Unsupervised Feature Selection Algorithm Based on $L_{2,p}$-norm Feature Reconstruction

Wei Liu [1,☯,*], Qian Ning[1,☯], Guangwei Liu[2], Haonan Wang[3], Yixin Zhu[1], Miao Zhong[1]

1 College of Science, Liaoning Technical University, Fuxin, Liaoning, China,
2 College of Mines, Liaoning Technical University, Fuxin, Liaoning, China,
3 Johns Hopkins University, Maryland, USA

☯These authors contributed equally to this work.
* liuwei@lntu.edu.cn
0009-0007-1989-7283

Abstract

Traditional subspace feature selection methods typically rely on a fixed distance to compute residuals between the original space and the feature reconstruction space. However, this approach struggles to adapt to diverse datasets and often fails to handle noise and outliers effectively. In this paper, we propose a novel unsupervised feature selection method named unsupervised feature selection algorithm based on $l_{2,p}$-norm feature reconstruction (NFRFS). By employing a flexible norm to represent both the original space and the spatial distance of feature reconstruction, this method enhances adaptability and broadens its applicability through the adjustment of p. Additionally, adaptive graph learning is integrated into the feature selection process to preserve the local geometric structure of the data. Features exhibiting sparsity and low redundancy are selected through the regularization constraint of the inner product in the feature selection matrix. To demonstrate the effectiveness of the method, numerical studies were conducted on 12 benchmark microarray datasets. The results indicate that the method outperforms 8 unsupervised feature selection algorithms in terms of clustering results.

Author summary

This work was supported in part by the National Natural Science Foundation of China (Grant No.52374123), in part by the Basic Scientific Research Project of the Liaoning Provincial Department of Education (Project No. LJ212410147013, LJ212410147019), and in part by LiaoNing Revitalization Talents Program (Project No.XLYC2211085).

Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use) The authors declare no competing financial interests.

1. Introduction

In this era of information explosion, traditional data processing methods are facing unprecedented challenges due to the vast amount of data and the high dimensionality. The efficient and accurate processing of these rapidly growing high-dimensional datasets and extract key information has become a focal point of attention and research in fields such as data mining [1], pattern recognition [2], and machine learning [3]. Feature

selection algorithms extract representative features from raw data, not only achieving dimensionality reduction but also preserving the physical significance of the data [4].

Based on whether the data includes label information, feature selection can be divided into three types: supervised, semi-supervised, and unsupervised [5]. Since unsupervised feature selection does not rely on label information, it identifies features that best represent the characteristics of the data by analyzing its intrinsic structure, making it of significant research importance and value [6]. According to evaluation criteria, feature selection methods can be classified into filter, wrapper, and embedded methods [7]. Embedded methods combine the advantages of both filter and wrapper methods, integrating the feature selection process into model training to enhance the performance of algorithms [8].

Graph structure is crucial for feature selection. Authors in [9] introduced the Laplacian Score algorithm, which is based on the relationships between data points. This algorithm evaluates the importance of each feature by calculating its Laplacian score, reflecting its ability to preserve local information. A study in [10] introduced a feature selection algorithm based on latent representation learning and manifold regularization. By combining latent representation learning using non-negative matrix factorization and graph-based manifold regularization, it performs feature selection in a robust latent space, capturing the intrinsic structure of the data and reducing the negative impact of noise. This approach, which relies on a fixed similarity graph and depends on the sample similarity matrix, separates the construction of the graph from the learning of the feature selection matrix. As a result, it is susceptible to the influence of noise or outliers. Therefore, literature [11] proposed an unsupervised feature selection algorithm based on adaptive structure learning, which simultaneously conducts feature selection and data structure learning to better capture both global and local structures of the data. In literature [12], the unsupervised feature selection method known as Self-Weighted Adaptive Graph-based Minimum-Redundant Subspace Learning is mentioned. This approach integrates adaptive self-weighted graph learning, minimum redundancy, and sparsity constraints into a comprehensive framework. Manifold regularization can preserve the inherent geometric structure of the data, so unsupervised feature selection algorithms that incorporate manifold regularization typically achieve better performance. Although these algorithms have made some improvements, they need to enhance their handling of redundant information during feature selection.

In recent years, regularizers are often used to constrain the feature selection matrix in dealing with redundant information [13]. A study in [14] combined spectral analysis with l_1-norm regularization and proposed the multi-cluster feature selection algorithm . Authors in [15] unified feature selection and similarity matrix construction into a single framework and used an $l_{2,0}$-norm constraint on the feature selection matrix to achieve feature selection.In Reference [16], the authors introduced non-negative constraints and applied an $l_{2,p}$-norm to the matrix of feature transformations. This approach, in comparison to the $l_{2,1}$-norm , offers a more tractable optimization process. The variable p allows for a flexible trade-off between row sparsity and the convexity of the model, potentially enhancing the model performance. Meanwhile, Reference [17] incorporated the absolute values of inner product outcomes between the vectors of the feature selection matrix as a regularization component, thus fully accounting for feature interdependencies in the pursuit of a more independent selection of the subset of features. Influenced by the norms used in the feature selection matrix, regularizers can also be added to the loss function to prevent model overfitting and promote sparsity. Common choices include F-norm and $l_{2,1}$-norm, but both assume a fixed distance between the original samples and predicted labels, which limits their ability to flexibly adjust this distance based on the data's structure. Therefore, Reference [18] proposed a feature selection method based on the $l_{2,p}$-norm and sample constraints, applied in the

diagnosis of Alzheimer's disease.

