IJICTR
International Journal of Information & Communication Technology Research

Volume 12- Number 3 – 2020 (38 -46)

Unsupervised Feature Selection Based on Low Dimensional Embedding and Subspace Learning

Hadi Zare*
Faculty of New Sciences and Technologies
University of Tehran
Tehran, Iran
h.zare@ut.ac.ir

Mohsen Ghasemi Parsa
Faculty of New Sciences and Technologies
University of Tehran
Tehran, Iran
mgparsa@ut.ac.ir

Mehdi Ghatee
Department of Computer Science
Amirkabir University of Technology
Tehran, Iran
ghatee@aut.ac.ir

Sasan H. Alizadeh
Department of Information Technology
IRAN Telecommunication Research Center
Tehran, Iran
s.alizadeh@itrc.ac.ir

Received: 15 March 2020 - Accepted: 7 May 2020

Abstract— **Nowadays, we face a huge number of high-dimensional data in different applications and technologies. To tackle the challenge, various feature selection methods have been recently proposed for reducing the computational complexity of the learning algorithms as well as simplifying the learning models. Maintaining the geometric structures and considering the discriminative information in data are two important factors that should be borne in mind particularly for unsupervised feature selection methods. In this paper, our aim is to propose a new unsupervised feature selection approach by considering global and local similarities and discriminative information. Furthermore, this unsupervised framework incorporates cluster analysis to consider the underlying structure of the samples. Moreover, the correlation of features and clusters is computed by an $\ell_{2,1}$-norm regularized regression to eliminate the redundant and irrelevant features. Finally, a unified objective function is presented as well as an efficient iterative optimization algorithm to solve the corresponding problem with some theoretical analysis of the convergence and the complexity of the algorithm. We compare the proposed approach with the state-of-the-art method based on clustering results on the various standard datasets including biology, image, voice, and artificial data. The experimental results have presented the strength and performance improvement of the proposed method by outperforming the well-known methods.**

Keywords- Unsupervised feature selection; Similarity preserving; Low dimensional embedding; Cluster analysis; Sparse learning

I. INTRODUCTION

Machine learning algorithms suffer from the curse of dimensionality, which may exponentially reduce the performance of the learning algorithms on high-dimensional data. Furthermore, the memory and computational requirements are significantly increased on high-dimensional data [1]. This challenge can be

* Corresponding Author

addressed based on two main viewpoints, feature extraction, and feature selection. Feature extraction techniques such as PCA [2] and LDA [3], transform the original features to a new low dimensional space, commonly by a linear or non-linear mapping. Due to the creation of new features on feature extraction techniques, the physical meaning of the new feature space is not specified. On the other hand, feature selection methods not only select a subset of original features, but also provide a better interpretation in the reduced space.

Feature selection is applied to different applications including multi-view learning [4], text mining [5], and complex network [6], [7] . Feature selection approach is utilized to deal with the curse of dimensionality [2] as well as to simplify the learned models [1].

In terms of feature subset evaluation, three typical categories in feature selection methods are mentioned including wrapper, filter, and embedded [8]. Wrapper approach [9] is based on the performance of the feature subset in a learning algorithm, while the evaluation measure in filter methods [10], [11] are based on the data itself without utilizing any machine learning method. Finally, feature selection process is combined with a learning algorithm in embedded methods [12].

In the label perspective, the family of feature selection methods is also partitioned into "Supervised" and "Unsupervised" approaches [13]. The most important factor in supervised methods is to consider the correlation between the features and the labels including information theory based methods [14], [15], statistical approaches [16], and sparse learning [17]. Unsupervised feature selection has recently received much attention due to the more applicability on a wide category of domains.

The unsupervised frameworks of feature selection are mainly initiated from the innate structural characteristics of the data [18], [19]. The well-known unsupervised feature selection categories are similarity preserving [10], data reconstruction [20], and sparse learning [21], [22].

In this paper, a novel unsupervised feature selection method is proposed based on sparse learning, named "Sparse Learning and Similarity Preserving" (SLSP), which preserves the global and local similarities as well as takes the discriminative information into account by low dimensional embedding, cluster analysis, and subspace learning. The proposed method is presented by an objective function based on an $L_{2,1}$-norm regularization to eliminate the redundant features as well as selecting the relevant features. The main contributions of the paper are summarized as follows.

- Introducing a joint framework to maintain global and local similarities as well as considering the discriminative information.
- Performing cluster analysis, subspace learning, linear low dimensional transformation, regression, and regularization in a unified objective function.
- Presenting an unsupervised feature selection algorithm and some theoretical analysis to show the convergence of the optimization process.

The rest of the paper is organized as follows. In Section II, we review the related works on unsupervised feature selection. The proposed method is presented in Section III based on an optimization algorithm. The convergence analysis and the computational complexity of the proposed algorithm are discussed in Section IV. The experimental results are presented in Section V and finally Section VI concludes the paper.

II. Related Works

In this section, we first present some notations. The earlier unsupervised feature selection methods are then reviewed in three subsections, similarity-based methods, reconstruction-based methods, and sparse learning-based methods. Finally, a comparison of well-known sparse learning-based methods is given.

A. Notations

Throughout this paper, the matrices are denoted by bold uppercase and vectors by bold lowercase characters. a_{ij} means the (i,j)-th element, $\mathbf{a_i}$ denotes the i-th row, and $\mathbf{a^j}$ is the j-th column of an arbitrary matrix $\mathbf{A}$. The Frobenius norm of the matrix $\mathbf{A}$ is denoted by $\| \mathbf{A} \|_F$, trace by $\mathrm{tr}(\mathbf{A})$, and transpose by $\mathbf{A}^\top$. $\| \mathbf{v} \|_2$ is the ℓ_2-norm of a vector $\mathbf{v}$, and the $\ell_{2,1}$-norm of the matrix $\mathbf{A}$ is defined as,

$$\| \mathbf{A} \|_{2,1} = \sum_i \sqrt{\sum_j a_{ij}^2}. \quad (1)$$

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

B. Similarity-based methods

Similarity preserving methods select features based on maintaining the geometric structure in data. Although the local similarities are preserved by this approach, eliminating the redundant features is neglected by these methods.

