopadhyay, S. Supervised feature selection using integration of densest subgraph finding with floating forward-backward search. *Inf. Sci.* **2021**, *566*, 1–18. [CrossRef]
12. Wang, Y.; Wang, J.; Pal, N. Supervised feature selection via collaborative neurodynamic optimization. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. [CrossRef] [PubMed]
13. Han, Y.; Yang, Y.; Yan, Y.; Ma, Z.; Sebe, N.; Zhou, X. Semisupervised feature selection via spline regression for video semantic recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *26*, 252–264. [CrossRef] [PubMed]
14. Chen, X.; Chen, R.; Wu, Q.; Nie, F.; Yang, M.; Mao, R. Semisupervised feature selection via structured manifold learning. *IEEE Trans. Cybern.* **2022**, *52*, 5756–5766. [CrossRef] [PubMed]
15. Li, Z.; Tang, J. Unsupervised feature selection via nonnegative spectral analysis and redundancy control. *IEEE Trans. Image Process.* **2015**, *24*, 5343–5355. [CrossRef] [PubMed]
16. Zhu, P.; Hou, X.; Tang, K.; Liu, Y.; Zhao, Y.; Wang, Z. Unsupervised feature selection through combining graph learning and ℓ2,0-norm constraint. *Inf. Sci.* **2023**, *622*, 68–82. [CrossRef]
17. Shang, R.; Kong, J.; Zhang, W.; Feng, J.; Jiao, L.; Stolkin, R. Uncorrelated feature selection via sparse latent representation and extended OLSDA. *Pattern Recognit.* **2022**, *132*, 108966. [CrossRef]
18. Zhang, R.; Zhang, H.; Li, X.; Yang, S. Unsupervised feature selection with extended OLSDA via embedding nonnegative manifold structure. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 2274–2280. [CrossRef]
19. Zhao, Z.; Liu, H. Spectral feature selection for supervised and unsupervised learning. In Proceedings of the 24th Annual International Conference on Machine Learning, Corvalis, OR, USA, 20–24 June 2007; pp. 1151–1157. [CrossRef]
20. Cai, D.; Zhang, C.; He, X. Unsupervised feature selection for multi-cluster data. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010; pp. 333–342. [CrossRef]
21. Hou, C.; Nie, F.; Li, X.; Yi, D.; Wu, Y. Joint embedding learning and sparse regression: A framework for unsupervised feature selection. *IEEE Trans. Cybern.* **2014**, *44*, 2168–2267. [CrossRef]
22. He, X.; Cai, D.; Niyogi, P. Laplacian score for feature selection. In *Advances in Neural Information Processing Systems 18*; The MIT Press: Cambridge, MA, USA, 2005; pp. 507–514.
23. Shang, R.; Wang, W.; Stolkin, R.; Jiao, L. Subspace learning-based graph regularized feature selection. *Knowl. Based Syst.* **2016**, *112*, 152–165. [CrossRef]
24. Liu, Y.; Ye, D.; Li, W.; Wang, H.; Gao, Y. Robust neighborhood embedding for unsupervised feature selection. *Knowl. Based Syst.* **2020**, *193*, 105462. [CrossRef]
25. Nie, F.; Zhu, W.; Li, X. Unsupervised feature selection with structured graph optimization. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 1302–1308.

26. Li, X.; Zhang, H.; Zhang, R.; Liu, Y.; Nie, F. Generalized uncorrelated regression with adaptive graph for unsupervised feature selection. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 1587–1595. [CrossRef]
27. Chen, H.; Nie, F.; Wang, R.; Li, X. Unsupervised feature selection with flexible optimal graph. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. [CrossRef] [PubMed]
28. Tang, C.; Bian, M.; Liu, X.; Li, M.; Zhou, H.; Wang, P.; Yin, H. Unsupervised feature selection via latent representation learning and manifold regularization. *Neural Netw.* **2019**, *117*, 163–178. [CrossRef] [PubMed]
29. Shang, R.; Wang, L.; Shang, F.; Jiao, L.; Li, Y. Dual space latent representation learning for unsupervised feature selection. *Pattern Recognit.* **2021**, *114*, 107873. [CrossRef]
30. Samaria, F.; Harter, A. Parameterisation of a stochastic model for human face identification. In Proceedings of the 2nd IEEE Workshop on Applications of Computer Vision, Princeton, NJ, USA, 19–21 October 1994; pp. 138–142.