Motivated by these considerations, we propose an efficient technique for feature selection, called unsupervised feature selection algorithm based on $l_{2,p}$-norm feature reconstruction (NFRFS). The distance between the original space and the reconstructed subspace can be flexibly adjusted through the $l_{2,p}$-norm. In this approach, graph embedding and feature selection interact to learn local structural information between data points. Inner product regularization is employed to select features that are both low in redundancy and sparse. The effectiveness of this method has been demonstrated on 12 benchmark datasets.

The main contributions of this paper are as follows:

- In the reconstruction error, a more flexible $l_{2,p}$-norm is used to measure the distance between the original samples and the reconstructed samples, and the value of p is adjusted to handle noise and outliers in the dataset.
- The feature selection matrix is sparsified by utilizing the inner product sparse regularization, selecting representative features with low redundancy.
- Comprehensive experiments on 12 benchmark datasets show that NFRFS is superior to several state-of-the-art feature selection methods. The experimental results validate the effectiveness and practicality of the model.

The rest of the paper is organized as follows. Section 2 explains some basic notions and definitions. In Section 3, we propose an optimization problem for feature selection and an iterative algorithm for solving the problem. In Section 4, various experimental results are analyzed. Conclusions are drawn in Section 5.

2. Related work

2.1. Notation and definition

In our work, for a data matrix $X \in \mathbb{R}^{n\times d}$ with n samples and d dimensions, the i-th row vector x_i and the j-th column vector x_j represent the i-th sample and the j-th feature, respectively. x_{ij} denotes the element at the i-th row and j-th column of X. $Tr(X)$ denotes the trace of matrix X. X^T is the transpose of X and the $l_{2,p}$-norm of X is defined as

$$\|X\|_{2,p} = \left(\sum_{i=1}^{d} \|x_i\|_2^p\right)^{1/p}$$

2.2. Feature Reconstruction in Subspace Using $l_{2,p}$-norm

Facing the challenges of high dimensionality and excessive redundant information, Wang [19] proposed an algorithm from the perspective of subspace learning. The algorithm extracts a low-dimensional subspace from high-dimensional data, which can represent the main information of the original feature space while removing redundancy and noise.

$$\begin{aligned} &argmin\, \|X - XWH\|_F^2 \\ &s.t. W.H \geq 0, W^T W = I_l \end{aligned} \tag{1}$$

Where $H \in \mathbb{R}^{l\times d}$ is the coefficient matrix used for reconstruction, which maps the learned subspace to the original space. l represents the number of selected features, and $I \in \mathbb{R}^{l\times l}$ is an identity matrix. $W \in \mathbb{R}^{d\times l}$ is the feature selection matrix, constrained

with orthogonality W to ensure that there is at most one non-zero value per row and column. Additionally, non-negative constraints are imposed on W to preserve its real-world physical meaning [20].

In existing subspace feature selection methods, the F-norm is typically used to measure the distance between the original data space and the reconstructed subspace [21]. However, for some datasets, using a fixed distance metric does not result in the optimal feature subset. Therefore, this paper uses an adaptive distance metric to effectively improve model performance, choosing the $l_{2,p}$-norm to constrain the distance between the original space and the reconstructed subspace. The $l_{2,p}$-norm allows for flexible adjustment of the size of parameter p , making the choice of the p parameter most favorable for feature selection. The model's application of the $l_{2,p}$-norm can be expressed as:

$$\begin{aligned} &argmin \, \|X - XWH\|_{2,p}^{p} \\ &s.t. W.H \geq 0, W^T W = I_l \end{aligned} \tag{2}$$

Choosing different p values for the significantly impacts the model's performance. **Fig 1** presents 3D surface plots for three different norms, showing that both F-norm and $l_{2,1}$-norm tend to optimize more towards the origin. However, during the optimization process, the $l_{2,1/2}$-norm is more inclined towards the coordinate axes. Therefore, using the $l_{2,1/2}$-norm to constrain the model effectively eliminates redundant features and selects more discriminative features.

Fig 1. 3D surface plot of F-norm, $l_{2,1}$-norm and $l_{2,1/2}$-norm regularization term.

2.3. Sparsity regularization term of the feature selection matrix

Choosing the appropriate constraints for the feature selection matrix can effectively enhance model sparsity and reduce redundancy [22]. For example, l_1-norm regularization may lead to underfitting in high-dimensional data, while $l_{2,0}$-norm regularization provides good sparsity but is difficult to solve, making it difficult to find an optimal solution. The $l_{2,1}$-norm regularization offers both good sparsity and optimization performance, making it widely used.Recent research has shown that inner product regularization can replace the $l_{2,1}$-norm. The inner product of the feature vectors gradually approaches the minimum during the optimization process, allowing the removal of features with high similarity, thereby reducing redundancy among features and improving clustering performance. The inner product regularization term for the feature selection matrix W is defined as:

$$\sum_{i,j=1,i\neq j}^{d} < w^i, w^j > \; = \sum_{i,j=1,i\neq j}^{d} w^i {w^j}^T = Tr(1_{d\times d} W W^T) - Tr(W W^T) \tag{3}$$

2.4. Adaptive Graph Learning

In the design of existing models, the global structure and constraints on the feature selection matrix have been thoroughly considered. To further enhance the generalizability of the model, local structural information is incorporated, and the most effective way to achieve this is by introducing graph learning [23]. Graph structures can effectively preserve the local neighborhood information of data, and when mapping from the original feature space to a lower-dimensional feature space, they can maintain the geometric structure of the samples [24]. Simply put, if two sample points are close to each other in the original space, they should also remain close in the feature-selected

projection space. Mathematically, this can be expressed as:

$$\min_{S} \sum_{i,j=1}^{n} ||W^T x_i - W^T x_j||_2^2 s_{ij} = Tr(W^T X^T L X W) \tag{4}$$