Laplacian score (LS) [10] aims to preserve geometric structure in data based on a laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{S}$, where $\mathbf{D}$ is a diagonal matrix as $d_{ii} = \sum_j s_{ij}$ and the similarity matrix $\mathbf{S}$ is as follows,

$$s_{ij} = \begin{cases} exp\left(-\frac{\|\mathbf{x_i} - \mathbf{x_j}\|_2^2}{\sigma}\right), & \text{if } \mathbf{x_i} \in \mathrm{N}_k(\mathbf{x_j}) \text{ or } \mathbf{x_j} \in N_k(\mathbf{x_i}) \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where $\mathrm{N}_k(\mathbf{x_i})$ represents the set of k-nearest neighbors of $\mathbf{x_i}$ and σ is the width parameter. Based on the laplacian matrix $\mathbf{L}$, a score is assigned to r-th feature as follows,

$$\tilde{f}_r = f_r - \frac{f_r^\top \mathbf{D1}}{\mathbf{1}^\top \mathbf{D1}} \mathbf{1}, \quad (3)$$

where $f_r = \mathbf{x}^r$,

$$L_r = \frac{\tilde{f}_r^\top \mathbf{L} \tilde{f}_r}{\tilde{f}_r^\top \mathbf{D} \tilde{f}_r}, \quad (4)$$

where $\mathbf{1}$ is a vector filled by ones. The larger L_r, the more likely is to select r-th feature.

Spectral feature selection (SPEC) [11] is another similarity preserving method based on the concept of consistency. A feature is consistent with the graph structure if it corresponds to similar values for close

samples. SPEC introduces three formulations for ranking the features based on a normalized laplacian matrix as,

$$\mathcal{L} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{D}-\mathbf{S})\mathbf{D}^{-\frac{1}{2}}. \quad (5)$$

C. Reconstruction-based methods

A bunch of unsupervised feature selection methods are based on reconstructing the data matrix. These methods are tried to eliminate redundant features without paying attention to clustering structure. Some well-known methods based on reconstruction are described in the following. Convex principal feature selection (CPFS) [20] re-expresses the data by a regularized linear transformation as,

$$\min_{\mathbf{Q}} \| \mathbf{X}-\mathbf{X}\mathbf{Q} \|_F^2 + \lambda \sum_{i=1}^{p} \| \mathbf{q}_i \|_\infty, \quad (6)$$

where $\mathbf{Q} \in \mathbb{R}^{p\times p}$ is the reconstruction matrix and $\|.\|_\infty$ denotes the infinity norm. Greedy feature selection (GreedyFS) [23], [24] proposed an algorithm to minimize the reconstruction error based on forward selection. In reconstruction-based feature selection (REFS) [19], a new reconstruction function from data is learned, instead of utilizing a linear function. Structure preserving unsupervised feature selection (SPUFS) [25] combined the reconstruction approach to a spectral analysis to preserve local similarities as follows,

$$\min_{\mathbf{Q}} \| \mathbf{X}-\mathbf{X}\mathbf{Q} \|_F^2 + \alpha \| \mathbf{Q} \|_{2,1} + \frac{\beta}{2} \operatorname{tr}(\mathbf{Q}^\top\mathbf{X}^\top\mathbf{L}\mathbf{X}\mathbf{Q}). \quad (7)$$

D. Sparse learning-based methods

A variety of feature selection approaches are designed based on the sparse learning idea [26], [27]. There are many interesting works in this fascinating category including multi-cluster feature selection (MCFS) [28], unsupervised discriminative feature selection (UDFS) [29], nonnegative discriminative feature selection (NDFS) [21], joint embedding learning and sparse regression (JELSR) [30], [31], local discriminative based sparse subspace learning (LDSSL) [32], subspace clustering feature selection (SCFS) [22], similarity preserving feature selection (SPFS) [33] and global and local similarity preserving (GLSPFS) [34].

MCFS aims to maintain multi-cluster structure of data in the selected features by solving the following eigen-problem,

$$\mathbf{L}y = \lambda\mathbf{D}y, \quad (8)$$

where $\mathbf{Y} = [y_1, \ldots, y_m]$ are m eigenvectors corresponds to the m smallest eigenvalues.

UDFS proposes a local discriminative feature selection algorithm for minimizing total scatter matrix as well as maximizing between class scatter matrix.

NDFS embeds the feature selection phase into spectral clustering as,

$$\begin{aligned} \min_{\mathbf{G},\mathbf{W}} \quad & \operatorname{tr}(\mathbf{G}^\top\mathbf{L}\mathbf{G}) + \alpha\left(\| \mathbf{X}\mathbf{W}-\mathbf{G} \|_F^2 + \beta \| \mathbf{W} \|_{2,1}\right) \\ \text{s.t.} \quad & \mathbf{G}^\top\mathbf{G} = \mathbf{I}\ ,\ \mathbf{G} \geq 0, \end{aligned} \quad (9)$$

where $\mathbf{G} \in \mathbb{R}^{n\times c}$ is the clustering matrix, and c is the number of clusters.

JELSR proposes a framework based on low dimensional embedding as,

$$\begin{aligned} \min_{\mathbf{R},\mathbf{W}} \quad & \operatorname{tr}(\mathbf{R}^\top\mathbf{L}\mathbf{R}) + \alpha\left(\| \mathbf{X}\mathbf{W}-\mathbf{R} \|_F^2 + \beta \| \mathbf{W} \|_{2,1}\right) \\ \text{s.t.} \quad & \mathbf{R}^\top\mathbf{R} = \mathbf{I}, \end{aligned} \quad (10)$$

where $\mathbf{R} \in \mathbb{R}^{n\times m}$ is an embedding matrix in $m \ll p$ dimension.

LDSSL proposes a sparse subspace learning method as,

$$\begin{aligned} \min_{\mathbf{W},\mathbf{H}} \quad & \| \mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{H} \|_F^2 + \alpha \operatorname{tr}(\mathbf{W}^\top\mathbf{X}^\top\mathbf{L}\mathbf{X}\mathbf{W}) + \beta \| \mathbf{W} \|_1 \\ \text{s.t.} \quad & \mathbf{W}^\top\mathbf{W} = \mathbf{I}\ ,\ \mathbf{W},\mathbf{H} \geq 0, \end{aligned} \quad (11)$$

where $\mathbf{W} \in \mathbb{R}^{p\times c}$ and $\mathbf{H} \in \mathbb{R}^{c\times p}$ are low-dimensional matrices.

SCFS is a sparse subspace clustering method based on implicit similarity learning as follows,

$$\begin{aligned} \min_{\mathbf{G},\mathbf{W}} \quad & \| \mathbf{X}-\mathbf{G}\mathbf{G}^\top\mathbf{X} \|_F^2 + \alpha \| \mathbf{X}\mathbf{W}-\mathbf{G} \|_F^2 + \beta \| \mathbf{W} \|_{2,1} \\ \text{s.t.} \quad & \mathbf{G} \geq 0, \mathbf{G}\mathbf{G}^\top\mathbf{1} = \mathbf{1}, \end{aligned} \quad (12)$$

where $\mathbf{1}$ is a matrix of ones.

Both of SPFS and GLSPFS are designed to preserve local structures, while the global similarities are also maintained by GLSPFS. In addition, local linear embedding (LLE) [35], linear preserve projection (LPP) [36], and local tangent space alignment (LTSA) [37] was utilized in GLSPFS.

