IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 3, 2020, accepted September 13, 2020, date of publication September 18, 2020,
date of current version September 28, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024690

Unsupervised Dual Learning for Feature and Instance Selection

LIANG DU[1,3], (Member, IEEE), XIN REN[1], PENG ZHOU[2], (Member, IEEE), AND ZHIGUO HU[1,3]

[1]School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
[2]School of Computer Science and Technology, Anhui University, Hefei 230601, China
[3]Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China

Corresponding authors: Liang Du (duliang@sxu.edu.cn) and Zhiguo Hu (646579354@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61502289 and Grant 61806003, in part by the Shanxi Province Key Research and Development Program under Grant 201803D31199, in part by the Natural Science Foundation of Shanxi Province, China, under Grant 201801D221163, and in part by the Scientific and Technological Innovation Programs (STIP) of Higher Education Institutions in Shanxi under Grant 2016101.

ABSTRACT Feature selection and instance selection are dual operations on a data matrix. Feature selection aims at selecting a subset of relevant and informative features from original feature space, while instance selection at identifying a subset of informative and representative instances. Most of previous studies address these two problems separately, such that irrelevant features (resp. outliers) may mislead the process of instance (resp. feature) selection. In this paper, we address the problem by doing feature and instance selection simultaneously. We propose a novel unified framework, which chooses instances and features simultaneously, such that 1)all the data can be reconstructed from the selected instances and features and 2) the global structure which is characterized by the sparse reconstruction coefficient is preserved. Experimental results on several benchmark data sets demonstrate the effectiveness of our proposed method.

INDEX TERMS Unsupervised feature selection, unsupervised instance selection, active learning, dual selection.

I. INTRODUCTION

Real world applications usually involve with big data with large volume and high dimensionality, presenting great challenges such as the curse of dimensionality, huge computation and storage cost. To tackle these difficulties, a lot of algorithms have been developed in the literature. The high dimension of features can be largely alleviated by feature selection techniques, which aims at keeping a few informative and relevant features [1]–[4]. Instead of being a passive recipient of data, instance selection (a.k.a, active learning) keeps the most informative data for labeling and further training [5]–[7]. These two tasks, i.e., feature selection and instance selection, are often solved separately. It's not strange that, the process of feature selection may be misled by the less informative instances, and meanwhile, the performance of instance selection may be degenerated with irrelevant features. Ideally, we should select features only on the most informative samples, and select instances only on relevant features. However, Most of the existing work addressed these two problems separately. Thus, in this paper, we consider the problem of dual selection on features and instances simultaneously.

With existing one-side only techniques, i.e., active learning methods and unsupervised feature selection algorithms, there are at least three different strategies to merge their results to achieve dual selection. Firstly, we can run one side algorithm on the original data set independently, and then manually merge their selected results. We can also run these one-side algorithm in sequence, i.e., feature selection first and then instance selection, or instance selection first and then feature selection. More specifically, we can run the feature selection on the original data and obtain the selected features, and then run the active learning method on data only with selected features to obtain the selected samples. We can also run instance selection method first to obtain the selected samples and then run feature selection method on data with all features and selected samples to obtain the selected features [8]. Apparently, The duality between feature selection and instance selection has been neglected by these tandem algorithms. Thus, these methods suffer from adverse effect from noisy features and outliers.

The associate editor coordinating the review of this manuscript and approving it for publication was Bohui Wang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Recently, Zhang *et al.* [9] proposed a unified feature and instance selection framework (UFI) based on A-optimal experimental design (AOD) [5]. The basic idea is to simultaneously select those features and instances that can minimize the size of parameter covariance matrix. Due to its combinatorial nature, UFI uses a greedy backward removal schema, which deletes the least informative features and instances sequentially. It has been shown that, UFI can achieve promising performance for data dual reduction. However, it evaluates the importance of each feature and instance individually and removes the less informative feature (instance) one by one. Such backward removal mechanism cannot make full use of the correlation among features and instances, and also incurs high computational costs.

In this paper, we propose a novel unsupervised dual learning framework to effectively diffuse the process of feature and instance selection (DFIS for short). From the view of instance selection, we resort to the intermediate results of feature selection instead of using all the relevant and irrelevant features. With the selected features, we keep these data points such that all the data can be well reconstructed by the selected ones. Similarly, we select the feature subset to best preserve the inherent structure of the data, where the structure is adaptively determined by the most informative instances from the result of instance selection rather than all the instances. By leveraging the interactions between these two selection tasks, it is believed that the dual learning method could achieve better results on both sides.

Our main contributions are highlighted as follows,

- we propose a novel method for dual selection, which chooses instances to reconstruct all the data in the reduced feature space and keeps the best features to preserve the global structure characterized by the sparse reconstruction coefficients among the selected instances.
- we present an effective algorithm to solve the optimization problem, which is non-greedy and proved to be converged.
- Experimental results on several benchmark data sets demonstrate the effectiveness and efficiency of the proposed method.