However, real-world data is often affected by noise, which can cause the k-nearest neighbor graph constructed by the aforementioned method to be susceptible to errors. To better preserve the manifold structure of the data, an adaptively updated similarity matrix is introduced. This allows low-dimensional embedding learning and manifold learning to be completed in a single step, thereby improving the effectiveness of feature selection.

$$\begin{aligned} \min_{S} Tr(W^T X^T L X W) + \gamma ||S||_F^2 \\ s.t. \sum\nolimits_{j=1}^{n} s_{ij} = 1, s_{ij} \geq 0 \end{aligned} \tag{5}$$

Where L is the Laplacian matrix, and $L = D - S$. $S \in R^{n \times n}$represents the similarity matrix of the samples, where the element s_{ij} denotes the similarity between the sample points x_i and x_j . $D \in R^{n \times n}$is a diagonal matrix and its diagonal elements are defined as:

$$d_{ii} = \sum\nolimits_{j=1}^{n} s_{ij} \tag{6}$$

3. Unsupervised Feature Selection Based on $l_{2,p}$-norm Feature Reconstruction

3.1. Model Construction

In constructing the subspace, the $l_{2,p}$-norm is used to flexibly measure the distance between the original samples and the reconstructed samples. The adaptive graph embedding learning takes into account the similarity relationships between samples, preserving the local geometric structure of the data. In addition, by applying an inner product constraint on the feature selection matrix, a more sparse solution can be obtained, which helps to select a representative subset of features. The final objective function is expressed as:

$$\begin{aligned} \min ||X - XWH||_{2,p}^p + \alpha Tr(W^T X^T L X W) + \\ \beta(Tr(1_{d \times d} W W^T) - Tr(W W^T)) + \gamma ||S||_F^2 \\ s.t. W \geq 0, H \geq 0, W^T W = I, \sum_{j=1}^{n} s_{ij} = 1, s_{ij} \geq 0 \end{aligned} \tag{7}$$

Where α ,β are regularization parameters, and γ is a coefficient that can be determined during the optimization process.

3.2. Model Solution

The objective function in eq (6) includes three variables, W, H , and S. To improve computational efficiency, this paper employs an alternate optimization method to optimize the objective function, that is, by fixing two variables each time and optimizing the other variable.

Define two Lagrange multipliers, θ and μ, to ensure the non-negativity of the matrices W and H. The resulting Lagrangian function is as follows:

$$\begin{aligned} L(W,H) =& ||X - XWH||_{2,p}^{p} + \alpha Tr(W^T X^T L X W) \\ &+ \beta(Tr(1_{d\times d} W W^T) - Tr(WW^T)) \\ &+ \frac{\lambda}{2}(W^T W - I) + Tr(\theta W^T) + Tr(\mu H^T) \end{aligned} \tag{8}$$

Define a diagonal matrix U, with diagonal elements being $u_{ii} = \frac{p}{2||(X-XWH)_i||_2^{2-p}}$.

1. Fix H , S , and Update W: By taking the partial derivative of eq 8 with respect to W, the following formula can be obtained:

$$\begin{aligned} \frac{\partial L}{\partial W} =& -X^T U X H^T + X^T U X W H H^T + \alpha X^T L X W + \beta(W W^T W - W) \\ &+ \lambda(1_{d\times d} W - W) + \theta \end{aligned} \tag{9}$$

By using the Karush–Kuhn–Tucker (KKT) conditions $\theta_{ij} W_{ij} = 0$, the obtained formula is as follows:

$$\begin{aligned} &(-X^T U X H^T + X^T U X W H H^T + \alpha X^T L X W \\ &+\beta(W W^T W - W) + \lambda(1_{d\times d} W - W))_{ij} W_{ij} = 0 \end{aligned} \tag{10}$$

Thus, the update rule for W is as follows:

$$W_{ij} \leftarrow W_{ij} \frac{\left[X U X^T H^T + \alpha X^T S X W + \beta W + \lambda W\right]_{ij}}{\left[X U X^T W H H^T + \alpha X^T D X W + \beta W W^T W + \lambda 1_{d\times d} W\right]_{ij}} \tag{11}$$

2.Fix W, S, and Update H: By taking the partial derivative of eq (8) with respect to H, the following formula can be obtained:

$$\frac{\partial L}{\partial H} = -W^T X^T U X + W^T X^T U X W H^T + \mu \tag{12}$$

By using the Karush–Kuhn–Tucker (KKT) conditions $\mu_{ij} H_{ij} = 0$, the obtained formula is as follows:

$$(-W^T X^T U X + W^T X^T U X W H^T)_{ij} H_{ij} = 0 \tag{13}$$

Thus, the update rule for H is as follows:

$$H_{ij} \leftarrow H_{ij} \frac{\left[W^T X^T U X\right]_{ij}}{\left[W^T X^T U X W H^T\right]_{ij}} \tag{14}$$

3.Fix W, H, and Update S: By taking the partial derivative of eq (8) with respect to S, the following formula can be obtained:

$$\begin{aligned} \min_{S} \sum_{i,j=1}^{n} &(\alpha ||W^T x_i - W^T x_j||_2^2 s_{ij} + \gamma s_{ij}^2) \\ &s.t. \sum\nolimits_{j=1}^{n} s_{ij} = 1, s_{ij} \geq 0 \end{aligned} \tag{15}$$

Denote $d_{ij} = ||W^T x_i - W^T x_j||_2^2$,so we can transform eq (14) to a vector form as

$$\begin{aligned} &\min_{s_i^T \mathbf{1}=1, s_i \geq 0} ||s_i + \frac{\alpha}{2\gamma} d_i||_2^2 \\ &s.t. \sum\nolimits_{j=1}^{n} s_{ij} = 1, s_{ij} \geq 0 \end{aligned} \tag{16}$$

