

Unsupervised feature selection based on adaptive similarity learning and subspace clustering

Mohsen Ghassemi Parsa^a, Hadi Zare^{a,*}, Mehdi Ghaee^b

^a Faculty of New Sciences and Technologies, University of Tehran, Iran

^b Department of Computer Science, Amirkabir University of Technology, Iran

ARTICLE INFO

Keywords:

Unsupervised feature selection
Graph learning
Subspace clustering
Sparse learning
Representation learning

ABSTRACT

Feature selection methods have an important role on the readability of data and the reduction of complexity of learning algorithms. In recent years, a variety of efforts are investigated on feature selection problems based on unsupervised viewpoint due to the laborious labeling task on large datasets. In this paper, we propose a novel approach on unsupervised feature selection initiated from the subspace clustering to preserve the similarities by representation learning of low dimensional subspaces among the samples. A self-expressive model is employed to implicitly learn the cluster similarities in an adaptive manner. The proposed method not only maintains the sample similarities through subspace clustering, but it also considers the underlying structure of data based on a regularized regression model. In line with the convergence analysis of the proposed method, the experimental results on benchmark datasets demonstrate the effectiveness of our approach as compared with the state-of-the-art methods.

1. Introduction

One of the most common approaches for dealing with high dimensional data is to select the appropriate features which is known as feature selection (FS) problem in machine learning community (Guyon and Elisseeff, 2003; Li et al., 2017a). FS techniques are widely applied in many domains including text mining (Liu et al., 2003), bioinformatics (Ding, 2003; Alirezanejad et al., 2020), social media (Tang and Liu, 2014), and ensemble learning (Abpeykar et al., 2019). On the one hand, FS approach provides a sparse representation for data with massive number of features to alleviate the curse of dimensionality effect on the learning performance (Bishop, 2006). On the other hand, the computational burden of facing with massive data could be decreased significantly through FS approach as compared with other techniques like feature extraction approaches (Guyon, 2006).

FS techniques can be categorized into wrapper, filter and embedded approaches by considering the evaluation criteria. In wrapper methods (Dy and Brodley, 2004; Kohavi and John, 1997; Roth and Lange, 2003), the evaluation process depends on the learning algorithms, in contrast to the filter methods (He et al., 2005; Nematzadeh et al., 2019) that only use data for evaluating without any learning phase. The embedded methods (Hou et al., 2014; Han et al., 2020), embed the process of selecting features in a learning algorithm.

On learning taxonomy, FS problem can be stated as “Supervised” and “Unsupervised”. Supervised FSs are primarily constructed from the

dependency among the features and the label information, including information theoretic approaches (Peng et al., 2005), statistical tests (Hall and Smith, 1999; Huan Liu and Setiono, 1995), sparse learning methods (Liu et al., 2009; Nie et al., 2010), and online feature selection based on structure learning (Zare and Niazi, 2016). These supervised approaches are mostly seek to detect the discriminative information as the main source of dependencies between the features and the label information (Peng et al., 2005). On the other hand, appropriate criterion on Unsupervised FS (UFS) problems is more challenging due to the ill-defined nature of the problem. There have been a variety of works on UFS (Solorio-Fernández et al., 2019) including metaheuristic approach (Tabakhi et al., 2014), graph clustering (Moradi and Rostami, 2015), feature-level reconstruction (Li et al., 2017b), manifold structure learning approach (Yang et al., 2011), and spectral clustering based (Li et al., 2012, 2014). These unsupervised approaches are mostly considered to preserve the underlying structure of data such as cluster analysis (Li et al., 2014), and graph learning (Moradi and Rostami, 2015).

One of the important aims in UFS approaches is to preserve the geometric structure in the selected features (He et al., 2005). To this end, the similarity preserving methods construct a graph similarities to maintain the geometric structure in the reduced space (Liu et al., 2014). The graph similarity computation is often performed independently from the feature selection, which may lead to a suboptimal solution. Adaptive structure learning methods (Du and Shen, 2015) explicitly

* Corresponding author.

E-mail address: h.zare@ut.ac.ir (H. Zare).

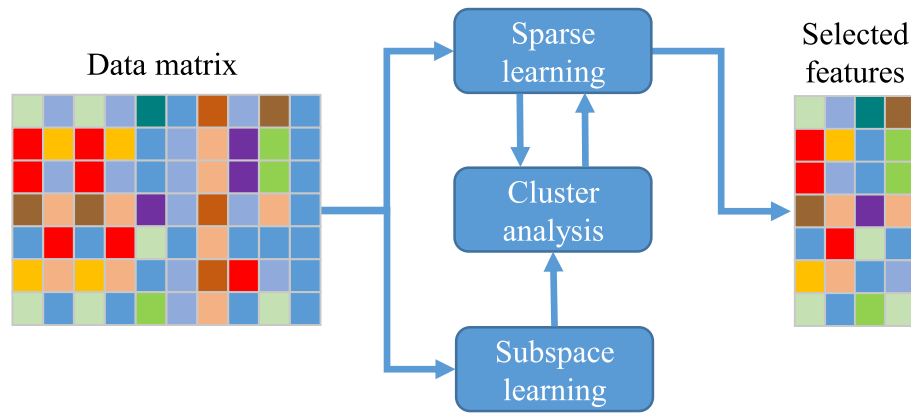


Fig. 1. The outline of the proposed method.

learn a similarity matrix which expand the search space of solutions. There are some UFS by considering the subspace learning to exploit the hidden multidimensional substructures of data (Wang et al., 2015). The main idea of the subspace learning is constructed from the representation of high dimensional data points based on the union of the subspaces (Soltanolkotabi et al., 2014). Subspace clustering refers to the process of clustering and detecting the low dimensional structure of the clusters at the same time (Vidal, 2011). The subspace learning UFS methods are considered a self-expressive model to reconstruct the data matrix based on either features (Shang et al., 2016, 2019), or samples (Zhu et al., 2017).

In this paper, we propose a novel UFS, “Subspace Clustering unsupervised Feature Selection” (SCFS), to adaptively maintain the similarity structure through an implicit similarity matrix computation. The proposed approach is constituted by a self-expressive model to include both of learning latent homogeneous structures and similarity matrix computation in a unified objective function aligned with an $\ell_{2,1}$ -norm to address a regularized regression model. In our approach, subspace learning is utilized to maintain the cluster similarities in the selected features and sparse learning is applied to learn a regularized regression model to measure the correlation between the features and the learned clustering information. The more a feature is related to the clusters, the more it is likely to be selected. The outline of the proposed method is presented in Fig. 1.

The main contributions of this paper are as follows,

- We propose a novel UFS method by applying subspace learning, cluster analysis and sparse learning to consider the sample similarities and underlying structure of data in the selected features in an integrated framework.
- We employ a self-expressive model on sample level to adaptively and implicitly learn the cluster similarities.
- A regularized regression approach is exploited to capture the correlation among features and clusters in a sparse manner.
- We introduce a unified optimization algorithm to address the proposed objective function.

This paper is organized as follows. The related works are introduced in Section 2. The proposed method and the corresponding optimization algorithm are presented in Section 3. Convergence analysis and computational complexity are discussed in Section 4. The connections and differences of the proposed method with the related algorithms are explained in the Section 5. Experimental setting and results are reported in Section 6. Finally, the conclusion of the paper is provided in Section 7.

2. Related works

Unsupervised feature selection methods generally select features based on the intrinsic structural characteristics of data. These methods

can be divided into four main categories, statistical (Krzanowski, 1987), similarity preserving (He et al., 2005), data reconstruction (Masaeli et al., 2010), and sparse learning approaches (Li et al., 2012).

Statistical methods such as MaxVar (Krzanowski, 1987) select features based on some statistical measures. The main focus of similarity preserving methods including Laplacian Score (He et al., 2005), SPEC (Zhao and Liu, 2007), and TraceRatio (Nie et al., 2008), is to maintain the local similarities. Reconstruction based methods employ a feature-level self-expressive model including convex principal feature selection, CPFS (Masaeli et al., 2010), $\ell_{2,1}$ -norm regularized self-representation, RSR (Zhu et al., 2015), embedded reconstruction based, REFS (Li et al., 2017b), and structure preserving, SPUS (Lu et al., 2018). Sparse unsupervised FS approaches are initiated from the ideas of the sparse machine learning (Bach et al., 2012). The core idea is to embed FS in a regularized learning model. There are several well-known approaches in this category including preserving the multi-cluster structure of data, MCFS (Cai et al., 2010), manifold structure learning approach using the scatter matrix, UDFS (Yang et al., 2011), kernel learning for local learning-based clustering, LLCFS (Zeng and Cheung, 2011), joint embedding learning and sparse regression, JELSR (Hou et al., 2014), nonnegative spectral clustering feature selection, NDFS (Li et al., 2012), global similarity preserving feature selection, SPFS (Zhao et al., 2013), global and local similarity preserving feature selection, GLSPFS (Liu et al., 2014), unsupervised feature selection with adaptive structure learning, FSASL (Du and Shen, 2015), and co-regularized unsupervised feature selection, CUFS (Zhu et al., 2018).

There are some UFS methods by considering the subspace learning idea. Feature-level reconstruction based approach was proposed in, MFFS (Wang et al., 2015) by exploiting the matrix factorization. In Shang et al. (2016), a graph regularized approach was introduced to maintain the local similarities. Spectral clustering is aligned with subspace learning by computing similarity matrix in an explicit way (Zhu et al., 2017). A sparse learning approach was suggested in Shang et al. (2019) to select features based on the local structure of the samples. While, UFS methods with the aid of subspace learning mainly reconstruct the data matrix in feature-level, the sample-level characteristics such as the cluster structures are not thoroughly incorporated in them.

There are several important properties that can be considered for a suitable UFS process including, self-expression, similarity preserving, adaptive similarity learning, joint learning, cluster analysis, regression and regularization. Self-expressive property indicates the ability of data reconstruction based on samples or features. Similarity preserving is related to maintain sample similarities in the reduced data matrix which can be performed explicitly or implicitly. The incremental learning of the similarity matrix through the feature selection process is called as adaptive similarity learning. Joint learning refers to perform both of subspace learning and feature selection in a unified manner.

Table 1

A comparison of the related unsupervised feature selection methods.

Algorithm	Self-expression	Similarity preserving	Adaptive similarity learning	Joint learning	Cluster analysis	Regression	Regularization
MaxVar (Krzanowski, 1987)	×	×	×	×	×	×	×
LS (He et al., 2005)	×	Explicit	×	✓	×	×	×
TraceRatio (Nie et al., 2008)	×	Explicit	×	✓	×	×	×
CPFS (Masaeli et al., 2010)	Feature	×	×	×	×	×	✓
RSR (Zhu et al., 2015)	Feature	×	×	×	×	×	✓
MCFS (Cai et al., 2010)	×	Explicit	×	×	✓	✓	✓
UDFS (Yang et al., 2011)	×	Explicit	×	✓	×	×	✓
LLCFS (Zeng and Cheung, 2011)	×	Explicit	✓	✓	✓	×	✓
NDFS (Li et al., 2012)	×	Explicit	×	✓	✓	✓	✓
GLSPFS (Liu et al., 2014)	×	Explicit	×	✓	×	✓	✓
MFFS (Wang et al., 2015)	Feature	×	×	×	×	×	×
FSASL (Du and Shen, 2015)	×	Explicit	✓	✓	×	✓	✓
SPUFS (Lu et al., 2018)	Feature	Explicit	×	✓	×	×	✓
LDSSL (Shang et al., 2019)	Feature	Explicit	×	✓	×	×	✓
SCUFS (Zhu et al., 2017)	Sample	Explicit	✓	✓	✓	✓	✓
SCFS	Sample	Implicit	✓	✓	✓	✓	✓

Some UFS methods exploit cluster analysis to select relevant features. The correlation between features and low dimensional representation matrix are approximated by a regression model that can be utilized to measure the importance of features. Finally, regularization indicates the employing of regularized terms to yield a sparse solution. The well-known UFS methods are compared by the aid of these properties in Table 1. While most of the UFS techniques are constructed from one or more of these properties, we propose a unified framework to consider all of these characteristics and implicit similarity preserving to result in a more robust and efficient UFS approach on the sample level.