Notations: In this paper, matrices are written as boldface uppercase letters and vectors are written as boldface lowercase letters. Given a matrix $\mathbf{H} = \{h_{ij}\}$, we denote its i-th row and j-th column as $\mathbf{h}_i$ and $\mathbf{h}^j$ respectively. The ℓ_p-norm of a vector $\mathbf{v} \in \mathcal{R}^n$ is defined as $||\mathbf{v}|| = (\sum_{i=1}^{n} |v_i|^p)^{\frac{1}{p}}$. The Frobenius norm of a matrix $\mathbf{H} \in \mathcal{R}^{n\times m}$ is defined as $||\mathbf{H}||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} h_{ij}^2} = \sqrt{\sum_{i=1}^{n} ||\mathbf{h}_i||_2^2}$. The $\ell_{2,1}$-norm of $\mathbf{H}$ is defined as $||\mathbf{H}||_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m} h_{ij}^2} = \sum_{i=1}^{n} ||\mathbf{h}_i||_2$. The $\ell_{2,0}$-norm of matrix $\mathbf{H}$ is defined as the number of nonzero rows. tr($\mathbf{M}$) is the trace of a squared matrix $\mathbf{M} \in \mathcal{R}^{n\times n}$.

II. RELATED WORK

A. UNSUPERVISED FEATURE SELECTION

Various methods have been developed for the task of feature selection in the unsupervised setting. Most of existing works distinguish these algorithms into three groups, i.e., filter [2], [4], [10], wrapper and embedded approaches [11]–[13], in terms of different selection strategy. Moreover, with the absent of supervised information, one of the key problem for unsupervised feature selection is to design the appropriate criterion to guide the search of relevant and informative features. In the previous literature, there are at least three type of criteria are well developed. A number of methods [1], [14]–[22] aim to exploit the intrinsic cluster structure of data, and use it as pseudo label for further feature selection task. Another line of work [15], [23]–[31] is to select those features which can be used to well reconstruct or approximate the whole data set. Besides, it has also been verified that the local structure of data is also vital important for unsupervised features selection [16], [32]–[34]. Most recently, several techniques have also been introduced to further improve unsupervised feature selection, such as the adaptive graph learning [35], [36], the ensemble of weak partitions [37]. It should be pointed out that unsupervised feature selection has also been extended to handle multi-view data.

Although a lot of algorithms have been developed, it is still worthwhile to point out that many existing works surfer from the annoying problem of hyper parameter selection in supervised setting [16], [38], [39]. Beside, it is believed that the low quality of data not only appeared on the feature side, but also the sample side.

Based on the above analysis, we aim to improve the feature selection with the help of dual selection mechanism. As a result, the process of unsupervised feature selection will be less influenced by low quality samples.

B. INSTANCE SELECTION

Instance selection targets at selecting the most informative instances from a large scale data set. Like the counterpart of unsupervised feature selection, one of the key issue in instance selection is to design the appropriate criterion, which is used to decide the usefulness of data samples. There are at least two strategies, i.e., representative [40], [41] and uncertainty sampling [42], to guide the search of informative instances. On the other hand, instance selection can also be categorized into early selection [43], where no labeled data is available, and normal selection which can access certain amount of labeled data.

In statistics, the problem of instance selection is referred to the Optimal Experimental Design (OED), which is to minimize the variance of a parameterized model. There are three typical types of design criteria: D-optimal design, A-optimal design and E-optimal design. Recently, [6], [41] proposed transductive experimental design, which can fully explore the available unlabeled data. There are also several extensions of transductive experimental design, including MAED [44], LROD [45], RRSS [7] and DTED [46]. The above methods perform feature and instance selection separately. It can be found that these OED series criteria aim to select the representative instances. Moreover, several methods have

also been developed in supervised scenario [47]–[50] and semi-supervised case [51], [52].

It is worthwhile to point out that all the above mentioned active learning method take all the input features to select sample. Actually, it has been widely verified that the low qualify and less discriminant features will degenerate the learning task, such as classification and clustering. Thus, it is also expected that the performance of instance selection could also be further improved by eliminating the side effect of noise features. However, there are only a few attempts to jointly select features and instances, such as UFI in [9], which takes high computational cost to obtain the greedy algorithm.

III. PRELIMINARIES

We first formulate the problem of dual selection. Let $\mathbf{X} \in \mathcal{R}^{d\times n}$ be the data matrix, where columns correspond to data instances and rows correspond to features. The goal is to simultaneously select h instances and l features such that, with the selected features as new representation and selected instances as training data, the prediction error on the testing data can be minimized.

Next, we briefly review the UFI method [9]. UFI aims at selecting a sub matrix $\mathbf{Z} \in \mathcal{R}^{l\times h}$ from the original matrix $\mathbf{X} \in \mathcal{R}^{d\times n}$ by solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{Z}} \quad & \mathrm{tr}(\mathbf{Z}\mathbf{Z}^{\mathbf{T}} + \lambda\mathbf{I})^{-1} \\ \text{s.t.} \quad & \mathbf{Z} \in \mathcal{R}^{l\times h} \text{ is a sub matrix of } \mathbf{X}. \end{aligned} \tag{1}$$

Since the above problem is NP-hard, UFI uses a greedy algorithm to solve it, where the importance of each feature and instance is evaluated individually and less informative feature (instance) is removed one by one. Such backward removal mechanism does not take special consideration on the correlations between features and instances. It also incurs high computational costs with time complexity of $O(n^4+d^4)$.