E. Comparison

We compare the well-known unsupervised feature selection methods in Table I, in terms of the main characteristics including preserving global and local similarities, clustering, and joint learning.

Maintaining the structure of the samples in the low-dimensional space is an important property for selecting features. However, considering either global or local similarities is not adequate for preserving the underlying structures in real-world applications. As represented in Table I, just a few methods preserve both of local and global similarities.

Furthermore, unsupervised feature selection methods can exploit a clustering technique for selecting relevant features in the lack of label information.

Finally, proposing a joint framework as a unified objective function is more prone to avoid sub-optimal solutions in contrast to two-step approaches.

Most of the existing unsupervised feature selection methods considered ad-hoc based approaches from one of these main characteristics. In this work, we propose a joint framework by maintaining both of the global and local similarities as well as cluster analysis to present a robust unsupervised feature selection method.

TABLE I: A comparison of well-known unsupervised feature selection methods

Algorithm	Similarity preserving		Clustering	Joint learning
	Local	Global		
MCFS [28]	✓	×	×	×
UDFS [29]	✓	×	×	✓
NDFS [21]	✓	×	✓	✓
SPFS [33]	×	✓	×	✓
GLSPFS [34]	✓	✓	×	✓
JELSR [31]	✓	×	×	✓
SPUFS [25]	✓	×	×	✓
LDSSL [32]	✓	×	✓	✓
SCFS [22]	✓	×	✓	✓
SLSP (Ours)	✓	✓	✓	✓

III. THE PROPOSED METHOD

In this section, we present details of the proposed method. Then, an optimization algorithmic framework for solving the main objective function is described.

A. The SLSP framework

The proposed framework is designed by maintaining similarities while performing clustering by a low dimensional embedding and regularized regression. First, we perform clustering based on the symmetric nonnegative matrix factorization (S-NMF) [38],

$$\min_{\mathbf{G}\geq 0} \| \mathbf{K} - \mathbf{G}\mathbf{G}^\top \|_F^2, \tag{13}$$

where $\mathbf{K} \in \mathbb{R}^{n\times n}$ is calculated by a Gaussian kernel with parameters σ_i and σ_j as the global similarity matrix,

$$k_{ij} = \exp\left(-\frac{\|\mathbf{x_i}-\mathbf{x_j}\|_2}{\sigma_i\sigma_j}\right). \tag{14}$$

The aim of exploiting S-NMF is to preserve the global similarities by a low dimensional embedding with the clustering purpose. The matrix $\mathbf{G}$ is interpreted as a clustering matrix in the lower dimension $c < \{n, p\}$, where the largest element of the i-th row specifies the cluster of the corresponding sample.

Based on the matrix $\mathbf{G}$, our primary goal is to construct a sparse transformation on the matrix $\mathbf{X}$ by performing the following regularized regression model,

$$\min_{\mathbf{W}} \| \mathbf{XW} - \mathbf{G} \|_F^2 + \beta \| \mathbf{W} \|_{2,1}, \tag{15}$$

where $\mathbf{W} \in \mathbb{R}^{p\times c}$ is a linear transformation matrix, and β is a regularization parameter. The regularized regression matrix $\mathbf{W}$ obtained by optimizing the objective function in Eq. (15) measures the correlation among the features and the clustering labels. We utilize the $\ell_{2,1}$-norm to impose sparsity on the rows of the matrix $\mathbf{W}$. The importance of the features is measured by descending order of ℓ_2-norm of the corresponding feature. If $\mathbf{w_i}$ is close to zero, the i-th feature can be eliminated as a less relevant feature.

For maintaining the local similarities among samples in the transformed matrix $\mathbf{XW}$, we employ a spectral analysis [39] as,

$$\min_{\mathbf{W}} \| \mathbf{XW} - \mathbf{G} \|_F^2 + \alpha \operatorname{tr}(\mathbf{W}^\top\mathbf{X}^\top\mathbf{LXW}) + \beta \| \mathbf{W} \|_{2,1}, \tag{16}$$

where α is a compromising parameter, the laplacian matrix $\mathbf{L}$ is obtained by $\mathbf{L} = \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{S})\mathbf{D}^{-1/2}$, $\mathbf{D}$ is a diagonal matrix as $d_{ii} = \sum_{j=1}^{n} s_{ij}$, and $\mathbf{S}$ is a similarity matrix that is calculated as,

$$s_{ij} = \begin{cases} k_{ij} & \text{if } \mathbf{x_i} \in \mathrm{N}_k(\mathbf{x_j}) \text{ or } \mathbf{x_j} \in \mathrm{N}_k(\mathbf{x_i}) \\ 0 & \text{otherwise,} \end{cases} \tag{17}$$

where $\mathrm{N}_k(\mathbf{x_i})$ denotes the set of k-nearest neighbors of $\mathbf{x_i}$.

We integrate the Eq. (13) and Eq. (16) in a unified objective function to obtain our final framework as,

$$\min_{\mathbf{G}\geq 0,\mathbf{W}} \| \mathbf{K} - \mathbf{G}\mathbf{G}^\top \|_F^2 + \lambda(\| \mathbf{XW} - \mathbf{G} \|_F^2 + \alpha \operatorname{tr}(\mathbf{W}^\top\mathbf{X}^\top\mathbf{LXW}) + \beta \| \mathbf{W} \|_{2,1}), \tag{18}$$

where λ is a weight parameter. The proposed method is enabled to preserve the global and local similarities as well as select relevant features by a sparse learning approach.

B. Optimization

We rewrite the Eq. (18) as,

$$\min_{\mathbf{G}\geq 0,\mathbf{W}} f(\mathbf{G},\mathbf{W}) = \| \mathbf{K} - \mathbf{G}\mathbf{G}^\top \|_F^2 + \lambda(\| \mathbf{XW} - \mathbf{G} \|_F^2 + \alpha\, tr(\mathbf{W}^\top\mathbf{X}^\top\mathbf{LXW}) + \beta \| \mathbf{W} \|_{2,1}). \tag{19}$$

For optimizing the objective function in Eq. (19), a numerical iterative process is employed to consider the main variable $\mathbf{G}$ and $\mathbf{W}$ and non-smoothness of $\ell_{2,1}$-norm regularization. First, by fixing $\mathbf{G}$, the following optimization function can be obtained,

$$\min_{\mathbf{W}} f(\mathbf{W}) = \| \mathbf{XW} - \mathbf{G} \|_F^2 + \alpha \operatorname{tr}(\mathbf{W}^\top\mathbf{X}^\top\mathbf{LXW}) + \beta \| \mathbf{W} \|_{2,1}. \tag{20}$$