Introduce Lagrange multipliers ω and μ to construct the Lagrangian function:

$$\mathcal{L}(s_i, \omega, \varphi_i) = ||s_i + \frac{\alpha}{2\gamma} d_i||_2^2 - \omega(s_i^T \mathbf{1} - 1) - \varphi_i^T s_i \tag{17}$$

According to the KKT conditions, the optimal solution is obtained.

$$s_{ij} = (-\frac{\alpha d_{ij}}{2\varepsilon_i} + \omega)_+ \tag{18}$$

In unsupervised feature selection algorithms, preserving the local geometric manifold structure of the data tends to be more effective than preserving the global structure. Therefore, only neighboring points k are considered to construct the similarity matrix. In the experiments of this paper, when the local structure of the data is maintained, five neighboring points are uniformly selected. The optimal solution can be represented as the average of all γ_i [25]. Assuming $d_{i1}, d_{i2}, \ldots, d_{in}$ is sorted from smallest to largest and satisfies the condition.Because s_i satisfies $s_{ik} > 0 \geq s_{i,k+1}$, we have then we have:

$$\begin{cases} s_{ik} > 0 \Rightarrow -\frac{\alpha d_{ik}}{2\varepsilon_i} + \omega > 0 \\ s_{i,k+1} \leq 0 \Rightarrow -\frac{\alpha d_{i,k+1}}{2\varepsilon_i} + \omega \leq 0. \end{cases} \tag{19}$$

According to eq (18) and the constraint $s_i^T 1 = 1$ we have

$$\sum_{j=1}^{k}(-\frac{\alpha d_{ij}}{2\gamma_i} + \omega) = 1 \Rightarrow \omega = \frac{1}{k} + \frac{\alpha}{2k\gamma_i}\sum_{i=1}^{k} d_{ij} \tag{20}$$

By substituting the value of ω in eq (19) into eq (18), we have

$$\frac{\alpha}{2}\left(kd_{ik} - \sum_{j=1}^{k} d_{ij}\right) < \gamma_i \leq \frac{\alpha}{2}\left(kd_{i,k+1} - \sum_{j=1}^{k} d_{ij}\right) \tag{21}$$

Therefore, in order to obtain an optimal solution of s_i that has exact k nonzero values, we set γ_i to be

$$\gamma_i = \frac{\alpha}{2}\left(kd_{i,k+1} - \sum_{j=1}^{k} d_{ij}\right) \tag{22}$$

and then the overall γ is set to the mean of γ_i as

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}(\frac{\alpha k}{2} d_{i,k+1} - \frac{\alpha}{2}\sum_{j=1}^{k} d_{ij}) \tag{23}$$

Finally, substitute eq (22) into eq (17), and consider only k neighboring points to construct the similarity matrix. In summary, the solution can be obtained by solving as

$$s_{ij} = \begin{cases} \frac{d_{i,k+1} - d_{ij}}{\alpha k d_{i,k+1} - \alpha\sum_{j=1}^{k} d_{ij}}, j \leq k \\ 0, j \geq k+1 \end{cases} \tag{24}$$

Algorithm 1: NFRFS

Input: Date matrix X, the coefficient α, β, and p, the dimension of the subspace l, select feature number m

Output: Calculate and sort $\|w_i\|_2$ in the descending order, then select the top m ranked features as the results of feature selection.

Initialize $t = 0$, W, H, maxIter
while $t \leq$ maxIter do
update W by eq (11)
update H by eq (14)
update S by eq (24)
end while

3.3. Convergence of Algorithm

In this section the convergence analysis of the proposed NFRFS algorithm is presented. Algorithm 1 solves eq (7) with iteratively updating of W, H, and S. In order to demonstrate this convergence, Lemma 1 is utilized subsequently, which is proposed and proven in [26].

Definition 1. If there is a function $J(h, h')$ that makes $C(h)$ satisfy the following conditions:

$$J(h, h') \geq C(h), J(h, h) = C(h) \tag{25}$$

Then C is non-increasing under the following update formula:

$$h^{(t+1)} = \arg\min_{h} J\left(h, h^{(t)}\right) \tag{26}$$

Where $J(h, h')$ is an auxiliary function of $C(h)$.

Proof. $C\left(h^{(t+1)}\right) \leq J\left(h^{(t+1)}, h^{(t)}\right) \leq J\left(h^{(t)}, h^{(t)}\right) = C\left(h^{(t)}\right)$

Since the monotonicity of the objective function eq (7) under the update rule of the variable H needs to be proved, the terms relating to the variable H in eq (7) are retained, and the following function is obtained:

$$C(H) = Tr(X - XWH)^T U(X - XWH) \tag{27}$$

By taking the first-order and the second-order partial derivatives of $C(H)$ with respect to H, the following formulas can be obtained:

$$C'_{ij} = \left[\frac{\partial C}{\partial H}\right]_{ij} = \left[-2W^T X^T UX + 2W^T X^T UXWH^T\right]_{ij} \tag{28}$$

$$C''_{ij} = 2\left[W^T X^T UXW\right]_{ij} \tag{29}$$

Lemma 1. Giving the auxiliary functions of C_{ij}, and the form is as follows:

$$\begin{aligned} J(H_{ij}, H_{ij}^{(t)}) =& C_{ij}(H_{ij}^{(t)}) + C'_{ij}(H_{ij}^{(t)})(H_{ij} - H_{ij}^{(t)}) \\ &+ \frac{[W^T X^T UXWH^{(t)}]_{ij}}{H_{ij}^{(t)}}(H_{ij} - H_{ij}^{(t)})^2 \end{aligned} \tag{30}$$