3. The proposed method

3.1. Notations

Throughout this paper, matrices are denoted by bold uppercase and vectors by bold lowercase characters. Let $\mathbf{B}$ be an arbitrary matrix, B_{ij} is its (i, j) -th element, and $\mathbf{b}_i$ denotes the i th row. The Frobenius norm, the trace and the transpose operators on matrix $\mathbf{B}$ are denoted by $\|\mathbf{B}\|_F$, $\text{tr}(\mathbf{B})$, and $\mathbf{B}^T$, respectively. The ℓ_2 -norm of a vector $\mathbf{v}$ is denoted as $\|\mathbf{v}\|_2$ and the $\ell_{2,1}$ -norm is defined as following,

$$\|\mathbf{B}\|_{2,1} = \sum_i \sqrt{\sum_j B_{ij}^2}.$$

$\mathbf{X} \in \mathbb{R}^{n \times p}$ denotes the data matrix, where n and p are the number of samples and features. $\mathbf{G} \in \mathbb{R}^{n \times c}$ represents the clustering matrix, where c is the number of the clusters.

3.2. The proposed method

At first, the similarity matrix is implicitly computed by subspace learning. The proposed self-expressive similarity representation is given in Eq. (1),

$$\begin{aligned} \min_{\mathbf{G}} \quad & \|\mathbf{X} - \mathbf{G}\mathbf{G}^T\mathbf{X}\|_F^2 \\ \text{s.t.} \quad & \mathbf{G} \geq 0, \mathbf{G}\mathbf{G}^T\mathbf{1} = \mathbf{1}, \end{aligned} \quad (1)$$

where $\mathbf{1}$ is an $n \times n$ matrix of ones and the constraint $\mathbf{G}\mathbf{G}^T\mathbf{1} = \mathbf{1}$ is imposed to normalize the similarity matrix. The symmetric nonnegative matrix $\mathbf{G}\mathbf{G}^T$ is learned such that the samples within common subspaces tend to attain large values in $\mathbf{G}\mathbf{G}^T$. In lines with consideration of $\mathbf{G}$ as a low dimensional representation of $\mathbf{X}$ by assuming $c < \{n, p\}$, $\mathbf{G}$ can also be interpreted as a clustering matrix. Moreover, $\mathbf{G}\mathbf{G}^T$, represents the pairwise sample similarities in terms of the clustering values.

The next stage is to construct a sparse transformation $\mathbf{W}$ on the data matrix $\mathbf{X}$ by employing the clustering matrix $\mathbf{G}$, joined with a regularization term,

$$\min_{\mathbf{W}} \quad \|\mathbf{X}\mathbf{W} - \mathbf{G}\|_F^2 + \beta\|\mathbf{W}\|_{2,1}, \quad (2)$$

where $\mathbf{W} \in \mathbb{R}^{p \times c}$ is a linear and low dimensional transformation matrix, and β is a regularization parameter. The objective function in Eq. (2) represents the linear transformation model to measure the association between features and clusters. The $\ell_{2,1}$ norm induces sparsity on the rows of the transformation matrix, $\mathbf{w}_i$'s. When $\mathbf{w}_i$'s are closer to zero, their correspondence features are less relevant and more likely to be eliminated from the final candidate set of the features.

By integrating Eq. (1) and (2) in a joint objective function, our final model is obtained as follows,

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{G}} \quad & \|\mathbf{X} - \mathbf{G}\mathbf{G}^T\mathbf{X}\|_F^2 + \alpha\|\mathbf{X}\mathbf{W} - \mathbf{G}\|_F^2 + \beta\|\mathbf{W}\|_{2,1} \\ \text{s.t.} \quad & \mathbf{G} \geq 0, \mathbf{G}\mathbf{G}^T\mathbf{1} = \mathbf{1}, \end{aligned} \quad (3)$$

where α is a tuning parameter. By solving the objective function in Eq. (3), $\mathbf{W}$ and $\mathbf{G}$ are iteratively updated in advance of achieving the optimal result.

3.3. Feature ranking

The weight matrix $\mathbf{W} = [\mathbf{w}_1; \mathbf{w}_2; \dots; \mathbf{w}_p]$ is obtained by our framework, where the i th row of $\mathbf{W}$, $\mathbf{w}_i \in \mathbb{R}^{1 \times c}$, corresponds to the importance of the i th feature. More formally, $\|\mathbf{w}_i\|_2$ is used to rank the most important features. Finally, the reduced data matrix $\mathbf{X}_{\text{new}} \in \mathbb{R}^{n \times m}$ is generated from the first m ranked features.

3.4. An illustrative example

We illustrate the steps of the proposed method in Fig. 2. $\mathbf{X}$ is a nonnegative artificial data matrix with seven samples and ten features. The bright entries of $\mathbf{X}$ are close to zero and the dark ones are far from zero. Initially, subspace learning stage provides the clustering matrix $\mathbf{G}$, which is used to construct the similarity matrix $\mathbf{G}\mathbf{G}^T$ among samples. Then, a sparse learning method is applied for learning the regularized coefficients $\mathbf{W}$ through a regression model to measure the importance of features. The $\mathbf{W}$ and $\mathbf{G}$ are optimized in an iterative process. Finally, the features are ranked by computing the ℓ_2 -norm on the rows in $\mathbf{W}$. In this example, $\|\mathbf{w}_6\|_2$, $\|\mathbf{w}_3\|_2$, and $\|\mathbf{w}_9\|_2$ are the top three weights which correspond to F_6 , F_3 , and F_9 . As we can see from Fig. 2, the learned hidden subspace is constituted from the top three selected features.

3.5. Optimization

The primary objective function in Eq. (3) can be considered as,

$$\min_{\mathbf{W}, \mathbf{G} \geq 0, \mathbf{G}\mathbf{G}^T\mathbf{1} = \mathbf{1}} \quad f(\mathbf{W}, \mathbf{G}) = \|\mathbf{X} - \mathbf{G}\mathbf{G}^T\mathbf{X}\|_F^2 + \alpha\|\mathbf{X}\mathbf{W} - \mathbf{G}\|_F^2 + \beta\|\mathbf{W}\|_{2,1}. \quad (4)$$

A gradient based procedure is utilized to solve this optimization problem by considering the main elements $\mathbf{W}$ and $\mathbf{G}$. It begins by fixing

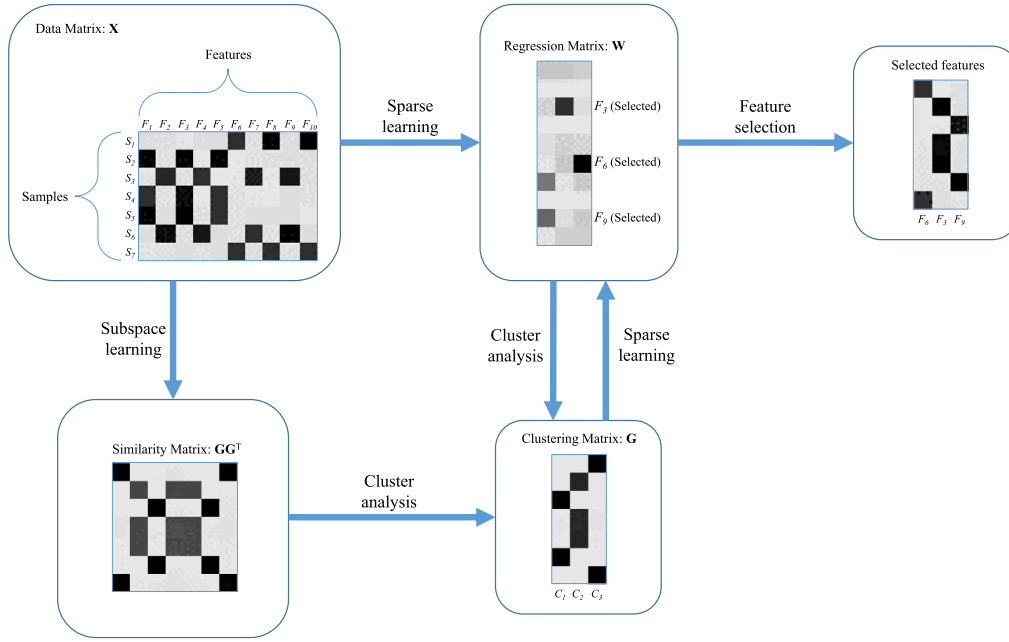


Fig. 2. An illustrative case study to describe the proposed method.

one element and finding the optimum value for the other one which is described in below. Initially, G is fixed to yield the following objective function,

$$\min_{\mathbf{W}} f(\mathbf{W}) = \alpha \|\mathbf{XW} - \mathbf{G}\|_F^2 + \beta \|\mathbf{W}\|_{2,1}. \quad (5)$$

Taking the derivative to calculate the $\nabla f(\mathbf{W})$ and setting it to zero,

$$\mathbf{W} = (\alpha \mathbf{X}^T \mathbf{X} + \beta \mathbf{D})^{-1} \alpha \mathbf{X}^T \mathbf{G}, \quad (6)$$

where $\mathbf{D}$ is a diagonal matrix with,

$$D_{ii} = \frac{1}{2\|\mathbf{w}_i\|_2 + \epsilon}, \quad (7)$$

where ϵ is a very small positive number to prevent the division by zero. Then, G is updated through the objective function in Eq. (8) by fixing $\mathbf{W}$,

$$\min_{\mathbf{G} \geq 0, \mathbf{G}\mathbf{G}^T \mathbf{1} = \mathbf{1}} f(\mathbf{G}) = \|\mathbf{X} - \mathbf{G}\mathbf{G}^T \mathbf{X}\|_F^2 + \alpha \|\mathbf{XW} - \mathbf{G}\|_F^2. \quad (8)$$

Eq. (8) is rewritten to relax the constraints as,

$$f(\mathbf{G}) = \|\mathbf{X} - \mathbf{G}\mathbf{G}^T \mathbf{X}\|_F^2 + \alpha \|\mathbf{XW} - \mathbf{G}\|_F^2 + \gamma \|\mathbf{G}\mathbf{G}^T \mathbf{1} - \mathbf{1}\|_F^2 + tr(\Phi \mathbf{G}^T), \quad (9)$$

where $\gamma > 0$ is a parameter to control the normalizing constraint and practically should be a large number. Φ is the Lagrange multiplier for $\mathbf{G} \geq 0$ constraint. Setting the derivative of $f(\mathbf{G})$ with respect to $\mathbf{G}$ to 0,

$$2\mathbf{M}\mathbf{G}^T \mathbf{G} + 2\mathbf{G}\mathbf{G}^T \mathbf{M} + 2\alpha \mathbf{G} - 4\mathbf{M} - 2\alpha \mathbf{XW} + \Phi = 0, \quad (10)$$

where $\mathbf{M} = (\mathbf{X}\mathbf{X}^T + n\gamma \mathbf{1}) \mathbf{G}$ and n is the number of samples. By applying the KKT condition (Kuhn and Tucker, 2014), the following updating rule is obtained,

$$G_{ij} = G_{ij} \frac{(2\mathbf{M} + \alpha \mathbf{XW})_{ij}}{(\mathbf{M}\mathbf{G}^T \mathbf{G} + \mathbf{G}\mathbf{G}^T \mathbf{M} + \alpha \mathbf{G})_{ij}}. \quad (11)$$

Therefore, by initializing the $\mathbf{G}$ and $\mathbf{D}$, in each iteration of the proposed formulation, first $\mathbf{W}$ is updated by Eq. (6), and then $\mathbf{G}$ and $\mathbf{D}$ is updated by Eq. (7) and (11). Algorithm 1 describes the optimization process of the proposed method.

4. The analysis of the proposed algorithm

This section presents the convergence behavior and computational complexity of the proposed method.

Algorithm 1 SCFS algorithm.

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and parameters α and β .

1: $t = 0$.

2: Initialize $\mathbf{G}_0 \in \mathbb{R}^{n \times c}$.

3: Initialize $\mathbf{D}_0$ as an identity matrix.

4: **repeat**

5: $\mathbf{W}_{t+1} = (\alpha \mathbf{X}^T \mathbf{X} + \beta \mathbf{D}_t)^{-1} \alpha \mathbf{X}^T \mathbf{G}_t$.

6: $\mathbf{M}_t = (\mathbf{X}\mathbf{X}^T + n\gamma \mathbf{1}) \mathbf{G}_t$.

7: $(G_{t+1})_{ij} = (G_t)_{ij} \frac{(2\mathbf{M}_t + \alpha \mathbf{XW}_{t+1})_{ij}}{(\mathbf{M}_t \mathbf{G}_t^T \mathbf{G}_t + \mathbf{G}_t \mathbf{G}_t^T \mathbf{M}_t + \alpha \mathbf{G}_t)_{ij}}$.