By calculating the gradient of $f(\mathbf{W})$ and setting $\nabla f(\mathbf{W})$ to zero,

$$\mathbf{W} = (\mathbf{X}^\top\mathbf{X} + \alpha\mathbf{X}^\top\mathbf{LX} + \beta\mathbf{D})^{-1}\mathbf{X}^\top\mathbf{G}, \tag{21}$$

where the diagonal matrix $\mathbf{D}$ is as,

$$d_{ii} = \frac{1}{2\|\mathbf{w_i}\|_2 + \epsilon}, \tag{22}$$

where ϵ is a small positive number to avoid dividing by zero. We rewrite the Eq. (19) by considering the equality $\operatorname{tr}(\mathbf{W}^\top\mathbf{DW}) = \| \mathbf{W} \|_{2,1}/2$,

$$\min_{\mathbf{G}\geq 0,\mathbf{W},\mathbf{D}} f(\mathbf{G},\mathbf{W},\mathbf{D}) = \| \mathbf{K} - \mathbf{G}\mathbf{G}^\top \|_F^2 + \lambda(\| \mathbf{XW} - \mathbf{G} \|_F^2 + \alpha\, tr(\boldsymbol{W}^\top\boldsymbol{X}^\top\boldsymbol{LXW}) + \beta\, tr(\boldsymbol{W}^\top\boldsymbol{DW})). \tag{23}$$

We have the following equation by putting the obtained $\mathbf{W}$ into the Eq. (23) and defining $\mathbf{M} = (\mathbf{X}^\top\mathbf{X} + \alpha\mathbf{X}^\top\mathbf{LX} + \beta\mathbf{D})$,

$$f(\mathbf{G}) = \| \mathbf{K} - \mathbf{G}\mathbf{G}^\top \|_F^2 + \lambda(\operatorname{tr}(\mathbf{G}^\top\mathbf{G}) - \operatorname{tr}(\mathbf{G}^\top\mathbf{XM}^{-1}\mathbf{XG})). \tag{24}$$

By taking the nonnegative constraint into account, the following objective function is obtained based on $\mathbf{G}$,

$$\min_{\mathbf{G}\geq 0} \quad f(\mathbf{G}) = \| \mathbf{K} - \mathbf{G}\mathbf{G}^\top \|_F^2 + \lambda \operatorname{tr}(\mathbf{G}^\top\mathbf{HG}), \tag{25}$$

where $\mathbf{H} = \mathbf{I_n} - \mathbf{XM}^{-1}\mathbf{X}^\top$. We employ the projected gradient method [40] as follows, to consider the constraint $\mathbf{G} \geq 0$,

$$\mathbf{G}^{t+1} = [\mathbf{G}^t - s^t\nabla f(\mathbf{G}^t)]^+, \tag{26}$$

where $[.]^+$ is a function for projecting the negative numbers to zero. We utilize the armijo rule [41] for setting the learning rate s^t in the t-th iteration. The $\nabla f(\mathbf{G}^t)$ based on Eq. (25) is as,

$$\nabla f(\mathbf{G}^t) = (\mathbf{G}^t(\mathbf{G}^t)^\top + \frac{\lambda}{2}\mathbf{H}^t - \mathbf{K})\mathbf{G}^t. \quad (27)$$

Algorithm 1 summarizes the proposed method, where G is updated by Eq. (26), and W is updated by Eq. (21) in an iterative manner.

IV. The Analysis of the Proposed Algorithm

In this section, we first analyze the convergence of the SLSP algorithm, and then we explain the computational complexity of the method.

A. Convergence Analysis

In this subsection, we prove the convergence of the Algorithm 1. First, a lemma is given according to [42],

Lemma 1. *By considering $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$ as two arbitrary nonzero vectors, we have the following inequality,*

$$\| \mathbf{u} \|_2 - \frac{\|\mathbf{u}\|_2^2}{2\|\mathbf{v}\|_2} \leq \| \mathbf{v} \|_2 - \frac{\|\mathbf{v}\|_2^2}{2\|\mathbf{v}\|_2}. \quad (28)$$

Theorem 1. *The behavior of the objective function in Eq. (19) is non-increasing, by utilizing the Algorithm 1.*

Proof. First, we fix $\mathbf{W}^t$ in $(t+1)$-th iteration. The following inequality is given based on non-increasing property of projected gradient method for updating $\mathbf{G}^{t+1}$ by appropriate step size s_t.

$$f(\mathbf{G}^{t+1}, \mathbf{W}^t, \mathbf{D}^t) \leq f(\mathbf{G}^t, \mathbf{W}^t, \mathbf{D}^t). \quad (29)$$

Now, we assume $\mathbf{G}^{t+1}$ to be fixed. The obtained $\mathbf{W}^{t+1}$ in Eq. (21) is the solution of the following objective function,

$$\min_{\mathbf{W}} \| \mathbf{XW} - \mathbf{G}^{t+1} \|_F^2 + \alpha \operatorname{tr}(\mathbf{W}^\top \mathbf{X}^\top \mathbf{L}\mathbf{X}\mathbf{W}) + \beta \operatorname{tr}(\mathbf{W}^\top \mathbf{D}^t \mathbf{W}). \quad (30)$$

Thus, the following inequality should be shown,

$$f(\mathbf{G}^{t+1}, \mathbf{W}^{t+1}, \mathbf{D}^t) \leq f(\mathbf{G}^{t+1}, \mathbf{W}^t, \mathbf{D}^t). \quad (31)$$

We write the inequality (31) as,

$$\begin{aligned} &\| \mathbf{XW}^{t+1} - \mathbf{G}^{t+1} \|_F^2 + \alpha \operatorname{tr}((\mathbf{W}^{t+1})^\top \mathbf{X}^\top \mathbf{L}\mathbf{X}\mathbf{W}^{t+1}) \\ &+ \beta \sum_{i=1}^p \left(\frac{\|\mathbf{w_i}^{t+1}\|_2^2}{2\|\mathbf{w_i}^t\|_2}\right) \\ &\leq \| \mathbf{XW}^t - \mathbf{G}^{t+1} \|_F^2 + \alpha \operatorname{tr}((\mathbf{W}^t)^\top \mathbf{X}^\top \mathbf{L}\mathbf{X}\mathbf{W}^t) \\ &\quad + \beta \sum_{i=1}^p \left(\frac{\|\mathbf{w_i}^t\|_2^2}{2\|\mathbf{w_i}^t\|_2}\right). \end{aligned} \quad (32)$$