Denoting the Taylor expansion of $C_{ij}(H_{ij})$ as follows:

$$\begin{aligned} C_{ij}(H_{ij}) =& C_{ij}(H_{ij}^{(t)}) + C'_{ij}(H_{ij}^{(t)})(H_{ij} - H_{ij}^{(t)}) \\ &+ \left\{\left[W^T X^T UXWH\right]_{ii}(H_{ij} - H_{ij}^{(t)})^2\right\} \end{aligned} \tag{31}$$

It can be seen from formulas eq (28) and eq (29) that $J(H_{ij}, H_{ij}^{(t)}) \geq C_{ij}(H_{ij})$ is equivalent to:

$$\frac{[W^T X^T U X W H^{(t)}]_{ij}}{H_{ij}^{(t)}} \geq \left[W^T X^T U X W\right]_{ii} \tag{32}$$

It is obvious that the following formula holds:

$$[W^T X^T U X W H^{(t)}]_{ij} = \sum_{b=1}^{l} \left[W^T X^T U X W\right]_{ib} H_{bj}^{(t)} \geq \left[W^T X^T U X W\right]_{ii} H_{ij}^{(t)} \tag{33}$$

So inequality eq (32) holds, that is, $J(H_{ij}, H_{ij}^{(t)}) \geq C_{ij}(H_{ij})$ holds. Obviously, the equation $J(H_{ij}, H_{ij}) = C_{ij}(H_{ij})$ also holds.

Then, we will prove that the update rule of variable H satisfies the update formula eq (26) that makes C_{ij} non-increasing.

By substituting $J(H_{ij}, H_{ij}^{(t)})$ in eq (30) into eq (26), the following formula can be obtained:

$$H_{ij}^{(t+1)} = H_{ij}^{(t)} - H_{ij}^{(t)} \frac{C'_{ij}(H_{ij}^{(t)})}{2[W^T X^T U X W H^{(t)}]_{ij}} \tag{34}$$

Substituting eq (28) into eq (34)gives the following expression:

$$H_{ij}^{(t+1)} = H_{ij}^{(t)} \frac{W^T X^T U X}{[W^T X^T U X W H^{(t)}]_{ij}} \tag{35}$$

It can be seen that eq (35) is the update rule of variable H, so C_{ij} is non-increasing under the update rule eq (12). Consequently, the non-increasing of updating rule of variable H can be concluded. Similar of variable H, it can be proved that updating rule of variable W is non-increasing.

3.4. Computational Complexity Analysis

This section analyzes the time complexity of the NFRFS algorithm. Let n represent the total number of samples, d represent the number of sample features, c represent the number of sample classes in the dataset, l represent the dimension of the subspace, and t represent the maximum number of iterations. According to eq (11), the time complexity for updating matrix W is $O(n^2 d + nd + dl)$, according eq (14), the time complexity for updating matrix H is $O(nd + d^2 n)$, and according to eq(24), the time complexity for updating matrix S is $O(n^2)$. Since the dimension of the subspace is less than the number of samples and the dimensionality of the data, the overall time complexity of the algorithm is $O(t(n^2 d + d^2 n))$.

4. Experiment

4.1 Datasets

This paper will test the performance of the feature selection model on 12 datasets. These include facial datasets (Yale, ORL, JAFFE, warpPIE10P, warpAR10P), biological datasets (lung, TOX-171, IsoIet, Lung_small), object dataset (COIL20), and digital image dataset (binalpha). Table 1 provides a detailed introduction to these datasets.

Table 1. Detail introduction to datasets.

Dataset	samples	features	classes
Yale	165	1024	15
lung	203	3312	5
COIL20	1440	1024	20
warpPIE10P	210	2420	10
warpAR10P	130	2400	10
ORL	400	1024	40
JAFFE	213	256	10
ATT40	400	1024	40
TOX-171	171	5748	4
Isolet	1560	617	26
binalpha	1404	320	36
Lung_small	73	325	7

4.2. Compared method

To validate the effectiveness of the proposed approach in UFS, the proposed approach is compared with a baseline method that performs clustering with all the original features and seven other representative existing UFS methods.

LS [9]: Use the local geometric information of data to select features, and calculate the score of each feature separately.

MCFS [14]: A multi-cluster feature selection for data, which first conducts spectral analysis and then selects features through sparse regression.

SPFS [27]: Integrates local and global structures into a unified framework and formulates the framework as a form of high-order matrix decomposition.

VSCDFS [28]: Selects a representative feature subset using variance-covariance information of the feature space.

AUFS [29]: Performs unsupervised feature selection by minimizing an objective function that includes self-representation reconstruction error, $l_{2,p}$-norm regularization term, and robust graph regularization term.

GLUFS [30]: Integrates the construction of similarity matrices and feature selection into a unified framework, introducing a sparse learning strategy with $l_{2,0}$-norm constraint.

HSL [31]: Simultaneously learns the projection matrix, first-order similarity information, and higher-order similarity information within a unified framework.

4.3. Evaluation Methodology

To verify the clustering performance of the algorithm, this article adopts two evaluation indicators, namely cluster accuracy (ACC) [32] and Normalized Mutual Information (NMI) [33]. Both values are within the range of [0, 1]. The higher the values of ACC and NMI, the better the clustering effect and the more representative the selected feature subset.

1. ACC

$$ACC = \frac{1}{n}\sum_{i=1}^{n}\delta(w_i, map(m_i)) \tag{36}$$

Where w_i denotes the ideal label, m_i represents the predicted label, $map(\cdot)$ denotes the optimal mapping function, and $\delta(\cdot)$ represents the indicator function. If $a = b$, then $\delta(a, b) = 1$, otherwise $\delta(a, b) = 0$.