Zhang *et al.* [9] proposed UFI to perform dual selection based on AOD, which is optimized by a greedy backward searching strategy. Different from UFI, we propose to evaluate the importance of a set of features and instances simultaneously and a non-greedy algorithm is also developed. The difference of our method and UFI are as follows, 1) different formulations; 2) UFI evaluates the importance of each feature and instance one by one, while our method can evaluate a subset of features and instances jointly, which leads to better performance.

IV. UNSUPERVISED DUAL LEARNING FOR FEATURE AND INSTANCE SELECTION

In this section, we introduce our framework DFIS for dual selection. As we have mentioned before, the key intuition behind DFIS is to effectively diffuse the process of dual selection into a unified framework. With the selected features, DFIS selects those data points such that the whole data set can be best approximated. Meanwhile, it selects features to best preserve the structure of the data, which is largely determined by the selected informative samples. As a result, our dual selection method is formulated as a data reconstruction problem from the view of instance selection, and a structure preserving problem from the view of feature selection. Concretely, the selected representative data points should have the ability to reconstruct the whole data set. Inspired by [6] and [7], we minimize the following data reconstruction problem

$$\min_{\mathbf{B}} \quad ||\mathbf{X} - \mathbf{X}\mathbf{B}||_F^2 \quad \text{s.t.} \quad ||\mathbf{B}||_{2,0} = k_1, \tag{2}$$

where $\mathbf{B} \in \mathcal{R}^{n\times n}$ is the instance selection matrix and k_1 is the number of selected instances. The quality of this reconstruction may be affected by noisy features in original feature space. To alleviate the adverse effects of noisy features on instance selection, we employ a transformation matrix $\mathbf{A} \in \mathcal{R}^{d\times c}$ for feature selection (i.e., to eliminate noisy features), where $\ell_{2,0}$-norm is imposed to achieve row-sparsity. Thus, we have the following optimization problem

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B}} \quad & ||\mathbf{A}^T\mathbf{X} - \mathbf{A}^T\mathbf{X}\mathbf{B}||_F^2 \\ \text{s.t.} \quad & ||\mathbf{B}||_{2,0} = k_1, ||\mathbf{A}||_{2,0} = k_2, \mathbf{A}^T\mathbf{X}(\mathbf{A}^T\mathbf{X})^T = \mathbf{I}, \end{aligned} \tag{3}$$

where k_2 is the number of selected features. The additional constraint, i.e., $\mathbf{A}^T\mathbf{X}(\mathbf{A}^T\mathbf{X})^T = \mathbf{I}$, is used to not only avoids the arbitrary scaling problem and the trivial solution with all zeros but also ensures that data on the subspace are statistically uncorrelated. Due to the presence of $\ell_{2,0}$-norm, the optimization problem in Eq. (3) is NP-hard. To resolve this, we appeal to a useful result from [53] where the optimization w.r.t $||\mathbf{A}||_{2,0}$ can be nearly identical or approximated by the minimization w.r.t $||\mathbf{A}||_{2,1}$. Therefore, the optimization problem in Eq. (3) can be relaxed to

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B}} \quad & ||\mathbf{A}^T\mathbf{X} - \mathbf{A}^T\mathbf{X}\mathbf{B}||_F^2 + \alpha||\mathbf{A}||_{2,1} + \beta||\mathbf{B}||_{2,1} \\ \text{s.t.} \quad & \mathbf{A}^T\mathbf{X}(\mathbf{A}^T\mathbf{X})^T = \mathbf{I} \end{aligned} \tag{4}$$

where α and β are parameters.

The above optimization problem involves both the feature selection matrix $\mathbf{A}$ and the instance selection matrix $\mathbf{B}$. It is worthwhile to further analyze the different role of these two variables in the dual learning task. When $\mathbf{A}$ is fixed and let $\mathbf{X}' = \mathbf{A}^T\mathbf{X}$, then Eq. (4) reduces to an instance selection method with the following form

$$\begin{aligned} \min_{\mathbf{B}} \quad & ||\mathbf{X}' - \mathbf{X}'\mathbf{B}||_F^2 + \beta||\mathbf{B}||_{2,1} \\ \text{s.t.} \quad & \mathbf{X}'(\mathbf{X}')^T = \mathbf{I}, \end{aligned} \tag{5}$$

which actually selects instances that can be used to best reconstruct the whole data set in a new feature space, where noisy and irrelevant features are eliminated by the row-sparsity constraint on $\mathbf{A}$. By alleviating the adverse effect from feature side, our method often lead to better performance for instance selection. When $\mathbf{B}$ is fixed, Eq. (4) can be simplified as the following feature selection problem

$$\min_{\mathbf{A}} \quad \mathrm{tr}(\mathbf{A}^T\mathbf{X}(\mathbf{I}-\mathbf{B})(\mathbf{I}-\mathbf{B})^T\mathbf{X}^T\mathbf{A}) + \alpha||\mathbf{A}||_{2,1}. \tag{6}$$

By denoting the graph Laplacian matrix $\mathbf{L} = (\mathbf{I}-\mathbf{B})(\mathbf{I}-\mathbf{B})^T$, the optimization problem in Eq. (6) selects those features that

can best preserve the global data structure captured by $\mathbf{L}$. Due to the row-sparsity of $\mathbf{B}$, the graph Laplacian is largely determined by the most informative instances, which alleviates the side effect of outliers from instance side. Moreover, unlike most existing unsupervised feature selection algorithms [1], [14], [54] using a pre-fixed graph Laplacain, the inherent structure within $\mathbf{L}$ in our method can be gradually improved by eliminating the less informative instances. Therefore, our method can also achieve better results on feature selection.