By rewriting the above inequality we have,

$$\begin{aligned} &\| \mathbf{XW}^{t+1} - \mathbf{G}^{t+1} \|_F^2 + \alpha \operatorname{tr}((\mathbf{W}^{t+1})^\top \mathbf{X}^\top \mathbf{L}\mathbf{X}\mathbf{W}^{t+1}) \\ &\quad + \beta \| \mathbf{W}^{t+1} \|_{2,1} - \beta \sum_{i=1}^p \left(\| \mathbf{w_i}^{t+1} \|_2 - \frac{\|\mathbf{w_i}^{t+1}\|_2^2}{2\|\mathbf{w_i}^t\|_2}\right) \\ &\leq \| \mathbf{XW}^t - \mathbf{G}^{t+1} \|_F^2 + \alpha \operatorname{tr}((\mathbf{W}^t)^\top \mathbf{X}^\top \mathbf{L}\mathbf{X}\mathbf{W}^t) \\ &\quad + \beta \| \mathbf{W}^t \|_{2,1} - \beta \sum_{i=1}^p \left(\| \mathbf{w_i}^t \|_2 - \frac{\|\mathbf{w_i}^t\|_2^2}{2\|\mathbf{w_i}^t\|_2}\right). \end{aligned} \quad (33)$$

The lemma 1 implies that,

$$\begin{aligned} &\| \mathbf{XW}^{t+1} - \mathbf{G}^{t+1} \|_F^2 + \alpha \operatorname{tr}((\mathbf{W}^{t+1})^\top \mathbf{X}^\top \mathbf{L}\mathbf{X}\mathbf{W}^{t+1}) + \beta \| \mathbf{W}^{t+1} \|_{2,1} \\ &\leq \| \mathbf{XW}^t - \mathbf{G}^{t+1} \|_F^2 + \alpha \operatorname{tr}((\mathbf{W}^t)^\top \mathbf{X}^\top \mathbf{L}\mathbf{X}\mathbf{W}^t) + \beta \| \mathbf{W}^t \|_{2,1}. \end{aligned} \quad (34)$$

■

Algorithm 1 SLSP algorithm

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and parameter $\lambda, \alpha, \beta, c, \epsilon$.
Output: Feature rank based on descending order of $\|\mathbf{w_i}\|_2$, $(i = 1...p)$.

1: $t = 0$.
2: Initialize $\mathbf{D}^t$ as a $p \times p$ identity matrix.
3: Initialize $\mathbf{G}^t$ as an $n \times c$ cluster indicator matrix.
4: Construct the global similarity matrix $\mathbf{K}$ and the laplacian matrix $\mathbf{L}$.
5: **repeat**
6: $\mathbf{M}^t = \mathbf{X}^\top \mathbf{X} + \alpha \mathbf{X}^\top \mathbf{L}\mathbf{X} + \beta \mathbf{D}^t$.
7: $\mathbf{H}^t = \mathbf{I_n} - \mathbf{X}(\mathbf{M}^t)^{-1}\mathbf{X}^\top$.
8: Calculate $\nabla f(\mathbf{G}^t) = (2\mathbf{G}^t(\mathbf{G}^t)^\top + \frac{\lambda}{2}\mathbf{H}^t - \mathbf{K})\mathbf{G}^t$.
9: Update $\mathbf{G}^{t+1}$ by projected gradient method as
$$\mathbf{G}^{t+1} = [\mathbf{G}^t - s^t \nabla f(\mathbf{G}^t)]^+.$$
10: Update $\mathbf{W}^{t+1} = (\mathbf{M}^t)^{-1}\mathbf{X}^\top \mathbf{G}^{t+1}$.
11: Update $\mathbf{D}^{t+1}$ as a diagonal matrix by
$$d_{ii}^{t+1} = \frac{1}{2\|\mathbf{w_i}^{t+1}\|_2 + \epsilon}.$$
12: $t = t + 1$.
13: **until** Convergence of objective function value in (25).

Therefore, based on the Eq. (34) and Eq. (29), Algorithm 1 has non-increasing behavior in the objective function in Eq. (19).

B. Computational Complexity

Updating $\mathbf{G}$ and $\mathbf{W}$ is the main steps in Algorithm 1. We require $O(np^2 + n^2p)$, $O(np^2 + n^2p + p^3)$, and $O(n^2c)$ time complexity to compute $\mathbf{M}$, $\mathbf{H}$, and $\nabla f(\mathbf{G})$. Therefore, computing $\mathbf{G}$ leads to $O(p^3 + np^2 + n^2p + n^2c)$ time complexity. On the other hand, the computational order of updating $\mathbf{W}$ is $O(p^3 + np^2 + npc)$. Eventually, $max\{O(p^3), O(n^2p), O(np^2), O(n^2c), O(npc)\}$ is the time cost of the Algorithm 1. Since, in almost all real world cases, the number of clusters is more smaller than the number of features, $c \ll p$, the final computational complexity of the proposed algorithm is $\max\{O(p^3), O(n^2p)\}$.

V. Experiments

In this section, we evaluate the proposed method, SLSP, by comparing with the well-known unsupervised feature selection methods on different standard datasets.

A. Datasets

We utilize some standard datasets in various applications including biological (ALLAML, Colon, GLIOMA, Lung), image (BA, COIL20, ORL, Yale), voice (Isolet), and artificial data (Madelon). The BA is available on https://cs.nyu.edu/~roweis/data.html, while all other datasets can be downloaded from [1]. The summery of the datasets is presented in Table II.

B. Evaluation measures

Accuracy (Acc) and normalized mutual information (NMI) are widely used as standard measures to evaluate the unsupervised methods. Let the ground truth and predicted label vectors are denoted by $\mathbf{y}$ and $\mathbf{z}$. Acc is presented as,

$$\text{Acc}(\mathbf{y}, \mathbf{z}) = \frac{1}{n}\sum_{i=1}^n \delta(y_i, \text{map}(z_i)), \quad (35)$$

where the function $\delta(a, b)$ equals to 1 if $a = b$ and 0, otherwise. For map(.) function, Kuhn-Munkres

TABLE II: The main characteristics of the used datasets.

Dataset	Samples	Features	Classes	Type
ALLAML	72	7129	2	Biology
Colon	62	2000	2	Biology
GLIOMA	50	4434	4	Biology
Lung	203	3312	5	Biology
BA	1404	320	36	Image
COIL20	1440	1024	20	Image
ORL	400	1024	40	Image
Yale	165	1024	15	Image
Isolet	1560	617	26	Voice
Madelon	2600	500	2	Artificial

approach [43] is employed to find the best permutation for matching the categories in vectors $\mathbf{y}$ and $\mathbf{z}$.