2. NMI

$$NMI = \frac{I(w, m)}{\sqrt{H(w)H(m)}} \tag{37}$$

Where w_i denotes the ideal label, m_i represents the predicted label, $I(\cdot)$ denotes the the mutual information, and $H(\cdot)$ represents the information entropy .

4.4. Experimental settings

Before conducting the experiment, it is necessary to specify several parameter values for the NFRFS method proposed in this paper and other comparison methods. The regularization parameters of all algorithms are set within the range $\{10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2, 10^4, 10^6\}$ using the grid search method, and the best results of all algorithms are recorded. The number of selected features in the dataset is $\{20, 40, 60, \cdots, 160, 180, 200\}$. The dimension of the subspace l is set to $\{\frac{d}{3}, \frac{d}{2}, \frac{2d}{3}\}$. The K-means algorithm is used to cluster the data points formed by the selected features in order to evaluate different methods. Considering that the K-means algorithm is sensitive to initialization, the experiment repeats the K-means algorithm 20 times to eliminate the influence of initial points on the clustering effect.

4.5. Analysis of Experimental Results

Table 2 and table 3 respectively present the best clustering accuracy (ACC) and normalized mutual information (NMI) scores achieved by NFRFS and other comparative algorithms on 12 datasets, along with the corresponding standard deviations (STD). The highest values among different algorithms for the same dataset are highlighted in bold black font in the tables. The specific results are shown in the tables.

From table 2 and table 3, it can be observed that NFRFS achieves the best ACC and NMI scores on all datasets except COIL20. However, on the COIL20 dataset, NFRFS still outperforms all other comparative algorithms except HSL. In terms of the ACC values from the experimental results, the improvement of NFRFS on the warpPIE10P dataset is the most significant, with an increase of 4.26% over the second-best algorithm and 32.5% over the method that uses all features for clustering. Compared to the SPFS algorithm, the proposed NFRFS algorithm achieves better performance, especially with improvements of over 9% in ACC on the lung, warpPIE10P, and TOX-171 datasets, and over 12% in NMI on the Yale and TOX-171 datasets. This is because NFRFS not only preserves local structural information by adaptively learning the manifold structure but also effectively adjusts the distance between samples in the original and reconstructed spaces through the subspace learning framework of matrix decomposition. Additionally, the inner product regularization term imposed on the feature selection matrix significantly enhances the model's performance.

Table 2. Best ACC for different methods on different datasets (mean±std%).

Dataset	baseline	LS	MCFS	SPFS	VSCDFS	AUFS	GLUFS	HSL	NFRFS
Yale	38.64±3.61	36.94±2.17	40.61±3.33	39.36±2.36	37.12±2.28	46.70±3.14	45.97±3.90	49.70±3.43	**50.39±3.18**
lung	72.46±10.20	57.91±7.48	72.02±7.57	75.37±6.71	60.47±7.43	79.68±4.52	81.55±10.03	79.85±3.94	**84.04±4.14**
COIL20	59.17±3.98	53.89±3.34	60.92±3.52	68.27±2.24	60.56±4.41	70.26±1.82	68.68±2.86	**71.05±3.32**	69.94±2.25
warpPIE10P	26.24±2.03	44.71±3.00	28.86±2.56	48.52±3.27	25.79±1.21	54.48±1.26	37.21±2.76	45.90±0.99	**58.74±3.96**
warpAR10P	23.58±3.94	33.08±3.08	30.54±3.23	48.92±2.86	26.88±3.39	40.92±1.72	40.23±2.64	43.15±3.30	**51.62±3.51**
ORL	51.79±3.37	40.09±2.20	52.71±3.02	51.01±2.26	48.21±2.97	53.70±1.92	54.43±2.90	54.89±2.24	**58.66±2.37**
JAFFE	67.28±6.18	66.38±6.15	70.26±6.35	78.22±4.30	69.86±7.89	80.80±3.52	81.20±3.77	80.00±4.49	**83.43±1.90**
ATT40	51.28±3.86	49.16±2.80	53.19±3.49	55.48±3.14	49.56±3.00	58.66±2.75	55.85±2.89	55.18±2.65	**59.31±2.37**
TOX-171	43.65±3.14	41.29±2.48	46.90±3.31	47.78±1.78	43.04±2.51	43.77±1.32	57.60±0.70	53.95±2.98	**58.89±1.30**
Isolet	57.35±3.44	58.09±2.72	57.73±2.80	67.27±2.25	61.52±2.97	70.72±1.58	65.94±2.24	60.32±2.38	**71.83±3.31**
binalpha	41.17±1.87	39.56±1.51	41.66±1.44	42.34±1.75	39.70±2.08	43.82±2.12	43.39±1.35	43.31±1.72	**45.08±1.36**
lung_small	67.81±7.42	67.74±5.56	69.93±6.83	79.59±5.41	65.00±6.49	82.26±5.47	80.00±5.09	78.15±5.29	**83.97±5.55**

Table 3. Best NMI for different methods on different datasets (mean±std%).