Starting from all features and instances, the key intuition behind DFIS is to effectively diffuse the process of dual selection into a unified framework. From the view of instance selection, we use the selected features detected by $\mathbf{A}^T\mathbf{X}$, instead of using all the relevant and irrelevant features. With the selected features, we select these data points such that all the data can be well reconstructed by the selected ones (see Eq. (5)). Similarly, we select the feature subset to best preserve the inherent structure of the data, where the structure is largely determined by the most informative instances (see Eq. (5)). By leveraging the interactions between these two tasks, the dual learning framework could achieve better results on both tasks.

A. ALGORITHM TO SOLVE DFIS

We present an efficient algorithm to solve the optimization problem in Eq. (4). Let

$$\mathcal{L}(\mathbf{A},\mathbf{B}) = ||\mathbf{A}^T\mathbf{X} - \mathbf{A}^T\mathbf{X}\mathbf{B}||_F^2 + \alpha||\mathbf{A}||_{2,1} + \beta||\mathbf{B}||_{2,1},$$

we have two variables $\mathbf{A}$ and $\mathbf{B}$. In order to deal with the non-smoothness of the sparsity induced ℓ_{21}-norm in Eq. (2), we develop an coordinate descent algorithm to alternatively minimize the above objective function with respect to $\mathbf{A}$ and $\mathbf{B}$ respectively. This process is repeated until convergence (see Algorithm 1).

1) OPTIMIZE A WITH FIXED B

The optimization problem for updating $\mathbf{A}$ is equivalent to minimize the following objective function

$$\mathcal{L}_1 = \mathrm{tr}(\mathbf{A}^T\mathbf{X}(\mathbf{I}-\mathbf{B})(\mathbf{I}-\mathbf{B})^T\mathbf{X}^T\mathbf{A}) + \alpha||\mathbf{A}||_{2,1} \quad (7)$$

with the constraint $\mathbf{A}^T\mathbf{X}(\mathbf{A}^T\mathbf{X})^T = \mathbf{I}$. Let $\mathbf{L} = \mathbf{X}(\mathbf{I}-\mathbf{B})(\mathbf{I}-\mathbf{B})^T\mathbf{X}^T$, then, Eq. (7) can be easily rewritten as

$$\mathcal{L}_1 = \mathrm{tr}(\mathbf{A}^T\mathbf{L}\mathbf{A}) + \alpha||\mathbf{A}||_{2,1} = \mathrm{tr}(\mathbf{A}^T(\mathbf{L}+\alpha\mathbf{S})\mathbf{A}) \quad (8)$$

where $\mathbf{S}$ is a diagonal matrix with $s_{ii} = \frac{1}{2||\mathbf{a}_i||_2}$. To avoid zero values, we use a very small constant ϵ to regularize $s_{ii} = \frac{1}{2||\mathbf{a}_i||_2+\epsilon}$. This problem can be solved by generalized eigen-decomposition $(\mathbf{L}+\alpha\mathbf{S})\mathbf{A} = \Lambda\mathbf{X}\mathbf{X}^T\mathbf{A}$, where Λ is a diagonal matrix whose diagonal elements are eigenvalues.

2) OPTIMIZE B WITH FIXED A

The optimization problem of updating $\mathbf{B}$ is equivalent to minimize the following objective function

$$\mathcal{L}_2 = ||\mathbf{A}^T\mathbf{X} - \mathbf{A}^T\mathbf{X}\mathbf{B}||_F^2 + \beta||\mathbf{B}||_{2,1}. \quad (9)$$

Let $\frac{\partial\mathcal{L}_2}{\partial\mathbf{B}} = 0$, that is,

$$\frac{\partial\mathcal{L}_2}{\partial\mathbf{B}} = -2\mathbf{X}^T\mathbf{A}\mathbf{A}^T\mathbf{X} + 2\mathbf{X}^T\mathbf{A}\mathbf{A}^T\mathbf{X}\mathbf{B} + 2\beta\mathbf{D}\mathbf{B} = 0,$$

where $\mathbf{D}$ is a diagonal matrix with $d_{ii} = \frac{1}{2||\mathbf{b}_i||_2}$. We also use a very small constant ε to regularize $d_{ii} = \frac{1}{2||\mathbf{b}_i||_2+\epsilon}$. Then we get the following close-form solution to $\mathbf{B}$,

$$\mathbf{B} = (\mathbf{X}^T\mathbf{A}\mathbf{A}^T\mathbf{X} + \beta\mathbf{D})^{-1}\mathbf{X}^T\mathbf{A}\mathbf{A}^T\mathbf{X}. \quad (10)$$

Note that the matrix $\mathbf{X}^T\mathbf{A}\mathbf{A}^T\mathbf{X}+\beta\mathbf{D}$ is a $n\times n$ matrix, so calculating the inverse matrix takes $O(n^3)$ time complexity.