The NMI is defined as,

$$\mathrm{NMI}(\mathbf{y},\mathbf{z}) = \frac{\mathrm{I}(\mathbf{y},\mathbf{z})}{\max(\mathrm{H}(\mathbf{y}),\mathrm{H}(\mathbf{z}))}, \quad (36)$$

where the entropy function is denoted by $\mathrm{H}(.)$ and the mutual information of $\mathbf{y}$ and $\mathbf{z}$ is presented by $\mathrm{I}(\mathbf{y},\mathbf{z})$ as,

$$\mathrm{I}(\mathbf{y},\mathbf{z}) = \sum_{y\in\mathbf{y}}\sum_{z\in\mathbf{z}} \mathrm{p}(y,z)\log\left(\frac{\mathrm{p}(y,z)}{\mathrm{p}(y)\,\mathrm{p}(z)}\right). \quad (37)$$

C. Experimental settings

We compare the proposed method with the state-of-the-art unsupervised feature selection algorithms including LS [10], MCFS [28], UDFS [29], NDFS [21], GLSPFS [34], SPUFS [25], and selecting all features namely All_Feat.

We set $k = 5$ for k-nearest neighbor algorithm and σ_i (or σ_j) is set to the distance between $\mathbf{x_i}$ (or $\mathbf{x_j}$) and its seventh neighbor based on [44]. For calculating the similarity matrix in other methods, we set the $\sigma = 1$. For NDFS method, we set $\gamma = 10^8$. A grid search strategy is employed to set the parameters α, β and λ from the set of $\{10^{-4}, 10^{-2}, 1, 10^2, 10^4\}$ candidates. For evaluating a method by NMI and Acc measures, we employ k-means algorithm in different number of selected features from {50,100,150,200,250,300} and the mean and standard deviation of NMI and Acc on 20 times repetitions.

D. Discussion and analysis of the results

By taking the results in Tables III and IV into consideration, we have the following conclusions.

- Comparing the results of All_Feat and other methods shows the better performance of selecting relevant features rather than all features. As a consequence, feature selection provides the ease-of-interpretation as well as better performance of the learning algorithms.
- The primary reason of obtaining more accurate results of SLSP is to take local and global structure into account. In the absence of label information, preserving the geometric structure of the samples is yielded to select more relevant features than the earlier methods.
- Furthermore, incorporating clustering in SLSP provides an effective strategy to select features based on underlying categories in data.
- Moreover, employing subspace learning in SLSP enables acquiring the discriminative information in the original data that leads to higher performance.

The proposed method obtains the best or at least the second best results in Tables III and IV in almost all the cases. While, GLSPFS and NDFS provide good results on some datasets, the proposed method SLSP perform very satisfactory on all of the datasets with negligible differences with the top results on some datasets. Furthermore, the "ALLAML" can be regarded as the most challenging dataset in our experimental setting where there are few samples, $n = 72$, and many features, $p = 7129$. As the obtained results revealed that there are a considerable difference between the attained results of SLSP and its competitors, which indicates the strength of the proposed approach. Moreover, the stability and good performance of the SLSP over the different datasets from a variety of applications show the robustness of the SLSP.

E. Sensitivity analysis

In this subsection, we consider the sensitivity of the proposed method, SLSP, in terms of setting the parameters. The main parameters in the objective function of SLSP in Eq. (18) are λ, α and β. By fixing the parameter $\alpha = 1$ for briefness, we investigate the sensitivity of the parameters λ and β. The experimental results of SLSP on Acc measure in $\log_{10}$ on the datasets in terms of different setting of the λ and β parameters are shown in Fig. 1. The slight sensitivity to the parameters in the performance of the proposed method (SLSP) is shown in Fig. 1.

VI. Conclusion

An unsupervised feature selection method based on sparse learning was presented in this work. We employed a symmetric nonnegative matrix factorization for cluster analysis as well as maintaining global similarities and low dimensional embedding. The correlation between the features and the clusters was measured by performing a linear regression model. The spectral analysis was employed for preserving local similarities in the selected feature space. The $\ell_{2,1}$ regularization was applied to the objective function to obtain a sparse feature representation. We presented a numerical optimization procedure to solve the proposed objective function and theoretically analyzed the convergence of the proposed algorithm. The experimental results showed that SLSP outperformed the well-known unsupervised feature selection methods due to propose a joint framework including clustering, global and local similarity maintaining in a sparse way.

There are some challenges and future directions on this interesting domain including the feature selection for online streaming data, enhancing the deep neural networks architectures by employing the selected features as the pre-train of the network, considering other numerical optimization algorithms, and applying a low-rank representation approach.

TABLE III: Clustering results (Acc% ± std) of some unsupervised feature selection methods on ten standard datasets. The best and the second best results are denoted by bold and underlined numbers.

Dataset	ALLAML	Colon	GLIOMA	Lung	BA	COIL20	ORL	Yale	Isolet	Madelon
All_Feat	73.19 ± 2.33	54.84 ± 0.00	58.30 ± 3.96	71.85 ± 7.48	42.70 ± 1.36	68.03 ± 3.86	59.35 ± 2.35	41.52 ± 2.44	62.21 ± 2.12	50.33 ± 0.08
LS	54.02 ± 0.90	54.05 ± 1.38	56.77 ± 2.19	55.45 ± 6.26	38.44 ± 5.09	58.31 ± 1.47	49.96 ± 3.51	39.72 ± 1.28	53.27 ± 4.33	54.74 ± 3.38
MCFS	69.98 ± 3.52	56.64 ± 1.99	57.23 ± 2.43	60.82 ± 6.54	40.90 ± 2.98	67.90 ± 1.92	57.22 ± 2.06	40.66 ± 2.33	49.61 ± 3.52	52.75 ± 1.69
UDFS	71.03 ± 2.22	59.41 ± 1.45	64.75 ± 1.44	58.96 ± 5.67	36.20 ± 4.46	62.46 ± 1.24	51.29 ± 1.40	39.42 ± 1.92	**70.89** ± 7.76	50.86 ± 0.42
NDFS	76.39 ± 0.00	56.21 ± 1.08	**66.83** ± 1.34	83.55 ± 0.25	43.32 ± 1.57	66.08 ± 2.13	59.68 ± 0.17	40.68 ± 1.24	69.33 ± 1.36	54.20 ± 3.55
GLSPFS	76.05 ± 0.52	**62.73** ± 0.68	60.28 ± 1.38	78.83 ± 4.55	43.24 ± 1.51	69.46 ± 0.77	**60.43** ± 1.17	40.62 ± 1.05	63.64 ± 3.26	57.73 ± 3.74
SPUFS	70.39 ± 0.97	60.05 ± 0.73	57.00 ± 0.78	67.33 ± 1.20	38.12 ± 5.21	67.86 ± 2.56	48.86 ± 2.65	35.04 ± 0.85	63.35 ± 3.11	51.19 ± 0.10
SLSP (Ours)	**86.15** ± 0.08	62.38 ± 2.82	65.45 ± 1.89	**84.66** ± 1.28	**43.41** ± 1.45	**69.95** ± 0.96	60.33 ± 0.18	**42.43** ± 0.94	65.21 ± 2.18	**57.81** ± 1.51

TABLE IV: Clustering results (NMI% ± std) of some unsupervised feature selection methods on ten standard datasets. The best and the second best results are denoted by bold and underlined numbers.