8: Update the diagonal matrix $\mathbf{D}$ as $(D_{t+1})_{ii} = \frac{1}{2\|(\mathbf{w}_{t+1})_i\|_2 + \epsilon}$.

9: $t = t + 1$.

10: **until** Convergence of the objective function in Eq. (3).

Output: Sort features by descending order of $\|\mathbf{w}_i\|_2$.

4.1. Convergence analysis

Our aim is to show the non-increasing behavior of the primary objective function in Eq. (3). First, a lemma is given (Nie et al., 2010).

Lemma 1. For any nonzero vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, the following holds,

$$\|\mathbf{u}\|_2 - \frac{\|\mathbf{u}\|_2^2}{2\|\mathbf{v}\|_2} \leq \|\mathbf{v}\|_2 - \frac{\|\mathbf{v}\|_2^2}{2\|\mathbf{v}\|_2}. \quad (12)$$

Theorem 1. The objective function in Eq. (3) is non-increasing in each iteration by employing the updating rules in Algorithm 1.

Proof. First, the objective function can be written as,

$$f(\mathbf{W}, \mathbf{G}) = \|\mathbf{X} - \mathbf{G}\mathbf{G}^T \mathbf{X}\|_F^2 + \alpha \|\mathbf{XW} - \mathbf{G}\|_F^2 + \beta \|\mathbf{W}\|_{2,1} + \gamma \|\mathbf{G}\mathbf{G}^T \mathbf{1} - \mathbf{1}\|_F^2. \quad (13)$$

By fixing $\mathbf{G}_t$, we should justify the following inequality,

$$f(\mathbf{W}_{t+1}, \mathbf{G}_t) \leq f(\mathbf{W}_t, \mathbf{G}_t). \quad (14)$$

Based on Eq. (5), inequality (14) can be written as,

$$\begin{aligned} & \|\mathbf{X}\mathbf{W}_{t+1} - \mathbf{G}_t\|_F^2 + \beta \sum_{i=1}^p \left(\frac{\|(\mathbf{w}_{t+1})_i\|_2^2}{2\|(\mathbf{w}_t)_i\|_2} \right) \\ & \leq \|\mathbf{X}\mathbf{W}_t - \mathbf{G}_t\|_F^2 + \beta \sum_{i=1}^p \left(\frac{\|(\mathbf{w}_t)_i\|_2^2}{2\|(\mathbf{w}_t)_i\|_2} \right). \end{aligned} \quad (15)$$

The inequality (15) is followed as,

$$\begin{aligned} & \|\mathbf{X}\mathbf{W}_{t+1} - \mathbf{G}_t\|_F^2 + \beta \|\mathbf{W}_{t+1}\|_{2,1} - \beta \sum_{i=1}^p \left(\|\mathbf{w}_{t+1}\|_2 - \frac{\|(\mathbf{w}_{t+1})_i\|_2^2}{2\|(\mathbf{w}_t)_i\|_2} \right) \\ & \leq \|\mathbf{X}\mathbf{W}_t - \mathbf{G}_t\|_F^2 + \beta \|\mathbf{W}_t\|_{2,1} - \beta \sum_{i=1}^p \left(\|\mathbf{w}_t\|_2 - \frac{\|(\mathbf{w}_t)_i\|_2^2}{2\|(\mathbf{w}_t)_i\|_2} \right). \end{aligned} \quad (16)$$

According to Lemma 1,

$$\|\mathbf{X}\mathbf{W}_{t+1} - \mathbf{G}_t\|_F^2 + \beta \|\mathbf{W}_{t+1}\|_{2,1} \leq \|\mathbf{X}\mathbf{W}_t - \mathbf{G}_t\|_F^2 + \beta \|\mathbf{W}_t\|_{2,1}. \quad (17)$$

Taking fixed $\mathbf{W}_{t+1}$, based on a similar approach in Lee and Seung (2000), it follows,

$$f(\mathbf{W}_{t+1}, \mathbf{G}_{t+1}) \leq f(\mathbf{W}_{t+1}, \mathbf{G}_t). \quad (18)$$

Hence,

$$f(\mathbf{W}_{t+1}, \mathbf{G}_{t+1}) \leq f(\mathbf{W}_{t+1}, \mathbf{G}_t) \leq f(\mathbf{W}_t, \mathbf{G}_t). \quad (19)$$

Therefore, Algorithm 1 will monotonically decrease the objective function in Eq. (3) based on the relations (17) and (18).

4.2. Computational complexity

The main steps of Algorithm 1 contains the updating $\mathbf{W}$ and $\mathbf{G}$ on each iteration. The update of $\mathbf{W}$ and $\mathbf{G}$ take $O(p^3 + np^2 + npc)$ and $O(n^2p + n^2c + npc)$ time complexity. Hence, the time complexity of the proposed algorithm is $\max\{O(p^3), O(np^2), O(n^2p), O(n^2c), O(npc)\}$. In most applied scenarios, $c \ll p$, that implies the time complexity of the proposed algorithm could be reduced to $\max\{O(p^3), O(n^2p)\}$.

5. Connections and differences with the related algorithms

In this section, we discuss the connections and differences between the proposed method and four algorithms, including RSR (Zhu et al., 2015), MCFS (Cai et al., 2010), NDFS (Li et al., 2012), and SCUFS (Zhu et al., 2017).

5.1. RSR

RSR (Zhu et al., 2015) is mainly a feature reconstruction approach, which attempts to learn a sparse transformation matrix by an $\ell_{2,1}$ -norm in both of loss and regularization term. The RSR objective function is specified as following,

$$\min_{\mathbf{Q}} \|\mathbf{X} - \mathbf{X}\mathbf{Q}\|_{2,1} + \lambda \|\mathbf{Q}\|_{2,1}. \quad (20)$$

By learning the reconstruction matrix $\mathbf{Q} \in \mathbb{R}^{p \times p}$, each feature is expressed as a linear combination of all features. If a row of the matrix $\mathbf{Q}$ becomes zero, then the corresponding feature can be regarded as a redundant feature. The main objectives in Eq. (20) and Eq. (3) imply that both RSR and SCFS exploit self-expression mechanisms. RSR reconstructs data in the level of features, unlike to SCFS, which is based on sample level reconstruction. Furthermore, RSR and SCFS differ in considering the underlying structure of data. While RSR neglects to maintain the sample similarities in the reduced space, SCFS considers to preserve the cluster similarities.

5.2. MCFS

MCFS (Cai et al., 2010) is a two-stage method to select features based on preserving geometric structure of the samples by a low dimensional embedding and linear regression. The formulation is as follows,

$$\begin{aligned} & \min_{\mathbf{G}^T \mathbf{G} = \mathbf{I}} \text{tr}(\mathbf{G}^T \mathbf{L} \mathbf{G}) \\ & \min_{\mathbf{W}} \|\mathbf{X}\mathbf{W} - \mathbf{G}\|_F + \lambda \|\mathbf{W}\|_{1,1}, \end{aligned} \quad (21)$$

where $\mathbf{L}$ is a Laplacian matrix. The main connections between MCFS and SCFS can be observed on learning low dimensional embedding and sparse regression matrix. The main differences between MCFS and SCFS, based on Eq. (21) and Eq. (3) are as follows. 1) MCFS computes the graph similarity matrix before optimizing its objective functions, while SCFS adaptively learn sample similarities. 2) MCFS explicitly calculate a similarity matrix, (3) MCFS proposes a two-stage method, while SCFS devises a joint framework. 4) MCFS employs $\ell_{1,1}$, while SCFS applies $\ell_{2,1}$ regularization.

5.3. NDFS

NDFS (Li et al., 2012) is a spectral clustering based method by the following formulation,

$$\begin{aligned} & \min_{\mathbf{W}, \mathbf{G}} \text{tr}(\mathbf{G}^T \mathbf{L} \mathbf{G}) + \alpha (\|\mathbf{X}\mathbf{W} - \mathbf{G}\|_F + \lambda \|\mathbf{W}\|_{2,1}) \\ & \text{s.t. } \mathbf{G} \geq 0, \mathbf{G}^T \mathbf{G} = \mathbf{I}, \end{aligned} \quad (22)$$

While NDFS computes the graph similarities independent from the feature selection process, the sample similarities is learned on SCFS framework to prevent a suboptimal solution. The SCFS exploits subspace clustering, as compared to utilize spectral clustering in NDFS to learn clustering matrix. NDFS and SCFS are similar in using regression model and $\ell_{2,1}$ -norm regularization on their objectives.

5.4. SCUFS

SCUFS (Zhu et al., 2017) proposed to select features based on subspace clustering and adaptive similarity learning as,

$$\begin{aligned} & \min_{\mathbf{W}, \mathbf{G}, \mathbf{S}} \|\mathbf{X} - \mathbf{S}\mathbf{X}\|_F^2 + \lambda_1 \text{tr}(\mathbf{G}^T \mathbf{L}_S \mathbf{G}) + \|\mathbf{X}\mathbf{W} - \mathbf{G}\|_{2,1} + \lambda_2 \|\mathbf{W}\|_{2,1} \\ & \text{s.t. } \mathbf{S}\mathbf{1} = \mathbf{1}, \mathbf{S}_{ii} = 0, \mathbf{G} \geq 0, \mathbf{G}^T \mathbf{G} = \mathbf{I}, \end{aligned} \quad (23)$$

where, $\mathbf{S}$ is a similarity matrix, and $\mathbf{L}_S$ is a Laplacian matrix based on $\mathbf{S}$. The connections between SCUFS and the proposed method, SCFS, are in adaptive similarity learning and cluster analysis. The difference between these methods is in how to learn the graph similarity matrix. SCUFS explicitly learns the similarity matrix $\mathbf{S}$ and performs spectral clustering based on $\mathbf{S}$, while our proposed approach learns concurrently the clustering matrix and similarity matrix $\mathbf{G}\mathbf{G}^T$. Moreover, SCUFS expands the search space by imposing to learn an explicit similarity matrix $\mathbf{S}$.

6. Experiments

In this section, the proposed method is evaluated using benchmark datasets by standard evaluation measures. A bunch of state-of-the-art FS methods are compared with SCFS where the results and experimental setting are presented in the following. The source codes of our method are released on GitHub, and the link is <https://github.com/mohsengh/SCFS>.

Table 2

The main properties of datasets in the experiments.

Dataset	#samples	#features	#classes	Type	Category
Lung	203	3312	5	Continuous	Biology
Lymphoma	96	4026	9	Discrete	Biology
Prostate-GE	102	5966	2	Continuous	Biology
ORL	400	1024	40	Discrete	Image
Isolet	1560	617	26	Continuous	Voice
BASEHOCK	1993	4862	2	Discrete	Text
BA	1404	320	36	Discrete	Image
GLIOMA	50	4434	4	Continuous	Biology
Madelon	2600	500	2	Continuous	Artificial

6.1. Datasets

We employed datasets from a variety of applied domains, including the lung carcinomas of human (Lung [Bhattacharjee et al., 2001](#)), the categories of human lymphoma (Lymphoma [Alizadeh et al., 2000](#)), gene expression of prostate cancer (Prostate-GE [Singh et al., 2002](#)), face image (ORL [Cai et al., 2006](#)), spoken letter recognition data (Isolet [Cole and Fandy, 1990](#)), newsgroup documents (BASEHOCK [Li et al., 2017a](#)), binary images of digits and capital alphabet letters (BA [Belhumeur et al., 1997](#)), malignant tumor of the brain (GLIOMA [Nutt et al., 2003](#)), and non sparse synthetic dataset (Madelon [Li et al., 2017a](#)). The BA can be accessed from <https://cs.nyu.edu/~roweis/data.html>. All other datasets are available on repository ([Li et al., 2017a](#)). Table 2 reports the main characteristics of datasets.

6.2. Evaluation measures

The performance is evaluated in terms of clustering by two widely used and standard measures, Accuracy (Acc) and Normalized Mutual Information (NMI). The performance of clustering is evaluated by Coefficient of Variation (CV), and the stability of the feature selection methods is assessed by Stability Index (SI) measure.