Algorithm 1 The Optimization Algorithm of DFIS

Input: data matrix $\mathbf{X}\in\mathcal{R}^{d\times n}$, parameters α and β.

1: Set $t = 0$, and initialize $\mathbf{S}\in\mathcal{R}^{d\times d}$ and $\mathbf{D}\in\mathcal{R}^{n\times n}$ as identity matrices
2: **repeat**
3: $\mathbf{B}^t = (\mathbf{X}^T\mathbf{A}\mathbf{A}^T\mathbf{X} + \beta\mathbf{D}^t)^{-1}\mathbf{X}^T\mathbf{A}\mathbf{A}^T\mathbf{X}$
4: Update the diagonal matrix $\mathbf{D}^{t+1}$ as
$$\mathbf{D}^{t+1} = \begin{pmatrix} \frac{1}{2||\mathbf{B}_1^{t+1}||_2+\epsilon} & & \\ & \cdots & \\ & & \frac{1}{2||\mathbf{B}_n^{t+1}||_2+\epsilon} \end{pmatrix}$$
5: $\mathbf{L}^t = \mathbf{X}(\mathbf{I}-\mathbf{B}^t)(\mathbf{I}-\mathbf{B}^t)^T\mathbf{X}^T$
6: $\mathbf{A}^t = [\mathbf{p}_1,\mathbf{p}_2,\mathbf{p}_3,\ldots,\mathbf{p}_c]$, where $\mathbf{p}_1,\mathbf{p}_2,\ldots,\mathbf{p}_c$ are the eigenvectors of $(\mathbf{L}+\alpha\mathbf{S})\mathbf{A} = \Lambda\mathbf{X}\mathbf{X}^T\mathbf{A}$ corresponding to the first c smallest eigenvalues
7: Update the diagonal matrix $\mathbf{S}^{t+1}$ as
$$\mathbf{S}^{t+1} = \begin{pmatrix} \frac{1}{2||\mathbf{A}_1^{t+1}||_2+\epsilon} & & \\ & \cdots & \\ & & \frac{1}{2||\mathbf{A}_d^{t+1}||_2+\epsilon} \end{pmatrix}$$
8: $t = t+1$
9: **until** Convergence;

Output: Sort all features and all instances according to $||\mathbf{a}_i||_2 \quad (i = 1,2,\ldots,d)$ and $||\mathbf{b}_j||_2 \quad (j = 1,2,\ldots,n)$ in descending order respectively and select top l ranked features and top h ranked instances.

B. CONVERGENCE ANALYSIS

We analyze the convergence of the proposed optimization algorithm.

Theorem 1: Updating $\mathbf{B}$ *using Eq.* (10) *will monotonically decrease the objective function in Eq.* (7).

Proof: Following a similar procedure in [55], we can prove the following inequation holds in the t-th step. Given $\mathbf{A}^t$ and $\mathbf{B}^t$ at t-th iteration, the optimal solution of $\mathbf{B}^{t+1}$ can be obtained as follows

$$\begin{aligned}\mathbf{B}^{t+1} &= \min_{\mathbf{B}} ||(\mathbf{A}^t)^T\mathbf{X} - (\mathbf{A}^t)^T\mathbf{X}\mathbf{B}||_F^2 + \beta\mathrm{tr}(\mathbf{B}^T\mathbf{D}^t\mathbf{B}) \\ &\le ||(\mathbf{A}^t)^T\mathbf{X} - (\mathbf{A}^t)^T\mathbf{X}\mathbf{B}^t||_F^2 + \beta\mathrm{tr}((\mathbf{B}^t)^T\mathbf{D}^t\mathbf{B}^t)\end{aligned}$$

According to the definition of $d_{ii}^t = \frac{1}{2(||\mathbf{b}_i^t||_2+\epsilon)}$, the following inequation holds

$$||(\mathbf{A}^t)^T\mathbf{X} - (\mathbf{A}^t)^T\mathbf{X}\mathbf{B}^{t+1}||_F^2 + \beta||\mathbf{B}^{t+1}||_{2,1}$$

$$
\begin{aligned}
&- \beta(||\mathbf{B}^{t+1}||_{2,1} - \sum_i \frac{||\mathbf{b}_i^{t+1}||_2^2}{2||\mathbf{b}_i^t||_2}) \\
&\leq ||(\mathbf{A}^t)^T\mathbf{X} - (\mathbf{A}^t)^T\mathbf{X}\mathbf{B}^t||_F^2 + \beta||\mathbf{B}^t||_{2,1} \\
&\quad - \beta(||\mathbf{B}^t||_{2,1} - \sum_i \frac{||\mathbf{b}_i^t||_2^2}{2||\mathbf{b}_i^t||_2})
\end{aligned}
$$