Dataset	ALLAML	Colon	GLIOMA	Lung	BA	COIL20	ORL	Yale	Isolet	Madelon
All_Feat	15.14 ± 2.64	00.60 ± 0.09	49.08 ± 2.15	63.91 ± 2.67	58.12 ± 0.72	79.08 ± 1.49	77.71 ± 1.13	49.43 ± 1.67	77.06 ± 1.02	00.00 ± 0.00
LS	00.40 ± 0.38	00.23 ± 0.39	47.85 ± 1.40	48.28 ± 7.35	53.83 ± 5.07	72.64 ± 1.99	70.89 ± 2.60	48.73 ± 1.38	70.36 ± 4.30	00.99 ± 1.30
MCFS	11.15 ± 4.43	00.78 ± 0.78	44.77 ± 2.42	51.88 ± 1.25	56.44 ± 2.67	77.40 ± 1.16	76.12 ± 1.48	48.66 ± 2.73	65.53 ± 4.19	00.30 ± 0.27
UDFS	14.77 ± 2.93	02.21 ± 0.74	47.63 ± 2.79	46.85 ± 4.65	52.26 ± 4.27	72.36 ± 0.88	72.29 ± 1.30	47.21 ± 2.27	70.89 ± 7.76	00.03 ± 0.03
NDFS	18.95 ± 0.00	00.95 ± 0.28	53.72 ± 0.46	68.50 ± 1.18	58.45 ± 1.65	77.24 ± 1.93	**78.28** ± 0.37	47.29 ± 1.03	**79.87** ± 1.80	00.88 ± 1.36
GLSPFS	19.11 ± 1.85	03.10 ± 0.27	53.96 ± 0.22	65.74 ± 2.04	**58.47** ± 1.40	78.33 ± 0.68	78.02 ± 0.61	47.12 ± 0.56	76.67 ± 2.54	**02.15** ± 1.58
SPUFS	10.80 ± 1.45	02.64 ± 0.29	53.68 ± 0.54	60.06 ± 1.98	53.43 ± 4.83	77.80 ± 1.97	70.24 ± 2.16	42.40 ± 0.76	72.86 ± 1.81	00.06 ± 0.01
SLSP (Ours)	**48.46** ± 0.20	**15.24** ± 3.61	**54.15** ± 0.51	**69.58** ± 0.64	58.46 ± 1.65	**79.57** ± 0.93	78.16 ± 0.36	**50.11** ± 0.87	77.72 ± 2.12	01.84 ± 0.73

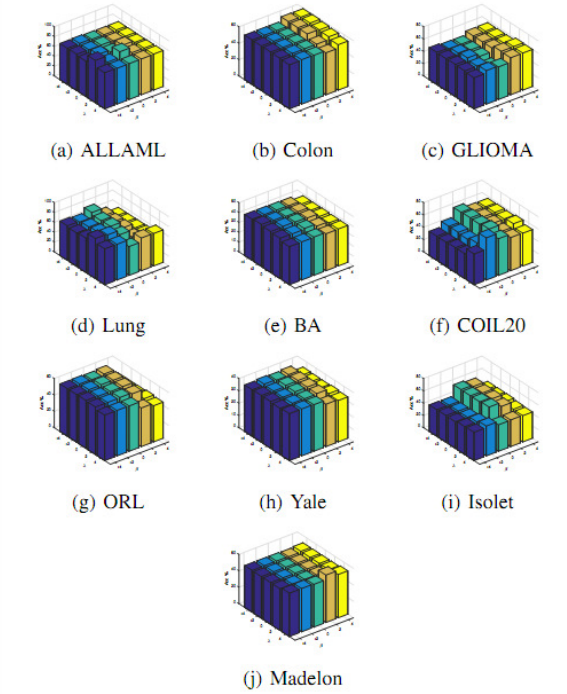

Fig. 1: Accuracy of SLSP with different values of the parameters λ and β.

REFERENCES

[1] J. Li et al., "Feature Selection: A Data Perspective, http://featureselection.asu.edu/," ACM Comput. Surv., vol. 50, no. 6, p. 94:1-94:45, 2017.

[2] C. Bishop, Pattern Recognition and Machine Learning. New York: Springer, 2007.

[3] K. Fukunaga, Introduction to statistical pattern recognition, 2nd ed. Boston: Academic Press, 1990.

[4] H. Shi, Y. Li, Y. Han, and Q. Hu, "Cluster Structure Preserving Unsupervised Feature Selection for Multi-View Tasks," Neurocomputing, vol. 175, pp. 686–697, Jan. 2016.

[5] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation on feature selection for text clustering," in Icml, 2003, vol. 3, pp. 488–495.

[6] J. Tang and H. Liu, "An Unsupervised Feature Selection Framework for Social Media Data," IEEE Trans. Knowl. Data Eng., vol. 26, no. 12, pp. 2914–2927, Dec. 2014.

[7] H. Zare, M. Hajiabadi, and M. Jalili, "Detection of Community Structures in Networks with Nodal Features based on Generative Probabilistic Approach," IEEE Trans. Knowl. Data Eng., pp. 1–1, 2019.

[8] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. Fco. Martínez-Trinidad, "A review of unsupervised feature selection methods," Artif. Intell. Rev., vol. 53, no. 2, pp. 907–948, Feb. 2020.

[9] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artif. Intell., vol. 97, no. 1, pp. 273–324, Dec. 1997.

[10] X. He, D. Cai, and P. Niyogi, "Laplacian Score for Feature Selection," in Proceedings of the 18th International Conference on Neural Information Processing Systems, Cambridge, MA, USA, 2005, pp. 507–514.

[11] Z. Zhao and H. Liu, "Spectral Feature Selection for Supervised and Unsupervised Learning," in Proceedings of the 24th International Conference on Machine Learning, New York, NY, USA, 2007, pp. 1151–1157.

[12] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J Mach Learn Res, vol. 3, pp. 1157–1182, Mar. 2003.

[13] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining. Springer Science & Business Media, 2012.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[15] H. Zare and M. Niazi, "Relevant based structure learning for feature selection," Eng. Appl. Artif. Intell., vol. 55, pp. 93–102, Oct. 2016.

[16] H. Liu and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," in Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence, Nov. 1995, pp. 388–391.

[17] J. Liu, S. Ji, and J. Ye, "Multi-Task Feature Learning via Efficient L2, 1-Norm Minimization," in Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, 2009, pp. 339–348.