By taking $\mathbf{y}$ as the ground truth label information, and $\mathbf{z}$ as the predicted ones, Acc is defined as,

$$\text{Acc}(\mathbf{y}, \mathbf{z}) = \frac{1}{n} \sum_{i=1}^n \delta(y_i, \text{map}(z_i)),$$

where $\delta(a, b)$ equals to 1 if $a = b$ and 0, otherwise. The best permutation of $\mathbf{z}$ to match $\mathbf{y}$ values is found by $\text{map}(\cdot)$ function based on the Kuhn–Munkres approach ([Lovasz, 1986](#)).

The definition of NMI is given as,

$$\text{NMI}(\mathbf{y}, \mathbf{z}) = \frac{I(\mathbf{y}, \mathbf{z})}{\max(H(\mathbf{y}), H(\mathbf{z}))},$$

where $H(\cdot)$ represents the entropy and $I(\mathbf{y}, \mathbf{z})$ is the mutual information of $\mathbf{y}$ and $\mathbf{z}$ defined as,

$$I(\mathbf{y}, \mathbf{z}) = \sum_{y \in \mathbf{y}} \sum_{z \in \mathbf{z}} p(y, z) \log \left(\frac{p(y, z)}{p(y)p(z)} \right).$$

Table 3Clustering results (Acc% $\pm$ std) of unsupervised feature selection methods on standard datasets. Bold and underlined numbers are the best and the second best.

Dataset	Lung	Lymphoma	Prostate-GE	ORL	Isolet	BASEHOCK	BA	GLIOMA	Madelon
Baseline	71.67 $\pm$ 6.86	58.75 $\pm$ 5.19	58.82 $\pm$ 0.00	<u>59.14</u> $\pm$ 2.11	63.19 $\pm$ 2.19	50.08 $\pm$ 0.00	42.04 $\pm$ 1.56	58.80 $\pm$ 3.25	50.34 $\pm$ 0.10
LS	61.46 $\pm$ 2.61	50.12 $\pm$ 2.26	60.82 $\pm$ 1.71	49.94 $\pm$ 3.62	53.31 $\pm$ 4.49	50.63 $\pm$ 0.23	41.70 $\pm$ 3.54	55.85 $\pm$ 2.42	50.34 $\pm$ 0.04
UDFS	56.65 $\pm$ 4.93	59.60 $\pm$ 2.50	61.36 $\pm$ 1.14	49.30 $\pm$ 3.90	57.97 $\pm$ 6.24	<u>51.57</u> $\pm$ 0.56	41.47 $\pm$ 3.51	58.48 $\pm$ 2.40	57.87 $\pm$ 0.57
NDFS	<u>83.71</u> $\pm$ 0.64	63.81 $\pm$ 0.33	60.07 $\pm$ 0.77	58.81 $\pm$ 1.19	<u>67.72</u> $\pm$ 1.51	50.18 $\pm$ 0.14	41.62 $\pm$ 2.91	62.15 $\pm$ 3.79	53.57 $\pm$ 3.64
SPUFS	68.33 $\pm$ 1.20	53.57 $\pm$ 3.42	61.09 $\pm$ 1.45	48.63 $\pm$ 2.68	63.65 $\pm$ 13.10	50.41 $\pm$ 0.40	38.21 $\pm$ 5.17	56.85 $\pm$ 0.84	51.18 $\pm$ 0.16
LDSSL	64.59 $\pm$ 5.14	58.11 $\pm$ 1.67	60.27 $\pm$ 0.74	57.78 $\pm$ 1.49	65.97 $\pm$ 3.64	50.49 $\pm$ 0.11	41.65 $\pm$ 3.02	57.02 $\pm$ 0.73	52.24 $\pm$ 1.57
MaxVar	70.35 $\pm$ 3.65	56.85 $\pm$ 6.61	60.66 $\pm$ 0.72	45.49 $\pm$ 4.55	57.77 $\pm$ 5.16	50.08 $\pm$ 0.00	38.03 $\pm$ 4.63	45.27 $\pm$ 1.20	50.29 $\pm$ 0.03
LLCFS	79.35 $\pm$ 4.33	55.08 $\pm$ 3.99	59.08 $\pm$ 0.54	54.39 $\pm$ 3.11	57.13 $\pm$ 4.29	50.08 $\pm$ 0.00	37.76 $\pm$ 3.73	49.47 $\pm$ 1.05	50.84 $\pm$ 1.25
TraceRatio	53.33 $\pm$ 5.54	54.89 $\pm$ 4.40	58.82 $\pm$ 0.57	47.21 $\pm$ 4.48	43.49 $\pm$ 12.56	50.08 $\pm$ 0.00	38.28 $\pm$ 5.07	58.93 $\pm$ 3.71	52.83 $\pm$ 1.99
MCFS	77.09 $\pm$ 3.47	63.32 $\pm$ 2.22	59.55 $\pm$ 1.49	58.01 $\pm$ 0.71	45.38 $\pm$ 7.56	50.32 $\pm$ 0.28	41.98 $\pm$ 2.11	61.05 $\pm$ 4.55	52.30 $\pm$ 0.98
SCUFS	69.71 $\pm$ 4.68	<u>64.03</u> $\pm$ 0.86	62.23 $\pm$ 0.76	57.29 $\pm$ 1.62	61.85 $\pm$ 7.50	50.92 $\pm$ 0.70	<u>42.47</u> $\pm$ 1.72	<u>63.77</u> $\pm$ 1.28	<u>58.87</u> $\pm$ 1.49
SCFS	86.70 $\pm$ 1.38	64.87 $\pm$ 1.59	<u>61.70</u> $\pm$ 0.73	59.19 $\pm$ 0.83	69.17 $\pm$ 1.03	51.95 $\pm$ 0.44	43.57 $\pm$ 1.96	64.93 $\pm$ 2.18	58.94 $\pm$ 1.56

By repeating the clustering algorithm, the mean (Mean) and standard deviation (STD) values of Acc and NMI are computed. Moreover, Coefficient of Variation (CV) is applied to measure the performance of clustering, where the smaller CV indicates the more robustness of the clustering algorithm. The definition of CV is as follows,

$$\text{CV} = \frac{\text{STD}}{\text{Mean}}.$$

The Stability Index (SI) ([Kuncheva, 2007](#)) is a measure to evaluate the stability of a feature selection method when a small perturbation is applied in data. Let $\{S_1, S_2, \dots, S_k\}$ be the selected feature sets by running k times of the feature selection process. SI is calculated as,

$$\text{SI} = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^k \frac{r_{ij}p - m^2}{m(p-m)},$$

where $r_{ij} = |S_i \cap S_j|$, $m = |S_i| = |S_j|$, and p is the size of the original feature set.

6.3. The experimental setting

The state-of-the-art UFS methods are applied such as LS ([He et al., 2005](#)), UDFS ([Yang et al., 2011](#)), NDFS ([Li et al., 2012](#)), SPUFS ([Lu et al., 2018](#)), LDSSL ([Shang et al., 2019](#)), MaxVar ([Krzanowski, 1987](#)), LLCFS ([Zeng and Cheung, 2011](#)), TraceRatio ([Nie et al., 2008](#)), MCFS ([Cai et al., 2010](#)), SCUFS ([Zhu et al., 2017](#)), and Baseline means to select all of the original features.

We set $k = 5$ on k-nearest neighbor algorithm, and $\sigma = 1$ for the bandwidth parameter in the Gaussian kernel for the methods based on explicit construction of the graph matrix. The $\gamma = 10^6$ is taken on our method and NDFS. The parameter c is set to the number of clusters in the clustering based methods, MCFS, LLCFS, NDFS, SCUFS, and SCFS. The grid search strategy is employed to choose the appropriate weight parameters α and β among the set of $\{10^{-4}, 10^{-2}, 1, 10^2, 10^4\}$ candidates. We limit data by selecting different number of features in the range of $\{50, 100, 150, 200, 250, 300\}$ and cluster each ones by k-means algorithm, and then evaluate the clustering results by Acc and NMI measures. The mean and standard deviation values of Acc and NMI are reported by repeating the experiments for 20 times.

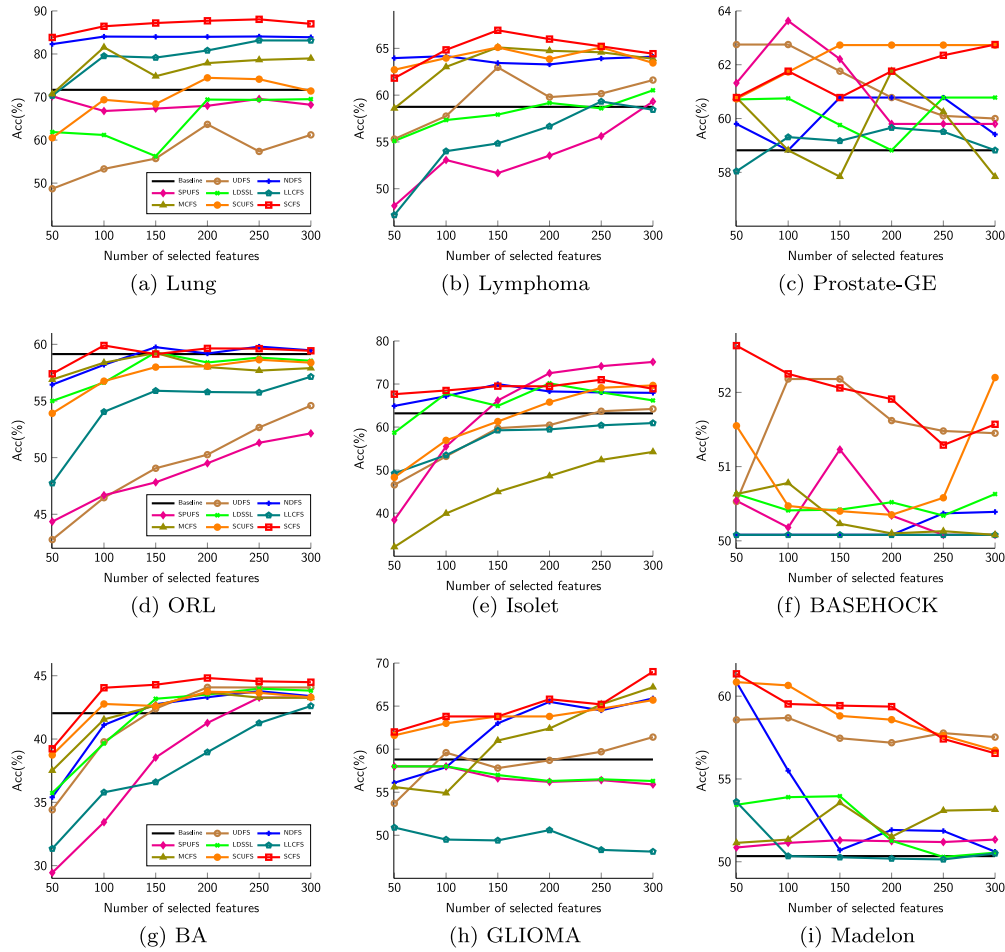
6.4. Experimental results

The performance of the feature selection algorithms are evaluated in terms of Acc, NMI, CV, and SI. The mean and standard deviation of the clustering result are reported in the [Tables 3 and 4](#). The best and the second best results are marked as bold and underline. By considering the [Tables 3 and 4](#), we have the following conclusions,

- The proposed method, SCFS, outperforms the Baseline method in most cases, which implies the efficacy of SCFS to select more relevant features rather than irrelevant and redundant ones.

Table 4Clustering results (NMI% $\pm$ std) of unsupervised feature selection methods on standard datasets. Bold and underlined numbers are the best and the second best.