According to the inequation $\sqrt{a} - \frac{a}{2\sqrt{b}} \leq \sqrt{b} - \frac{b}{2\sqrt{b}}$ in [53], we have

$$
\begin{aligned}
&||(\mathbf{A}^t)^T\mathbf{X} - (\mathbf{A}^t)^T\mathbf{X}\mathbf{B}^{t+1}||_F^2 + \beta||\mathbf{B}^{t+1}||_{2,1} \\
&\quad \leq ||(\mathbf{A}^t)^T\mathbf{X} - (\mathbf{A}^t)^T\mathbf{X}\mathbf{B}^t||_F^2 + \beta||\mathbf{B}^t||_{2,1} \qquad (11)
\end{aligned}
$$

By combining Equations (7) and (11), we have

$$
\mathcal{L}(\mathbf{A}^t, \mathbf{B}^{t+1}) \leq \mathcal{L}(\mathbf{A}^t, \mathbf{B}^t). \qquad (12)
$$

Thus, Theorem 1 is proved. □

Theorem 2: Updating $\mathbf{A}$ *using the procedure in Algorithm 1 monotonically decreases the objective function in Eq.* (7).

Proof: Following a similar procedure in [3], we can prove the following inequation holds in the t-th step. Given $\mathbf{A}^t$ and $\mathbf{B}^t$ at t-th iteration, the optimal solution of $\mathbf{A}^{t+1}$ can be obtained as follows

$$
\begin{aligned}
\mathbf{A}^{t+1} &= \min_{\mathbf{A}} \mathrm{tr}(\mathbf{A}^T[\mathbf{X}(\mathbf{I} - \mathbf{B}^t)(\mathbf{I} - \mathbf{B}^t)^T\mathbf{X}^T + \alpha\mathbf{S}]\mathbf{A}) \\
&\leq \mathrm{tr}((\mathbf{A}^t)^T[\mathbf{X}(\mathbf{I} - \mathbf{B}^t)(\mathbf{I} - \mathbf{B}^t)^T\mathbf{X}^T + \alpha\mathbf{S}]\mathbf{A}^t)
\end{aligned}
$$

According to the definition of $s_{ii}^t = \frac{1}{2(||\mathbf{a}_i^t||_2 + \epsilon)}$ and $\mathbf{L} = (\mathbf{I} - \mathbf{B}^t)(\mathbf{I} - \mathbf{B}^t)^T$, the following inequation holds

$$
\begin{aligned}
&\mathrm{tr}((\mathbf{A}^{t+1})^T\mathbf{L}\mathbf{A}^{t+1}) + \alpha||\mathbf{A}^{t+1}||_{2,1} \\
&\quad - \alpha(||\mathbf{A}^{t+1}||_{2,1} - \sum_i \frac{||\mathbf{a}_i^{t+1}||_2^2}{2||\mathbf{a}_i^t||_2}) \\
&\leq \mathrm{tr}((\mathbf{A}^t)^T\mathbf{L}\mathbf{A}^t) + \alpha||\mathbf{A}^t||_{2,1} \\
&\quad - \alpha(||\mathbf{A}^t||_{2,1} - \sum_i \frac{||\mathbf{a}_i^t||_2^2}{2||\mathbf{a}_i^t||_2})
\end{aligned}
$$

According to the inequation $\sqrt{a} - \frac{a}{2\sqrt{b}} \leq \sqrt{b} - \frac{b}{2\sqrt{b}}$ in [53], we have

$$
\begin{aligned}
&\mathrm{tr}((\mathbf{A}^{t+1})^T\mathbf{L}\mathbf{A}^{t+1}) + \alpha||\mathbf{A}^{t+1}||_{2,1} \\
&\quad \leq \mathrm{tr}((\mathbf{A}^t)^T\mathbf{L}\mathbf{A}^t) + \alpha||\mathbf{A}^t||_{2,1} \qquad (13)
\end{aligned}
$$

By combining Eq. (7) and Eq. (13), we have

$$
\mathcal{L}(\mathbf{A}^{t+1}, \mathbf{B}^{t+1}) \leq \mathcal{L}(\mathbf{A}^t, \mathbf{B}^{t+1}). \qquad (14)
$$

□

Theorem 3: The alternating update rules in Algorithm 1 monotonically decrease the objective function of Eq. (7).

Proof: Combining Eq. (12) and Eq. (14), we can get

$$
\mathcal{L}(\mathbf{A}^{t+1}, \mathbf{B}^{t+1}) \leq \mathcal{L}(\mathbf{A}^t, \mathbf{B}^{t+1}) \leq \mathcal{L}(\mathbf{A}^t, \mathbf{B}^t). \qquad (15)
$$

Thus, the objective function in Eq. (7) monotonically decreases by updating rules in algorithm 1 and Theorem 3 is proved. □

Since the function in Eq. (7) is non-negative, it is lower bounded. Based on Theorem 3, the alternating update rules in Algorithm 1 monotonically decrease the objective function of Eq. (7), thus the proposed optimization algorithm is converged.