[18] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu, "Clustering-Guided Sparse Structural Learning for Unsupervised Feature Selection," IEEE Trans. Knowl. Data Eng., vol. 26, no. 9, pp. 2138–2150, Sep. 2014.

[19] J. Li, J. Tang, and H. Liu, "Reconstruction-based Unsupervised Feature Selection: An Embedded Approach," pp. 2159–2165, 2017.

[20] M. Masaeli, Y. Yan, Y. Cui, G. Fung, and J. Dy, "Convex Principal Feature Selection," in Proceedings of the 2010 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, 2010, pp. 619–628.

[21] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised Feature Selection Using Nonnegative Spectral Analysis," in Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto, Ontario, Canada, 2012, pp. 1026–1032.

[22] M. G. Parsa, H. Zare, and M. Ghatee, "Unsupervised feature selection based on adaptive similarity learning and subspace clustering," Eng. Appl. Artif. Intell., vol. 95, p. 103855, Oct. 2020.

[23] A. K. Farahat, A. Ghodsi, and M. S. Kamel, "Efficient greedy feature selection for unsupervised learning," Knowl. Inf. Syst., vol. 35, no. 2, pp. 285–310, May 2013.

[24] A. K. Farahat, A. Ghodsi, and M. S. Kamel, "An Efficient Greedy Method for Unsupervised Feature Selection," in 2011 IEEE 11th International Conference on Data Mining, Dec. 2011, pp. 161–170.

[25] Q. Lu, X. Li, and Y. Dong, "Structure preserving unsupervised feature selection," Neurocomputing, vol. 301, pp. 36–45, Aug. 2018.

[26] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with Sparsity-Inducing Penalties," Found. Trends® Mach. Learn., vol. 4, no. 1, pp. 1–106, 2012.

[27] T. Hastie, R. Tibshirani, and M. Wainwright, Statistical Learning with Sparsity: The Lasso and Generalizations, 1 edition. Boca Raton: Chapman and Hall/CRC, 2015.

[28] D. Cai, C. Zhang, and X. He, "Unsupervised Feature Selection for Multi-cluster Data," in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2010, pp. 333–342.

[29] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "L2,1-norm Regularized Discriminative Feature Selection for Unsupervised Learning," in Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two, Barcelona, Catalonia, Spain, 2011, pp. 1589–1594.

[30] C. Hou, F. Nie, D. Yi, and Y. Wu, "Feature Selection via Joint Embedding Learning and Sparse Regression," in Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two, Barcelona, Catalonia, Spain, 2011, pp. 1324–1329.

[31] C. Hou, F. Nie, X. Li, D. Yi, and Y. Wu, "Joint Embedding Learning and Sparse Regression: A Framework for Unsupervised Feature Selection," IEEE Trans. Cybern., vol. 44, no. 6, pp. 793–804, Jun. 2014.

[32] R. Shang, Y. Meng, W. Wang, F. Shang, and L. Jiao, "Local discriminative based sparse subspace learning for feature selection," Pattern Recognit., vol. 92, pp. 219–230, Aug. 2019.

[33] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On Similarity Preserving Feature Selection," IEEE Trans. Knowl. Data Eng., vol. 25, no. 3, pp. 619–632, Mar. 2013.

[34] X. Liu, L. Wang, J. Zhang, J. Yin, and H. Liu, "Global and Local Structure Preservation for Feature Selection," IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 6, pp. 1083–1095, Jun. 2014.

[35] S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," Science, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[36] X. He and P. Niyogi, "Locality Preserving Projections," in Proceedings of the 16th International Conference on Neural Information Processing Systems, Cambridge, MA, USA, Dec. 2003, pp. 153–160.

[37] J. Wang, "Local Tangent Space Alignment," in Geometric Structure of High-Dimensional Data and Dimensionality Reduction, J. Wang, Ed. Berlin, Heidelberg: Springer, 2012, pp. 221–234.

[38] D. Kuang, C. Ding, and H. Park, "Symmetric Nonnegative Matrix Factorization for Graph Clustering," in Proceedings of the 2012 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, 2012, pp. 106–117.

[39] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an algorithm," in Advances in Neural Information Processing Systems 14, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002, pp. 849–856.

[40] C.-J. Lin, "Projected Gradient Methods for Nonnegative Matrix Factorization," Neural Comput., vol. 19, no. 10, pp. 2756–2779, Aug. 2007.

[41] D. P. Bertsekas, Nonlinear programming, Third edition. Belmont, Massachusetts: Athena Scientific, 2016.

[42] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and Robust Feature Selection via Joint l2,1-Norms Minimization," in Advances in Neural Information Processing Systems 23, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 1813–1821.

[43] M. D. Plummer and L. Lovász, Matching Theory. Elsevier Science, 1986.

[44] L. Zelnik-manor and P. Perona, "Self-Tuning Spectral Clustering," in Advances in Neural Information Processing Systems 17, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, 2005, pp. 1601–1608.

Hadi Zare received his B.Sc. degree in Statistics from Shiraz University, Iran, and his M.Sc. degree in Statistics from Amirkabir University of Technology, Tehran, Iran, and his Ph.D. in Applied Mathematics from Amirkabir University of Technology, Tehran, Iran. Now he is an assistant professor of the Faculty of New Sciences and Technologies, University of Tehran, Iran. His research interests include Machine Learning, Optimization, Graphical Model, Data Science, and Social Networks.

Mohsen Ghassemi Parsa received his B.Sc. degree in Computer Engineering (Software) from Payame Noor University of Shiraz, Iran, and his M.Sc. degree in Computer Engineering (Artificial Intelligence) from Iran University of Science and Technology (IUST), Tehran, Iran. Now he is a Ph.D. candidate on Information Engineering in University of Tehran, Iran. His research interests include Artificial Intelligence, Sparse Machine Learning, High-Dimensional Data, Feature Selection, Complex Networks, and Evolutionary Computation.

Mahdi Ghatee received his B.Sc. degree in Applied Mathematics from Shiraz University, Iran, and his M.Sc. degree in Applied Mathematics from Amirkabir University of Technology, Tehran, Iran, and his Ph.D. in Computer Science from Amirkabir University of Technology, Tehran, Iran. Now he is an associated professor of the Department of Computer Science, Amirkabir University of Technology, Tehran, Iran. His research interests include Data Mining and Data Science, Neural Networks, Intelligent Transportation Systems, and Network Optimization.

Sasan H. Alizadeh received his B.Sc. degree in Computer Engineering from Shiraz University, Iran, and his M.Sc. and Ph.D. degrees in Computer Science from Amirkabir University of Technology, Tehran, Iran. Now he is an assistant professor of the Department of Information Technology, IRAN Telecommunication Research Center, Tehran, Iran. His research interests include Recommender Systems, Machine Learning, Artificial Intelligence, and Stochastic Processes.