Dataset	Lung	Lymphoma	Prostate-GE	ORL	Isolet	BASEHOCK	BA	GLIOMA	Madelon
Baseline	62.90 $\pm$ 2.76	68.95 $\pm$ 3.63	2.55 $\pm$ 0.00	77.90 $\pm$ 0.86	77.61 $\pm$ 1.12	0.63 $\pm$ 0.00	<u>57.89</u> $\pm$ 0.74	49.94 $\pm$ 1.70	0.00 $\pm$ 0.00
LS	50.16 $\pm$ 5.95	55.88 $\pm$ 2.52	4.46 $\pm$ 1.65	70.90 $\pm$ 2.67	70.41 $\pm$ 4.52	2.54 $\pm$ 0.82	57.11 $\pm$ 2.99	49.84 $\pm$ 2.35	0.00 $\pm$ 0.00
UDFS	45.73 $\pm$ 4.02	69.46 $\pm$ 3.36	5.06 $\pm$ 1.12	70.75 $\pm$ 2.91	71.03 $\pm$ 6.07	1.02 $\pm$ 0.77	57.23 $\pm$ 3.16	37.23 $\pm$ 2.97	2.45 $\pm$ 0.37
NDFS	<u>67.68</u> $\pm$ 0.85	<u>73.70</u> $\pm$ 0.71	5.42 $\pm$ 0.48	77.61 $\pm$ 0.72	<u>79.40</u> $\pm$ 1.72	1.20 $\pm$ 0.81	57.44 $\pm$ 2.40	52.85 $\pm$ 0.34	0.70 $\pm$ 1.11
SPUFS	60.28 $\pm$ 1.81	63.24 $\pm$ 2.45	5.17 $\pm$ 0.45	70.25 $\pm$ 2.24	72.81 $\pm$ 11.14	1.86 $\pm$ 1.25	53.47 $\pm$ 4.66	51.47 $\pm$ 0.32	0.06 $\pm$ 0.01
LDSSL	52.31 $\pm$ 5.19	64.64 $\pm$ 1.66	5.66 $\pm$ 0.13	76.41 $\pm$ 1.26	77.26 $\pm$ 3.37	1.67 $\pm$ 0.39	56.96 $\pm$ 2.52	51.64 $\pm$ 0.48	0.22 $\pm$ 0.20
MaxVar	59.90 $\pm$ 2.94	66.97 $\pm$ 6.18	4.71 $\pm$ 0.54	68.00 $\pm$ 3.23	73.13 $\pm$ 4.77	0.63 $\pm$ 0.00	53.03 $\pm$ 4.79	22.39 $\pm$ 2.69	0.00 $\pm$ 0.00
LLCFS	65.14 $\pm$ 2.65	62.88 $\pm$ 6.15	5.48 $\pm$ 0.46	73.50 $\pm$ 2.13	72.85 $\pm$ 3.83	0.65 $\pm$ 0.02	52.65 $\pm$ 4.24	28.89 $\pm$ 1.81	0.07 $\pm$ 0.14
TraceRatio	40.45 $\pm$ 6.95	66.35 $\pm$ 3.52	5.22 $\pm$ 0.68	70.40 $\pm$ 3.13	59.62 $\pm$ 11.63	0.63 $\pm$ 0.00	54.14 $\pm$ 4.13	50.16 $\pm$ 1.19	0.47 $\pm$ 0.57
MCFS	61.99 $\pm$ 3.04	71.50 $\pm$ 3.07	3.28 $\pm$ 1.41	76.73 $\pm$ 0.66	61.48 $\pm$ 8.18	1.36 $\pm$ 0.73	57.58 $\pm$ 2.07	38.10 $\pm$ 8.22	0.18 $\pm$ 0.13
SCUFS	59.44 $\pm$ 3.89	71.80 $\pm$ 1.98	5.89 $\pm$ 0.84	76.31 $\pm$ 1.25	75.07 $\pm$ 6.11	<u>2.97</u> $\pm$ 0.82	57.38 $\pm$ 2.18	<u>52.94</u> $\pm$ 0.35	<u>2.46</u> $\pm$ 0.45
SCFS	70.17 $\pm$ 0.90	73.73 $\pm$ 0.75	<u>5.85</u> $\pm$ 0.46	<u>77.71</u> $\pm$ 0.44	79.43 $\pm$ 1.62	3.73 $\pm$ 0.50	58.59 $\pm$ 1.91	53.70 $\pm$ 0.39	2.51 $\pm$ 0.42

**Fig. 3.** The obtained results in terms of Acc with different numbers of selected features.

- Clustering based methods such as NDFS, SCUFS and SCFS commonly attain better results in an unsupervised manner.
- SCFS achieves the best performance according to Acc on the eight datasets.
- The obtained NMI results indicate the superiority of SCFS as compared with other methods.
- Moreover, SCFS outperforms the earlier subspace learning based approach, LDSSL, due to employing the sample-level self-expression, and adaptive learning of the cluster similarities.

Furthermore, we demonstrate the performance of the proposed method in different number of selected features from 50 to 300. Figs. 3

and 4 represent the obtained results according to these scenarios, where the three methods including LS, MaxVar, and TraceRatio are ignored due to weak results. On the one hand, SCFS obtained satisfactory performance with the small number of selected features. On the other hand, these results indicate that the proposed approach attains better performance than its well-known competitors on by varying the number of selected features.

In addition, the results based on Coefficient of Variation (CV) is reported in Tables 5 and 6. By average, SCFS achieves the best robustness in CV using Acc and NMI measures.

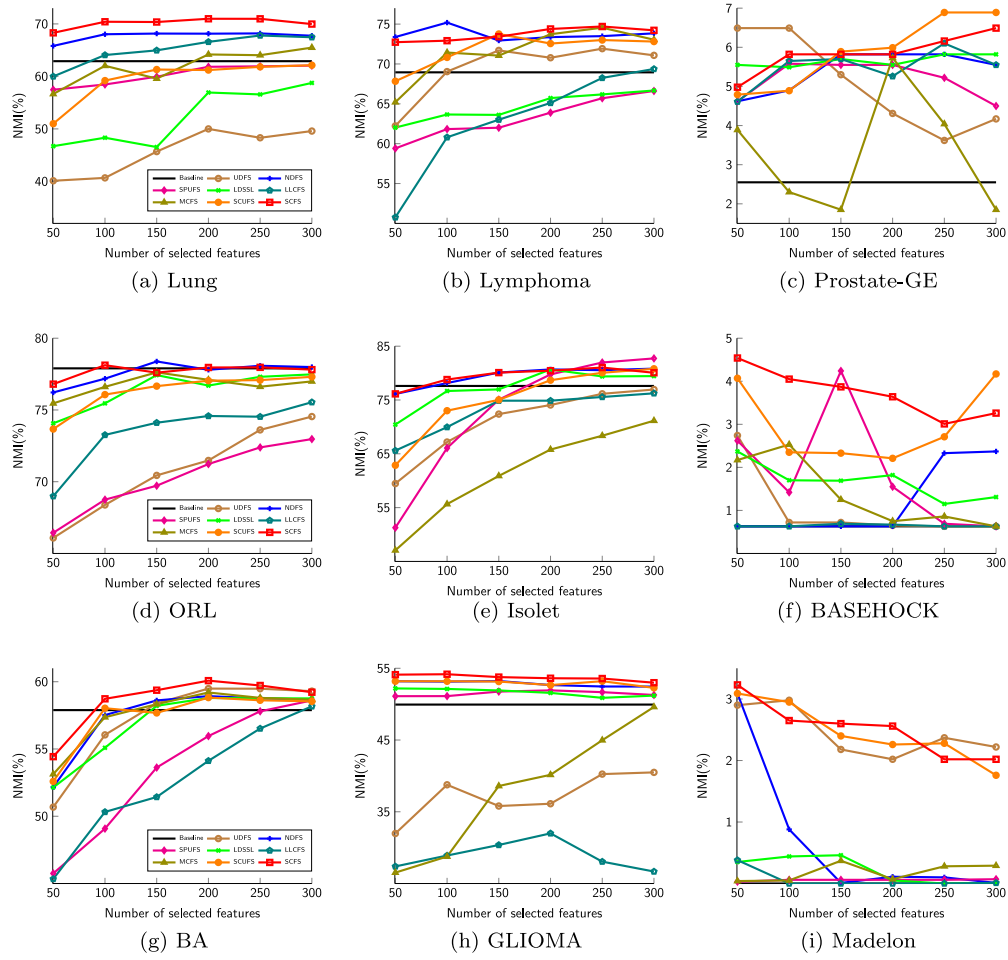


Fig. 4. The obtained results in terms of NMI with different numbers of selected features.

Table 5

Coefficient of variation (CV%) using Acc. Bold and underlined numbers are the best and the second best on average.

Dataset	Lung	Lymphoma	Prostate-GE	ORL	Isolet	BASEHOCK	BA	GLIOMA	Madelon	Average
Baseline	0.10	0.09	0.00	0.04	0.03	0.00	0.04	0.06	0.00	0.04
LS	0.04	0.05	0.03	0.07	0.08	0.00	0.08	0.04	0.00	0.05
UDFS	0.09	0.04	0.02	0.08	0.11	0.01	0.08	0.04	0.01	0.05
NDFS	0.01	0.01	0.01	0.02	0.02	0.00	0.07	0.06	0.07	<u>0.03</u>
SPUFS	0.02	0.06	0.02	0.06	0.21	0.01	0.14	0.01	0.00	0.06
LDSSL	0.08	0.03	0.01	0.03	0.06	0.00	0.07	0.01	0.03	0.04
MaxVar	0.05	0.12	0.01	0.10	0.09	0.00	0.12	0.03	0.00	0.06
LLCFS	0.05	0.07	0.01	0.06	0.08	0.00	0.10	0.02	0.02	0.05
TraceRatio	0.10	0.08	0.01	0.09	0.29	0.00	0.13	0.06	0.04	0.09
MCFS	0.04	0.04	0.02	0.01	0.17	0.01	0.05	0.07	0.02	0.05
SCUFS	0.07	0.01	0.01	0.03	0.12	0.01	0.04	0.02	0.03	0.04
SCFS	0.02	0.02	0.01	0.01	0.01	0.01	0.04	0.03	0.03	0.02

Table 6

Coefficient of variation (CV%) using NMI. Bold and underlined numbers are the best and the second best on average.

Dataset	Lung	Lymphoma	Prostate-GE	ORL	Isolet	BASEHOCK	BA	GLIOMA	Madelon	Average
Baseline	0.04	0.05	0.00	0.01	0.01	0.00	0.01	0.03	0.53	<u>0.08</u>
LS	0.12	0.05	0.37	0.04	0.06	0.32	0.05	0.05	0.22	0.14
UDFS	0.09	0.05	0.22	0.04	0.09	0.76	0.06	0.08	0.15	0.17
NDFS	0.01	0.01	0.09	0.01	0.02	0.67	0.04	0.01	1.59	0.27
SPUFS	0.03	0.04	0.09	0.03	0.15	0.67	0.09	0.01	0.22	0.15
LDSSL	0.10	0.03	0.02	0.02	0.04	0.23	0.04	0.01	0.93	0.16
MaxVar	0.05	0.09	0.11	0.05	0.07	0.00	0.09	0.12	0.19	0.09
LLCFS	0.04	0.10	0.08	0.03	0.05	0.03	0.08	0.06	2.13	0.29
TraceRatio	0.17	0.05	0.13	0.04	0.20	0.00	0.08	0.02	1.23	0.21
MCFS	0.05	0.04	0.43	0.01	0.13	0.53	0.04	0.22	0.73	0.24
SCUFS	0.07	0.03	0.14	0.02	0.08	0.28	0.04	0.01	0.18	0.09
SCFS	0.01	0.01	0.08	0.01	0.02	0.14	0.03	0.01	0.17	0.05

Table 7

The Stability index (SI) from 10 runs by perturbing the 20% of the samples.

Dataset	Lung	Lymphoma	Prostate-GE	ORL	Isolet	BASEHOCK	BA	GLIOMA	Madelon	Average
LS	0.59	1.00	0.15	1.00	0.68	0.50	0.64	0.32	1.00	0.65
UDFS	0.30	0.09	0.16	0.38	0.57	0.24	0.28	0.24	0.58	0.31
NDFS	0.88	0.82	0.90	0.85	0.85	0.87	0.82	0.85	0.85	0.86
SPUFS	0.37	0.23	0.36	0.17	0.07	0.48	0.05	0.30	0.15	0.24
LDSSL	0.98	0.00	0.06	0.97	0.02	0.27	0.89	0.32	1.00	0.50
MaxVar	0.99	0.92	0.98	0.97	0.99	0.97	0.98	0.97	1.00	0.98
LLCFS	0.79	0.38	0.37	0.79	0.90	0.63	0.65	0.67	0.96	0.68
TraceRatio	0.99	0.88	0.99	0.99	0.99	0.93	0.98	0.99	1.00	0.97
MCFS	0.42	0.26	0.45	0.46	0.35	1.00	0.20	0.36	0.43	0.44
SCUFS	0.68	0.50	0.51	0.39	0.51	0.43	0.61	0.53	0.48	0.52
SCFS	0.80	0.76	0.77	0.77	0.81	0.82	0.86	0.80	0.86	0.81

Table 8

The running time of SCFS compared with the other methods (in seconds).