C. COMPLEXITY ANALYSIS

When updating $\mathbf{B}$, the most consuming process is to compute the inverse matrix of a $n \times n$ matrix, which leads to $O(n^3)$ time complexity. The time complexity of updating $\mathbf{A}$ is $O(d^3)$. So the overall time complexity of our algorithm is $O(n^3)$ with respect to n and $O(d^3)$ with respect to d.

D. DISCUSSION

We discuss the relationships of our proposed model with existing feature and instance selection models.

1) CONNECTION WITH TED

TED [6] aims at minimizing the assessed uncertainty of predictions on given unlabeled data, while solving the following problem

$$
\min_{\beta, \boldsymbol{\alpha}_i \in \mathcal{R}^n} \quad \sum_{i=1}^n ||\mathbf{x}_i - \mathbf{X}^T\boldsymbol{\alpha}_i||^2 + \sum_{j=1}^n \frac{\alpha_{i,j}^2}{\beta_j} + \gamma||\beta||_1 \qquad (16)
$$

If we fix $\mathbf{A}$ and let $\mathbf{G} = \mathbf{A}^T\mathbf{X}$, our optimization problem in Eq. (4) can be reduced to minimize the following objective function

$$
\min_{\mathbf{B}} \quad ||\mathbf{G} - \mathbf{G}\mathbf{B}||_F^2 + \beta||\mathbf{B}||_{2,1} \qquad (17)
$$

Both our method and TED select those data points to best reconstruct all the data samples via the sparsity-induced norm (i.e., ℓ_{21} and ℓ_1). However, our method actually selects instances in a reduced feature space, which avoids the adverse effect of noisy and irrelevant features. Thus, it's expected to achieve a better performance.

2) CONNECTION WITH FEATURE SELECTION METHODS

If we fix B, our objective is reduced to

$$
\begin{aligned}
&\min_{\mathbf{A}} \quad \mathrm{tr}(\mathbf{A}^T\mathbf{X}(\mathbf{I} - \mathbf{B})(\mathbf{I} - \mathbf{B})^T\mathbf{X}^T\mathbf{A}) + \alpha||\mathbf{A}||_{2,1} \\
&\text{s.t.} \quad \mathbf{A}^T\mathbf{X}(\mathbf{A}^T\mathbf{X})^T = \mathbf{I}. \qquad (18)
\end{aligned}
$$

When compared with Laplacian Score [2] and MCFS [1], our method is superior as it selects a subset of feature simultaneously rather than individually, which can better model the correlation among features. When compared with UDFS [3] and NDFS [55], which construct the graph Laplacian by local information, our method uses sparse reconstruction coefficient to capture the global intrinsic structure. What's more, the graph Laplacian of our method can be adaptively updated to better capture the data structure in the learning process.

3) CONNECTION WITH UFI

The objective function of UFI is given in Eq. (1). An equivalent form of our DFIS framework is given as follows

$$\begin{aligned} \max_{\mathbf{Z}'} \quad & \mathrm{tr}[\mathbf{X}'^{T}\mathbf{Z}'(\mathbf{Z}'^{T}\mathbf{Z}' + \nu\mathbf{I})^{-1}\mathbf{Z}'^{T}\mathbf{X}'] \\ \text{s.t.} \quad & \mathbf{Z} \in \mathbf{X}, \mathbf{Z}' = \mathbf{A}^{T}\mathbf{Z}, \mathbf{X}' = \mathbf{A}^{T}\mathbf{X}, |\mathbf{Z}| = k_1, |\mathbf{A}| = k_2, \end{aligned} \tag{19}$$

As shown in Eq. (1), UFI aims at selecting informative features and instances to minimize the size of parameter covariance matrix. Similar to [41], our dual selection method can be interpreted as jointly selecting features and instances to minimize the uncertainty of prediction on given unlabeled data. UFI evaluates the importance of each feature and instance individually. Our method evaluates the importance of a subset of features and instances jointly, which can better model the correlations among features and instances. The non-greedy optimization scheme of DFIS is also more efficient than the greedy backward selection mechanism of UFI.

V. EXPERIMENTS

In this section, we evaluate the performance of DIFS. Following the similar experimental protocol in [9], we train classification model on the selected instances and make predictions on the remaining instances. The prediction accuracy is used to measure the performance of each method.

A. DATA SETS

We conduct experiments on two face image data sets, i.e. ORL and YALE data and one document data set CSTR. We give brief description about these data sets.

Yale database contains 165 gray scale images in GIF format from 15 individuals. There are 11 images per subject, one per different facial expression or configuration.

ORL database contains 400 images from 40 distinct subjects. The images were taken at different times, varying the lighting, facial expressions and facial details.

CSTR database contains 476 abstracts of technical reports published in the Department of Computer Science at the University of Rochester between 1991 and 2002. The abstracts are divided into four research areas: Natural Language Processing, Robotics/Vision, Systems and Theory.

B. EXPERIMENTAL SETTINGS

We compare our proposed method with 12 carefully designed baselines and UFI [9] method. We first choose LS [2] and MCFS [1] as the feature selection methods, and choose AOD [5] and cTED [6] as active learning methods. By adopting different ways to couple the results of feature selection and active learning, we have the following 12 combinations for dual selection.