Dataset	baseline	LS	MCFS	SPFS	VSCDFS	AUFS	GLUFS	HSL	NFRFS
Yale	46.48±2.88	44.15±1.67	49.23±2.79	46.17±2.27	44.35±1.92	56.34±2.87	51.37±1.67	57.39±7.79	**59.10±2.15**
lung	60.37±5.38	47.04±3.14	59.87±5.85	65.04±0.88	53.47±5.36	64.74±1.30	66.49±4.01	64.36±1.89	**66.78±3.16**
COIL20	75.58±1.64	70.53±1.45	74.25±2.04	79.73±1.35	75.13±1.31	79.76±1.02	78.65±1.29	**80.91±1.45**	80.04±1.04
warpPIE10P	25.36±3.18	50.09±3.36	30.74±3.57	56.41±2.54	22.62±2.18	59.87±2.10	39.78±3.42	48.19±2.30	**63.10±1.47**
warpAR10P	20.28±5.42	35.23±2.91	29.47±2.78	51.68±1.67	22.05±3.52	43.57±1.67	42.49±3.60	44.77±2.39	**53.18±2.71**
ORL	74.26±1.82	63.93±1.46	74.81±1.71	72.59±1.04	71.68±1.32	74.76±1.20	75.82±1.19	75.59±1.59	**78.19±1.05**
JAFFE	73.14±3.55	71.27±3.52	75.77±3.26	82.39±1.68	76.68±3.62	82.32±2.04	82.75±1.78	83.26±2.65	**84.67±1.50**
ATT40	74.02±1.79	72.23±1.39	75.30±1.56	75.94±1.57	72.41±1.63	78.13±0.95	76.32±1.27	76.13±0.61	**78.53±1.31**
TOX-171	15.87±4.44	16.44±1.33	22.69±4.12	23.48±1.15	12.19±1.53	14.97±0.96	30.01±0.77	34.52±3.69	**36.96±1.20**
Isolet	75.07±1.71	74.14±1.14	75.29±1.18	77.98±0.87	75.86±1.29	80.01±1.09	78.69±1.07	73.99±0.69	**81.58±0.89**
binalpha	57.71±0.87	55.85±1.24	58.29±0.66	57.58±0.67	55.25±1.03	59.05±0.85	58.53±0.76	59.19±0.51	**59.68±0.93**
lung_small	65.15±7.16	64.70±4.42	66.62±5.86	74.48±3.34	62.72±4.66	76.61±4.23	74.52±3.14	74.30±3.39	**78.36±4.33**

Fig 2 and **3** show the relationship between the number of selected features and the best accuracy and normalized mutual information, respectively, for nine unsupervised clustering methods on 12 datasets. The horizontal axis represents the number of selected features. On the binalpha dataset, NFRFS clustering performance is not very good when the number of selected features is small, but as the number of selected features increases, NFRFS performance improves and ultimately surpasses the GLUFS and AUFS algorithms. Compared to direct K-means clustering, NFRFS achieves higher performance even when a smaller number of features are selected. On the Yale, warpAR10P, ORL, JAFFE, TOX-171, and lung_small datasets, NFRFS clustering performance corresponding to all numbers of selected features is higher than that of other comparative algorithms, demonstrating that NFRFS feature selection is effective and feasible.

Fig 2. ACC with different number of features on different datasets.

Fig 3. NMI with different number of features on different datasets.

4.6. Parameter sensitivity

To consider the impact of parameter variations on the model, this section conducts a parameter sensitivity experiment analysis. From the objective function, there are four regularization parameters α, β, γ, and p, but in the experiment, only three regularization parameters α, β, and p need to be adjusted. First, the sensitivity of parameters α and β is discussed, and the clustering accuracy (ACC) and normalized mutual information (NMI) under parameter combinations within a certain range $\{10^{-6}, 10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}, 10^{6}\}$ are made. The three-dimensional histograms of ACC and NMI values on twelve datasets are shown in **Fig 4** and **5** ,respectively. Next, the effect of the parameter $p = \{0.01, 0.05, 0.1, 0.5, 1\}$ on the model performance is discussed separately, and the results are plotted in a line graph.

From the experimental results, parameters α and β show relatively stable changes in the clustering performance of most datasets, but they are more sensitive on the lung, warpPIE10P, warpAR10P, and TOX-171 datasets. This is because the number of features in these datasets is much larger than the number of samples, but the NFRFS algorithm can still achieve good clustering results under certain combinations of parameters. To further analyze the impact of parameters on the experimental results, the parameter $p = \{0.01, 0.05, 0.1, 0.5, 1\}$ is selected for analysis, and the results are shown in **Fig 6** and **7**. The default parameters α and β are set to fixed values of 1. For the lung dataset, different values of p result in significant differences in clustering performance. The ACC and NMI of the Yale, ATT40, and Isolate datasets are maximized at $p = 0.01$, while the ACC and NMI of the lung, warpPIE10P, JAFFE, and binalpha datasets are maximized at $p = 1$. This proves that for different datasets, the distances in the original space and the reconstructed space should be flexibly adjusted to find the optimal feature subse.

Fig 6. ACC of NFRFS with different values of p on different datasets.

Fig 7. NMI of NFRFS with different values of p on different datasets.

4.7. Convergence test

Additionally, the convergence speed of the method was studied through some numerical results. **Fig 8** and shows the convergence results. In **Fig 8**, the horizontal axis represents the number of iterations, and the vertical axis represents the value of the objective function. The results indicate that the proposed method is effective, with the objective function value decreasing very quickly and not increasing in subsequent iterations, which verifies the convergence of the proposed method.

Fig 8. Convergence curves of NFRFS on different datasets.

5. Conclusion

In this work, we propose an unsupervised feature selection algorithm based on $l_{2,p}$-norm feature reconstruction, which employs the $l_{2,p}$-norm to flexibly adjust the distance between the original space and the reconstructed subspace, thereby enhancing the model's robustness to noise and outliers. By leveraging inner product sparse regularization, the rows and columns of the feature selection matrix are sparsified to select representative and low-redundancy features. Incorporating adaptive structure learning into the feature selection objective function helps preserve the local structure of the data. Experimental results demonstrate that the NFRFS algorithm exhibits excellent performance in feature selection across different datasets. However, determining the optimal hyperparameter p in the feature reconstruction term in a more theoretical and efficient manner remains a direction for future research.