Dataset	Lung	Lymphoma	Prostate-GE	ORL	Isolet	BASEHOCK	BA	GLIOMA	Madelon	Average
LS	0.39	0.41	33.15	5.10	38.67	4.59	0.54	0.15	25.03	12.00
UDFS	821.77	2445.84	22720.17	62.17	133.99	3403.93	35.73	2978.69	270.80	3652.57
NDFS	1715.43	542.29	15616.99	107.62	72.55	601.63	14.31	612.98	255.38	2171.02
SPUFS	17.41	12.22	19.59	49.60	0.12	109.32	0.02	9.84	84.94	33.67
LDSSL	195.32	7184.70	1548.83	642.50	710.74	6125.77	15.57	312.33	9.84	1860.62
MaxVar	0.02	0.03	0.64	0.63	3.72	0.08	0.03	0.03	1.44	0.73
LLCFS	20.19	9.31	24.90	35.27	359.88	852.95	82.76	4.11	1375.66	307.23
TraceRatio	37.48	10.91	132.23	40.03	17.60	65.67	1.69	10.30	13.04	36.55
MCFS	0.08	0.04	12.58	0.07	0.82	3.65	0.27	0.02	7.11	2.74
SCUFS	76.52	3.13	15.38	141.74	556.75	1559.20	190.71	1.61	1763.96	478.78
SCFS	2.35	142.86	26.32	0.35	10.67	7.16	21.53	8.02	2.68	24.66

The stability of different used methods is evaluated by SI measure with 10 runs and the 20% perturbation applied on the samples. Table 7 presents SI results. Although, SCFS is not the best one based on SI measure, but it generally stands in the four best stable methods on average.

Finally, we investigate the running time of the proposed method compared to the other methods. The running time of different methods are reported in Table 8. Although, the running time of SCFS is not the best, but the average time cost is in the top four methods.

The proposed approach resulted in better performance based on a variety of evaluation measures due to the following considerations,

- Implicitly learn the similarity matrix to reduce the search space.
- Adaptively compute the sample similarities to prevent suboptimal result.
- Perform clustering to learn underlying structure of data.

6.5. Parameter sensitivity and convergence study

First, the sensitivity of parameters α and β in our model are investigated. The experimental results on Acc and NMI criteria for all of datasets are presented on Figs. 5 and 6. For all candidate of α and β parameters, the logarithms base 10 is taken. Figs. 5 and 6 demonstrate that there is a slight sensitivity to these parameters. By considering the results on the different parameters, we set $\alpha = 1$ and $\beta = 10^2$ as the default values. Furthermore, Fig. 7 represents the performance results based on the default and their best parameters. As shown in Fig. 7, using the default parameters setting, the proposed method obtains relatively good results in comparison to the best parameters setting, except on GLIOMA dataset in NMI measure with a significant gap between the default and the best parameters setting. This may be due to the mechanism of selecting default parameters which is based on the majority process among all nine datasets. The default parameters, $\alpha = 1$ and $\beta = 10^2$, are resulted in overall good results on all of datasets except NMI on GLIOMA dataset. Therefore, the proposed method can be practically applied based on the default parameters setting.

Next, we experimentally study the convergence behavior of the proposed algorithm. Fig. 8 presents the speed of the convergence according to the objective values with respect to the number of iterations on different datasets. The stopping criteria is set as $\frac{|obj(t) - obj(t-1)|}{obj(t)} < 10^{-5}$, where $obj(t)$ is the objective function value of Eq. (3) in the t th iteration. As shown in the Fig. 8, the proposed algorithm monotonically decreases the objective function in a few iteration. A small point is that the proposed algorithm is converged too slowly in lymphoma dataset as compared with the other employed datasets.

6.6. Selected features

Here, we investigate the selected features obtained by the 11 methods (LS, UDFS, NDFS, SPUFS, LDSSL, MaxVar, LLCFS, TraceRatio, MCFS, SCUFS, and SCFS), when 300 features are selected. Table 9 shows the most commonly selected features by the different applied methods. This can be applied to illustrate features are providing more advantage rather than the others.

The overlapping feature space among the methods are represented in Fig. 9. Part (a) to (i) represent the number of common selected features of the employed methods on the datasets of Table 2. Furthermore, part (j) of Fig. 9 is dedicated to the average number of overlapping features. It can be observed from Fig. 9 that there are similarities among the proposed method, SCFS, and state-of-the-art methods, UDFS, MaxVar, SCUFS, LLCFS, and SPUFS on choosing the common features. This fact indicates the reliability and confidence of the SCFS that can be applied on practical situations.

7. Conclusion

In this paper, we proposed a novel unsupervised feature selection framework initiated from the subspace learning and regularized regression to maintain sample similarities and take underlying structure into account in the selected features. The proposed method, SCFS, was designed to implicitly learn the cluster similarities in an adaptive manner. Furthermore, a unified objective function was constituted

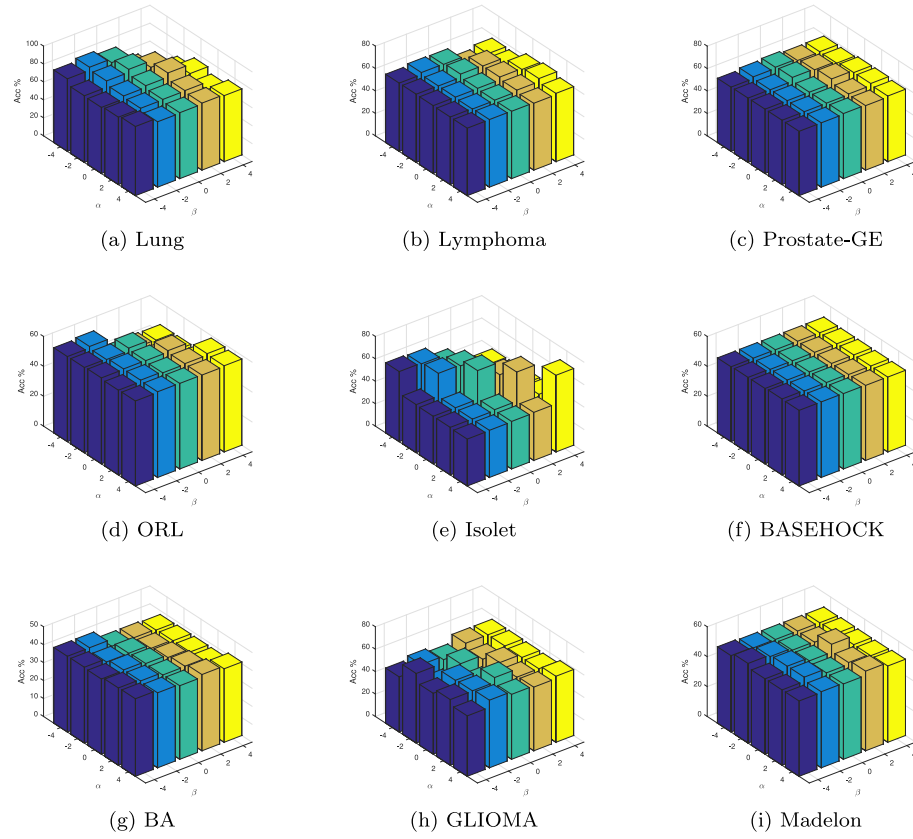


Fig. 5. The performance of SCFS in terms of Acc using different values of the parameters α and β .

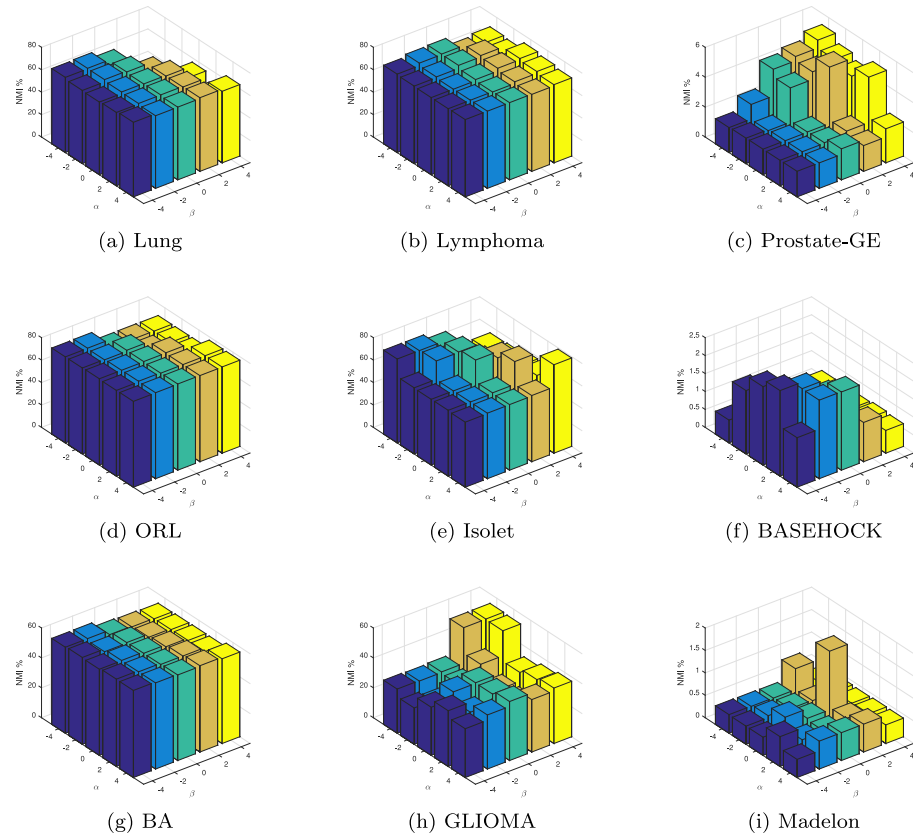


Fig. 6. The performance of SCFS in terms of NMI using different values of the parameters α and β .

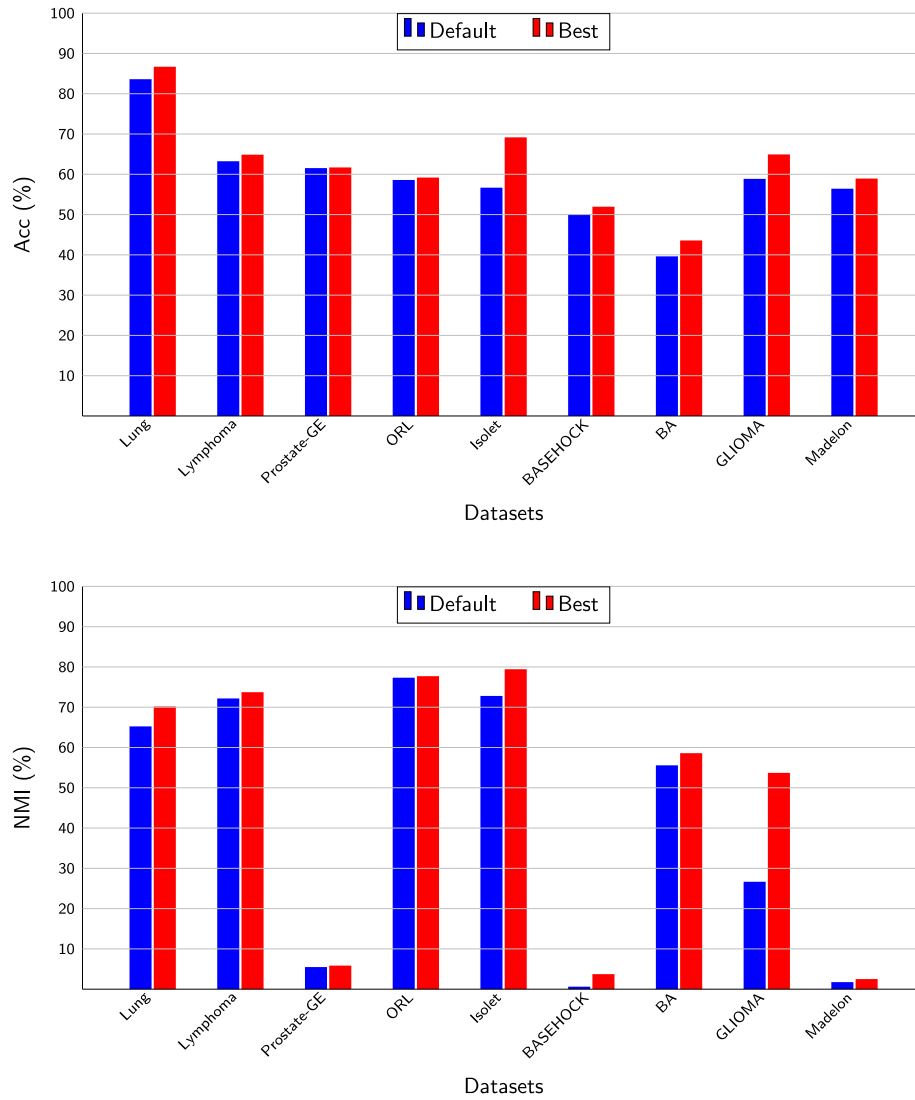


Fig. 7. Comparison between the default parameters setting ($\alpha = 1$ and $\beta = 10^2$) and the best setting.