31. Yang, F.; Mao, K.; Lee, G.; Tang, W. Emphasizing minority class in LDA for feature subset selection on high-dimensional small-sized problems. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 88–101. [CrossRef]
32. Tao, H.; Hou, C.; Nie, F.; Jiao, Y.; Yi, D. Effective discriminative feature selection with nontrivial solution. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 796–808. [CrossRef]
33. Pang, T.; Nie, F.; Han, J.; Li, X. Efficient feature selection via ℓ2,0-norm constrained sparse regression. *IEEE Trans. Knowl. Data Eng.* **2019**, *31*, 880–893. [CrossRef]
34. Zhao, S.; Wang, M.; Ma, S.; Cui, Q. A feature selection method via relevant-redundant weight. *Expert Syst. Appl.* **2022**, *207*, 117923. [CrossRef]
35. Nouri-Moghaddam, B.; Ghazanfari, M.; Fathian, M. A novel multi-objective forest optimization algorithm for wrapper feature selection. *Expert Syst. Appl.* **2021**, *175*, 114737. [CrossRef]
36. Maldonado, S.; Weber, R. A wrapper method for feature selection using support vector machines. *Inf. Sci.* **2009**, *179*, 2208–2217. [CrossRef]
37. Shi, D.; Zhu, L.; Li, J.; Zhang, Z.; Chang, X. Unsupervised adaptive feature selection with binary hashing. *IEEE Trans. Image Process.* **2023**, *32*, 838–853. [CrossRef] [PubMed]
38. Nie, F.; Wang, Z.; Tian, L.; Wang, R.; Li, X. Subspace Sparse Discriminative Feature Selection. *IEEE Trans. Cybern.* **2022**, *52*, 4221–4233. [CrossRef] [PubMed]
39. Roffo, G.; Melzi, S.; Castellani, U.; Vinciarelli, A.; Cristani, M. Infinite feature selection: A graph-based feature filtering approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 4396–4410. [CrossRef] [PubMed]
40. Yang, Y.; Shen, H.; Ma, Z.; Huang, Z.; Zhou, X. ℓ2,1-norm regularized discriminative feature selection for unsupervised learning. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011; pp. 1589–1594. Available online: https://dl.acm.org/doi/10.5555/2283516.2283660 (accessed on 8 August 2022).
41. Xue, B.; Zhang, M.; Browne, W. Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach. *IEEE Trans. Cybern.* **2013**, *43*, 1656–1671. [CrossRef] [PubMed]
42. Ding, D.; Yang, X.; Xia, F.; Ma, T.; Liu, H.; Tang, C. Unsupervised feature selection via adaptive hypergraph regularized latent representation learning. *Neurocomputing* **2020**, *378*, 79–97. [CrossRef]
43. Shang, R.; Kong, J.; Feng, J.; Jiao, L. Feature selection via non-convex constraint and latent representation learning with Laplacian embedding. *Expert Syst. Appl.* **2022**, *208*, 118179. [CrossRef]
44. He, Z.; Xie, S.; Zdunek, R.; Zhou, G.; Cichocki, A. Symmetric nonnegative matrix factorization: Algorithms and applications to probabilistic clustering. *IEEE Trans. Neural Netw.* **2011**, *22*, 2117–2131. [CrossRef]
45. Shang, R.; Wang, W.; Stolkin, R.; Jiao, L. Non-negative spectral learning and sparse regression-based dual-graph regularized feature selection. *IEEE Trans. Cybern.* **2018**, *48*, 793–806. [CrossRef]
46. Huang, D.; Cai, X.; Wang, C. Unsupervised feature selection with multi-subspace randomization and collaboration. *Knowl. Based Syst.* **2019**, *182*, 104856. [CrossRef]
47. Xiao, J.; Zhu, X. Some properties and applications of Menger probabilistic inner product spaces. *Fuzzy Sets Syst.* **2022**, *451*, 398–416. [CrossRef]
48. Cai, D.; He, X.; Han, J. Locally consistent concept factorization for document clustering. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 902–913. [CrossRef]
49. Pan, V.; Soleymani, F.; Zhao, L. An efficient computation of generalized inverse of a matrix. *Appl. Math. Comput.* **2018**, *316*, 89–101. [CrossRef]
50. Luo, C.; Zheng, J.; Li, T.; Chen, H.; Huang, Y.; Peng, X. Orthogonally constrained matrix factorization for robust unsupervised feature selection with local preserving. *Inf. Sci.* **2022**, *586*, 662–675. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.