References

1. Nießl C, Herrmann M, Wiedemann C, Casalicchio G, Boulesteix AL. Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2022;12:e1441.

2. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. bioinformatics. 2007;23:2507–2517.

3. Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In: 2014 science and information conference. IEEE; 2014. p. 372–378.

4. Reddy GT, Reddy MPK, Lakshmanna K, Kaluri R, Rajput DS, Srivastava G, et al. Analysis of dimensionality reduction techniques on big data. Ieee Access. 2020;8:54776–54788.

5. Huang H, Shi G, He H, Duan Y, Luo F. Dimensionality reduction of hyperspectral imagery based on spatial–spectral manifold learning. IEEE transactions on cybernetics. 2019;50:2604–2616.

6. Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF. A review of unsupervised feature selection methods. Artificial Intelligence Review. 2020;53:907–948.

7. Wang S, Tang J, Liu H. Embedded unsupervised feature selection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 29; 2015.

8. Wang S, Zhu W. Sparse graph embedding unsupervised feature selection. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2016;48:329–341.

9. He X, Cai D, Niyogi P. Laplacian score for feature selection. Advances in neural information processing systems. 2005;18.

10. Tang C, Bian M, Liu X, Li M, Zhou H, Wang P, et al. Unsupervised feature selection via latent representation learning and manifold regularization. Neural Networks. 2019;117:163–178.

11. Du L, Shen YD. Unsupervised feature selection with adaptive structure learning. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining; 2015. p. 209–218.

12. Ma Z, Wei Y, Huang Y, Wang J. Unsupervised feature selection based on minimum-redundant subspace learning with self-weighted adaptive graph. Digital Signal Processing. 2024; p. 104738.

13. Emmert-Streib F, Dehmer M. High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection. Machine Learning and Knowledge Extraction. 2019;1:359–383.

14. 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; 2010. p. 333–342.

15. Nie F, Dong X, Tian L, Wang R, Li X. Unsupervised feature selection with constrained $l_{2,0}$-Norm and optimized graph. IEEE transactions on neural networks and learning systems. 2020;33:1702–1713.

16. Shang R, Liu C, Zhang W, Li Y, Xu S. Unsupervised feature selection method based on dual manifold learning and dual spatial latent representation. Expert Systems with Applications. 2024;255:124696.

17. Qi M, Wang T, Liu F, Zhang B, Wang J, Yi Y. Unsupervised feature selection by regularized matrix factorization. Neurocomputing. 2018;273:593–610.

18. Zhang M, Yang Y, Zhang H, Shen F, Zhang D. $l_{2,p}$-norm and sample constraint based feature selection and classification for AD diagnosis. Neurocomputing. 2016;195:104–111.

19. Wang S, Pedrycz W, Zhu Q, Zhu W. Subspace learning for unsupervised feature selection via matrix factorization. Pattern Recognition. 2015;48:10–19.

20. Shang R, Xu K, Shang F, Jiao L. Sparse and low-redundant subspace learning-based dual-graph regularized robust feature selection. Knowledge-Based Systems. 2020;187:104830.

21. Gong X, Yu L, Wang J, Zhang K, Bai X, Pal NR. Unsupervised feature selection via adaptive autoencoder with redundancy control. Neural Networks. 2022;150:87–101.

22. Sun Z, Xie H, Liu J, Yu Y. Multi-label feature selection via adaptive dual-graph optimization. Expert Systems with Applications. 2024;243:122884.

23. Zhang R, Zhang Y, Li X. Unsupervised feature selection via adaptive graph learning and constraint. IEEE Transactions on neural networks and learning systems. 2020;33:1355–1362.

24. Nie F, Zhu W, Li X. Unsupervised feature selection with structured graph optimization. In: Proceedings of the AAAI conference on artificial intelligence. vol. 30; 2016.

25. Bai H, Huang M, Zhong P. Precise feature selection via non-convex regularized graph embedding and self-representation for unsupervised learning. Knowledge-Based Systems. 2024;296:111900.

26. Moslemi A, Ahmadian A. Dual regularized subspace learning using adaptive graph learning and rank constraint: Unsupervised feature selection on gene expression microarray datasets. Computers in Biology and Medicine. 2023;167:107659.

27. Wang S, Chen J, Guo W, Liu G. Structured learning for unsupervised feature selection with high-order matrix factorization. Expert Systems with Applications. 2020;140:112878.

28. Karami S, Saberi-Movahed F, Tiwari P, Marttinen P, Vahdati S. Unsupervised feature selection based on variance–covariance subspace distance. Neural Networks. 2023;166:188–203.

29. Cao Z, Xie X, Sun F. Adaptive unsupervised feature selection with robust graph regularization. International Journal of Machine Learning and Cybernetics. 2024;15:341–354.

30. Zhu P, Hou X, Tang K, Liu Y, Zhao YP, Wang Z. Unsupervised feature selection through combining graph learning and $l_{2,0}$-norm constraint. Information Sciences. 2023;622:68–82.

31. Mi Y, Chen H, Luo C, Horng SJ, Li T. Unsupervised feature selection with high-order similarity learning. Knowledge-Based Systems. 2024;285:111317.

32. Zhou P, Du L, Li X, Shen YD, Qian Y. Unsupervised feature selection with adaptive multiple graph learning. Pattern Recognition. 2020;105:107375.

33. Tang C, Zheng X, Zhang W, Liu X, Zhu X, Zhu E. Unsupervised feature selection via multiple graph fusion and feature weight learning. Science China Information Sciences. 2023;66:152101.

Supporting information

S1 Data.