Table 9

Selected features by the most methods on standard datasets.

Dataset	Selected features	#Methods
Lung	597, 775, 2110, 2579, 2750, 3021, 3128	8
Lymphoma	1103, 1141	7
Prostate-GE	159, 1362, 1828, 1992, 2594, 2984, 3724, 4486	6
ORL	22, 46, 57, 65, 100, 188, 200, 262, 530, 531, 562, 993, 1017	8
Isolet	20, 70, 83, 462	10
BASEHOCK	250, 483, 769, 1184, 1366, 1722, 2005, 2472, 2631, 2965, 3038, 3233, 3302, 4428, 4806	6
BA	22, 23, 26, 27, 36, 37, 38, 39, 40, 48, 49, 50, 51, 52, 53, 54, 55, 56, 59, 60, 62, 63, 67, 69, 71, 72, 74, 75, 76, 77, 79, 80, 82, 83, 87, 89, 90, 92, 93, 94, 97, 99, 100, 103, 105, 106, 107, 108, 109, 110, 115, 118, 119, 121, 123, 124, 125, 126, 127, 128, 129, 131, 134, 135, 136, 137, 138, 139, 141, 142, 144, 145, 147, 150, 152, 155, 156, 157, 158, 159, 164, 165, 168, 169, 170, 175, 176, 177, 178, 180, 183, 184, 185, 186, 187, 188, 189, 191, 192, 194, 195, 196, 197, 198, 199, 200, 202, 203, 204, 205, 206, 207, 208, 210, 212, 213, 214, 215, 216, 217, 218, 219, 221, 222, 223, 225, 226, 227, 228, 229, 231, 234, 235, 236, 239, 240, 241, 243, 245, 246, 247, 249, 251, 252, 253, 255, 258, 259, 260, 263, 265, 266, 267, 269, 270, 271, 272, 273, 274, 276, 278, 280, 282, 283, 284, 285, 286, 287, 289, 290, 291, 292, 293, 295, 296, 297, 298, 299, 300, 301, 303, 304, 305, 306, 307, 312, 313, 315, 316, 317, 318, 319	11
GLIOMA	3707, 4120, 4390	8
Madelon	80, 158, 481	11

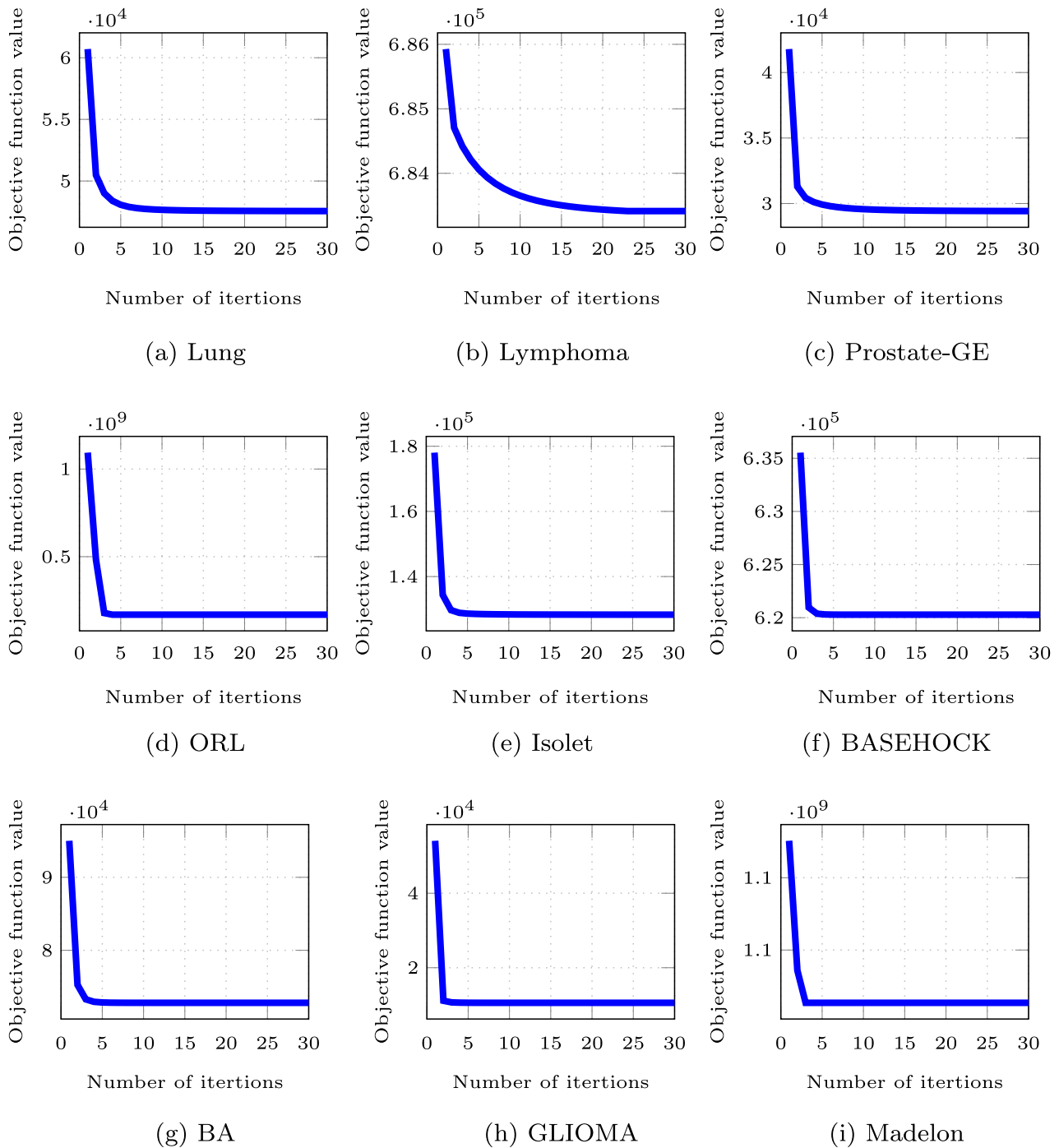


Fig. 8. Convergence curve of SCFS on different datasets.

from the main underlying characteristics of the proposed method. The optimization algorithm was proposed to obtain the solutions in an efficient way. In line with the computational complexity of the proposed algorithm, its convergence was investigated through an empirical study on real datasets. Extensive experiments on a variety of datasets was performed to show the effectiveness of the proposed method. Furthermore, the robustness and stability of the proposed approach were approved through a variety of standard evaluation measures.

CRediT authorship contribution statement

Mohsen Ghassemi Parsa: Conception and design of study, Acquisition of data, Analysis and/or interpretation of data, Writing - original

draft, Writing - review & editing. **Hadi Zare:** Conception and design of study, Analysis and/or interpretation of data, Writing - original draft, Writing - review & editing. **Mehdi Ghatte:** Conception and design of study, Analysis and/or interpretation of data, Writing - original draft, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Method	LS	UDFS	NDFS	SPUFS	LDSSL	MaxVar	LLCFS	TraceRatio	MCFS	SCUFS	SCFS
LS	300										
UDFS	13	300									
NDFS	16	6	300								
SPUFS	11	169	4	300							
LDSSL	64	80	0	82	300						
MaxVar	72	119	5	168	74	300					
LLCFS	66	125	5	171	83	228	300				
TraceRatio	17	0	46	0	0	0	0	300			
MCFS	31	59	10	68	39	69	63	1	300		
SCUFS	17	127	8	168	88	116	115	0	57	300	
SCFS	55	105	5	127	81	143	150	1	67	108	300

(a) Lung

Method	LS	UDFS	NDFS	SPUFS	LDSSL	MaxVar	LLCFS	TraceRatio	MCFS	SCUFS	SCFS
LS	300										
UDFS	0	300									
NDFS	12	31	300								
SPUFS	0	77	55	300							
LDSSL	13	23	30	25	300						
MaxVar	75	30	24	56	30	300					
LLCFS	14	43	27	57	30	80	300				
TraceRatio	28	13	34	11	25	16	22	300			
MCFS	5	44	28	42	26	26	26	17	300		
SCUFS	16	32	71	68	26	33	23	82	36	300	
SCFS	5	46	48	58	28	27	39	15	33	27	300

(b) Lymphoma

Method	LS	UDFS	NDFS	SPUFS	LDSSL	MaxVar	LLCFS	TraceRatio	MCFS	SCUFS	SCFS
LS	300										
UDFS	0	300									
NDFS	1	58	300								
SPUFS	2	112	110	300							
LDSSL	0	2	0	0	300						
MaxVar	55	46	48	66	0	300					
LLCFS	1	41	38	55	3	46	300				
TraceRatio	60	5	4	1	0	46	36	300			
MCFS	0	22	14	22	47	7	22	10	300		
SCUFS	0	81	88	160	0	44	46	0	23	300	
SCFS	1	45	38	58	0	55	29	2	15	42	300

(c) Prostate-GE

Method	LS	UDFS	NDFS	SPUFS	LDSSL	MaxVar	LLCFS	TraceRatio	MCFS	SCUFS	SCFS
LS	300										
UDFS	46	300									
NDFS	52	102	300								
SPUFS	47	94	109	300							
LDSSL	52	95	106	154	300						
MaxVar	54	109	88	159	229	300					
LLCFS	13	96	101	170	182	202	300				
TraceRatio	200	64	52	31	4	8	2	300			
MCFS	66	97	126	113	93	94	99	62	300		
SCUFS	55	106	112	96	93	90	92	52	105	300	
SCFS	55	107	117	116	92	93	97	61	147	111	300

(d) ORL

Method	LS	UDFS	NDFS	SPUFS	LDSSL	MaxVar	LLCFS	TraceRatio	MCFS	SCUFS	SCFS
LS	300										
UDFS	184	300									
NDFS	189	250	300								
SPUFS	93	29	27	300							
LDSSL	141	132	144	153	300						
MaxVar	242	145	154	137	149	300					
LLCFS	227	144	149	141	147	274	300				
TraceRatio	121	181	175	117	138	86	90	300			
MCFS	162	199	200	85	139	141	143	159	300		
SCUFS	187	233	232	56	138	149	150	165	198	300	
SCFS	171	185	184	97	139	147	145	165	167	180	300

(e) Isolet

Fig. 9. The number of common selected features of the employed methods in our study on the different datasets from part (a) to part (i). The average number of overlapping on all of datasets is shown in part (j).

Method	LS	UDFS	NDFS	SPUFS	LDSSL	MaxVar	LLCFS	TraceRatio	MCFS	SCUFS	SCFS
LS	300										
UDFS	22	300									
NDFS	9	23	300								
SPUFS	27	7	6	300							
LDSSL	26	19	12	72	300						
MaxVar	56	23	3	194	89	300					
LLCFS	25	12	1	199	90	197	300				
TraceRatio	110	4	2	32	32	30	31	300			
MCFS	20	17	16	13	12	14	12	14	300		
SCUFS	17	30	26	3	17	5	4	5	16	300	
SCFS	26	47	12	48	42	70	57	42	13	19	300

(f) BASEHOCK

Method	LS	UDFS	NDFS	SPUFS	LDSSL	MaxVar	LLCFS	TraceRatio	MCFS	SCUFS	SCFS
LS	300										
UDFS	284	300									
NDFS	283	282	300								
SPUFS	280	280	280	300							
LDSSL	286	280	282	281	300						
MaxVar	293	280	281	280	286	300					
LLCFS	289	280	280	283	284	294	300				
TraceRatio	298	284	282	280	286	294	289	300			
MCFS	281	283	284	280	280	280	280	281	300		
SCUFS	282	281	282	281	280	281	281	282	284	300	
SCFS	284	283	281	280	281	282	281	283	282	283	300

(g) BA

Method	LS	UDFS	NDFS	SPUFS	LDSSL	MaxVar	LLCFS	TraceRatio	MCFS	SCUFS	SCFS
LS	300										
UDFS	2	300									
NDFS	2	94	300								
SPUFS	1	154	128	300							
LDSSL	0	16	20	8	300						
MaxVar	8	113	88	152	0	300					
LLCFS	9	111	91	146	0	228	300				
TraceRatio	85	0	0	0	0	0	0	300			
MCFS	2	65	59	76	48	66	70	1	300		
SCUFS	0	132	103	184	21	110	105	0	67	300	
SCFS	6	66	93	109	59	102	94	0	57	90	300

(h) GLIOMA

Method	LS	UDFS	NDFS	SPUFS	LDSSL	MaxVar	LLCFS	TraceRatio	MCFS	SCUFS	SCFS
LS	300										
UDFS	124	300									
NDFS	132	220	300								
SPUFS	274	118	130	300							
LDSSL	183	183	172	182	300						
MaxVar	285	120	133	281	182	300					
LLCFS	287	120	133	281	181	296	300				
TraceRatio	100	240	231	113	175	100	100	300			
MCFS	115	221	222	111	180	116	116	246	300		
SCUFS	110	236	229	104	178	108	107	257	238	300	
SCFS	150	198	213	145	180	152	151	209	212	202	300

(i) Madelon

Method	LS	UDFS	NDFS	SPUFS	LDSSL	MaxVar	LLCFS	TraceRatio	MCFS	SCUFS	SCFS
LS	300										
UDFS	75	300									
NDFS	77	118	300								
SPUFS	82	116	94	300							
LDSSL	85	92	85	106	300						
MaxVar	127	109	92	166	115	300					
LLCFS	103	108	92	167	111	205	300				
TraceRatio	113	88	92	65	73	64	63	300			
MCFS	76	112	107	90	96	90	92	88	300		
SCUFS	76	140	128	124	93	104	103	94	114	300	
SCFS	84	120	110	115	100	119	116	86	110	118	300

(j) Average

Fig. 9. (continued).

References

- Abpeykar, S., Ghatsee, M., Zare, H., 2019. Ensemble decision forest of RBF networks via hybrid feature clustering approach for high-dimensional data classification. *Comput. Statist. Data Anal.* 131, 12–36.
- Alirezanejad, M., Enayatifar, R., Motameni, H., Nematzadeh, H., 2020. Heuristic filter feature selection methods for medical datasets. *Genomics* 112 (2), 1173–1181.
- Alizadeh, A.A., et al., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403 (6769), 503–511.
- Bach, F., Jenatton, R., Mairal, J., Obozinski, G., 2012. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.* 4 (1), 1–106.
- Belhumeur, P., Hespanha, J., Kriegman, D., 1997. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* 19 (7), 711–720.
- Bhattacharjee, A., et al., 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci.* 98 (24), 13790–13795.
- Bishop, C.M., 2006. *Pattern Recognition and Machine Learning* (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.
- Cai, D., He, X., Han, J., Zhang, H.-J., 2006. Orthogonal Laplacianfaces for face recognition. *IEEE Trans. Image Process.* 15 (11), 3608–3614.
- Cai, D., Zhang, C., He, X., 2010. Unsupervised feature selection for multi-cluster data. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. In: KDD '10, ACM, New York, NY, USA, pp. 333–342.
- Cole, R., Fanty, M., 1990. Spoken letter recognition. In: *Proceedings of the Workshop on Speech and Natural Language*. In: HLT '90, Association for Computational Linguistics, Hidden Valley, Pennsylvania, pp. 385–390.
- Ding, C.H.Q., 2003. Unsupervised feature selection via two-way ordering in gene expression analysis. *Bioinformatics* 19 (10), 1259.
- Du, L., Shen, Y.-D., 2015. Unsupervised feature selection with adaptive structure learning. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. In: KDD '15, ACM, New York, NY, USA, pp. 209–218.
- Dy, J.G., Brodley, C.E., 2004. Feature selection for unsupervised learning. *J. Mach. Learn. Res.* 5, 845–889.
- Guyon, I. (Ed.), 2006. *Feature extraction: foundations and applications*. In: *Studies in Fuzziness and Soft Computing*, no. 207, Springer, Berlin.
- Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.
- Hall, M.A., Smith, L.A., 1999. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In: *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*. AAAI Press, pp. 235–239.
- Han, X., Liu, P., Wang, L., Li, D., 2020. Unsupervised feature selection via graph matrix learning and the low-dimensional space learning for classification. *Eng. Appl. Artif. Intell.* 87, 103283.
- He, X., Cai, D., Niyogi, P., 2005. Laplacian score for feature selection. In: *Proceedings of the 18th International Conference on Neural Information Processing Systems*. In: NIPS'05, MIT Press, Cambridge, MA, USA, pp. 507–514.
- Hou, C., Nie, F., Li, X., Yi, D., Wu, Y., 2014. Joint embedding learning and sparse regression: A framework for unsupervised feature selection. *IEEE Trans. Cybern.* 44 (6), 793–804.
- Huan Liu, Setiono, R., 1995. Chi2: feature selection and discretization of numeric attributes. In: *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*. pp. 388–391.
- Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. *Artificial Intelligence* 97 (1–2), 273–324.
- Krzanowski, W.J., 1987. Selection of variables to preserve multivariate data structure, using principal components. *J. R. Stat. Soc. Ser. C. Appl. Stat.* 36 (1), 22–33.
- Kuhn, H.W., Tucker, A.W., 2014. Nonlinear programming. In: Giorgi, G., Kjeldsen, T.H. (Eds.), *Traces and Emergence of Nonlinear Programming*. Springer Basel, Basel, pp. 247–258.
- Kuncheva, L.I., 2007. A stability index for feature selection. In: *Proceedings of the 25th Conference on Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications*. In: AIAP'07, ACTA Press, Innsbruck, Austria, pp. 390–395.
- Lee, D.D., Seung, H.S., 2000. Algorithms for non-negative matrix factorization. In: *Proceedings of the 13th International Conference on Neural Information Processing Systems*. In: NIPS'00, MIT Press, Cambridge, MA, USA, pp. 535–541.
- Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H., 2017a. Feature selection: A data perspective. *ACM Comput. Surv.* 50 (6), 94:1–94:45, <http://featureselection.asu.edu/>.
- Li, Z., Liu, J., Yang, Y., Zhou, X., Lu, H., 2014. Clustering-guided sparse structural learning for unsupervised feature selection. *IEEE Trans. Knowl. Data Eng.* 26 (9), 2138–2150.
- Li, J., Tang, J., Liu, H., 2017b. Reconstruction-based unsupervised feature selection: An embedded approach. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. In: IJCAI'17, AAAI Press, pp. 2159–2165.
- Li, Z., Yang, Y., Liu, J., Zhou, X., Lu, H., 2012. Unsupervised feature selection using nonnegative spectral analysis. In: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. In: AAAI'12, AAAI Press, pp. 1026–1032.
- Liu, J., Ji, S., Ye, J., 2009. Multi-task feature learning via efficient L2, 1-norm minimization. In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. In: UAI '09, AUAI Press, Arlington, Virginia, United States, pp. 339–348.
- Liu, T., Liu, S., Chen, Z., Ma, W.-Y., 2003. An evaluation on feature selection for text clustering. In: *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*. In: ICML'03, AAAI Press, pp. 488–495.
- Liu, X., Wang, L., Zhang, J., Yin, J., Liu, H., 2014. Global and local structure preservation for feature selection. *IEEE Trans. Neural Netw. Learn. Syst.* 25 (6), 1083–1095.
- Lovasz, L., 1986. *Matching Theory* (North-Holland Mathematics Studies). Elsevier Science Ltd., Oxford, UK, UK.
- Lu, Q., Li, X., Dong, Y., 2018. Structure preserving unsupervised feature selection. *Neurocomputing* 301 (C), 36–45.
- Masaeli, M., Yan, Y., Cui, Y., Fung, G., Dy, J., 2010. Convex principal feature selection. In: *Proceedings of the 2010 SIAM International Conference on Data Mining*. In: *Proceedings, Society for Industrial and Applied Mathematics*, pp. 619–628.
- Moradi, P., Rostami, M., 2015. A graph theoretic approach for unsupervised feature selection. *Eng. Appl. Artif. Intell.* 44, 33–45.
- Nematzadeh, H., Enayatifar, R., Mahmud, M., Akbari, E., 2019. Frequency based feature selection method using whale algorithm. *Genomics* 111 (6), 1946–1955.
- Nie, F., Huang, H., Cai, X., Ding, C., 2010. Efficient and robust feature selection via joint L2,1-norms minimization. In: *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*. In: NIPS'10, Curran Associates Inc., USA, pp. 1813–1821.
- Nie, F., Xiang, S., Jia, Y., Zhang, C., Yan, S., 2008. Trace ratio criterion for feature selection. In: *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*. In: AAAI'08, AAAI Press, Chicago, Illinois, pp. 671–676.
- Nutt, C.L., et al., 2003. Gene expression-based classification of malignant Gliomas correlates Better with Survival than Histological classification. *Cancer Res.* 63 (7), 1602–1607.
- Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (8), 1226–1238.
- Roth, V., Lange, T., 2003. *Feature Selection in Clustering Problems*. MIT Press, pp. 473–480.
- Shang, R., Meng, Y., Wang, W., Shang, F., Jiao, L., 2019. Local discriminative based sparse subspace learning for feature selection. *Pattern Recognit.* 92, 219–230.
- Shang, R., Wang, W., Stolkin, R., Jiao, L., 2016. Subspace learning-based graph regularized feature selection. *Knowl.-Based Syst.* 112 (C), 152–165.
- Singh, D., et al., 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1 (2), 203–209.
- Solorio-Fernández, S., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., 2019. A review of unsupervised feature selection methods. *Artif. Intell. Rev.* 1–42.
- Soltanolkotabi, M., Elhamifar, E., Candès, E.J., 2014. Robust subspace clustering. *Ann. Statist.* 42 (2), 669–699.
- Tabakhi, S., Moradi, P., Akhlaghian, F., 2014. An unsupervised feature selection algorithm based on ant colony optimization. *Eng. Appl. Artif. Intell.* 32, 112–123.
- Tang, J., Liu, H., 2014. An unsupervised feature selection framework for social media data. *IEEE Trans. Knowl. Data Eng.* 26 (12), 2914–2927.
- Vidal, R., 2011. Subspace clustering. *IEEE Signal Process. Mag.* 28 (2), 52–68.
- Wang, S., Pedrycz, W., Zhu, Q., Zhu, W., 2015. Subspace learning for unsupervised feature selection via matrix factorization. *Pattern Recognit.* 48 (1), 10–19.
- Yang, Y., Shen, H.T., Ma, Z., Huang, Z., Zhou, X., 2011. L2,1-norm regularized discriminative feature selection for unsupervised learning. In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Two*. In: IJCAI'11, AAAI Press, pp. 1589–1594.
- Zare, H., Niazi, M., 2016. Relevant based structure learning for feature selection. *Eng. Appl. Artif. Intell.* 55, 93–102.
- Zeng, H., Cheung, Y.-m., 2011. Feature selection and kernel learning for local learning-based clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (8), 1532–1547.
- Zhao, Z., Liu, H., 2007. Spectral feature selection for supervised and unsupervised learning. In: *Proceedings of the 24th International Conference on Machine Learning*. In: ICML '07, ACM, New York, NY, USA, pp. 1151–1157.
- Zhao, Z., Wang, L., Liu, H., Ye, J., 2013. On similarity preserving feature selection. *IEEE Trans. Knowl. Data Eng.* 25 (3), 619–632.
- Zhu, P., Xu, Q., Hu, Q., Zhang, C., 2018. Co-regularized unsupervised feature selection. *Neurocomputing* 275 (C), 2855–2863.
- Zhu, P., Zhu, W., Hu, Q., Zhang, C., Zuo, W., 2017. Subspace clustering guided unsupervised feature selection. *Pattern Recognit.* 66 (C), 364–374.
- Zhu, P., Zuo, W., Zhang, L., Hu, Q., Shiu, S.C., 2015. Unsupervised feature selection by regularized self-representation. *Pattern Recognit.* 48 (2), 438–446.