- **LS+AOD**. This combination selects subset of features based on all instances via Laplacian Score [2], and selects subset of samples based on all features via AOD [5]. The SVM classifier is finally trained and evaluated based on the selected samples and features.
- **LS+cTED**. Similar with LS+AOD, except that the instances are selected via cTED [6] method.
- **MCFS+AOD**. Similar with LS+AOD, except that the features are selected via MCFS [1] method.
- **MCFS+cTED**. Similar with LS+AOD, except that the features are selected via MCFS [1] method and the instances are selected via cTED [6] method.
- **LS2AOD**. This combination first selects subset of features based on all instances via Laplacian Score [2], and then selects subset of samples only based on selected features via AOD [5]. The SVM classifier is finally trained and evaluated based on the selected samples and features.
- **LS2cTED**. Similar with LS2AOD, except that the instances are selected via cTED [6] method.
- **MCFS2AOD**. Similar with LS2AOD, except that the features are selected via MCFS [1] method.
- **MCFS2cTED**. Similar with LS2AOD, except that the features are selected via MCFS [1] method and the instances are selected via cTED [6] method.
- **AOD2LS**. This combination first selects subset of instances based on all features via AOD [5], and then selects subset of features only based on selected instances via Laplacian Score [2]. The SVM classifier is finally trained and evaluated based on the selected samples and features.
- **AOD2MCFS**. Similar with AOD2LS, except that the features are selected via MCFS [1] method.
- **cTED2LS**. Similar with AOD2LS, except that the instances are selected via cTED [6] method.
- **cTED2MCFS**. Similar with AOD2LS, except that the instances are selected via cTED [6] method and the features are selected via MCFS [1] method.

Similar with [9], these 12 baselines can be categorized into three groups. The first category have 4 methods, i.e., LS+cTED, MCFS+cTED, LS+AOD and MCFS+AOD. These combinations perform feature selection and instance selection independently. The selected features and instances are combined finally. The second category also have 4 combinations, i.e., LS2AOD, LS2cTED, MCFS2AOD and MCFS2cTED. These methods first select subset of features based on all samples and then select informative instances only based on selected features. The third category also have 4 methods, i.e., AOD2LS, AOD2MCFS, cTED2LS and cTED2MCFS. These combinations first select subset of instances based on all features and then select informative features only based on selected instances.

Once the subset of features and subset of samples are selected, we use these selected samples represented by selected features and their labels to train the linear SVM model, and evaluate the classification accuracy of SVM classifier on the rest unlabeled data which are also represented by selected features. We conduct one-against-all classification to handle the problem of multi-class classification as UFI did. If there are c classes in the data set, c binary classifiers are

trained. In the prediction stage, all these c classifiers are used for each instance. The class label of each instance is determined by the classifier with the largest output value. SVM [56] with linear kernel is used in our experiments. To fairly compare the above unsupervised algorithms, we tune the parameters for all methods by a "grid-search strategy" from a large range of $\{10^{-3}, 10^{-2}, \ldots, 10^{6}\}$ and report the best result each method can achieve.

C. EXPERIMENTAL RESULTS

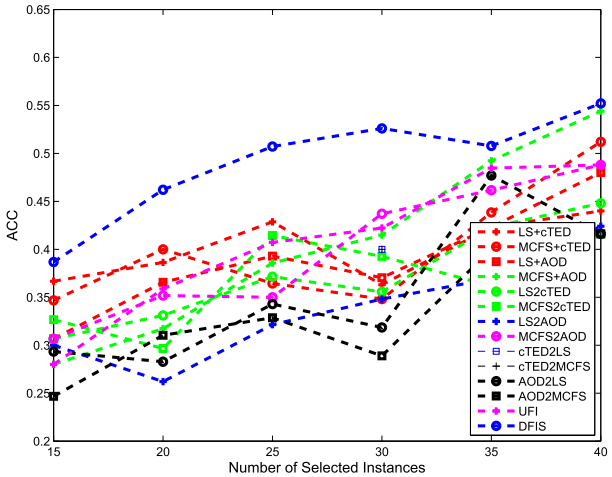
On the YALE data set, we apply all the comparing methods to select $k = \{300, 500, 700\}$ features and $m = \{15, 20, \ldots, 40\}$ instances. The classification results are shown in Figure 1. At a first glance, our method can achieve the best result. In most of the cases, UFI can achieve the second best performance. If we fix the number of selection instance as 25, we can see in Figure 1(a) that our method achieve 0.5 classification accuracy with 300 features. To achieve the same performance, the best baseline requires 700 features, as shown in Figure 1(c). This indicates that our method can select more informative features than other methods. If we fix the selected features as 500, we see in Figure 1(b) that our method can achieve 0.55 classification accuracy with 25 instances. To achieve the same performance, the best baseline needs 40 instances. This indicates that the selected instances by our method are more informative. We also observe that, our method can achieve the best classification accuracy (about 0.6) with 500 features and 30 instances. It needs the best baseline method select 700 features and 40 instances to approximate 0.6. On ORL and CSTR data sets, our method can also achieve the best performance as shown in Figure 2 and Figure 3. In summary, our proposed method can achieve better performance than other methods in most of the cases.

D. EFFECT OF DUAL SELECTION STRATEGY

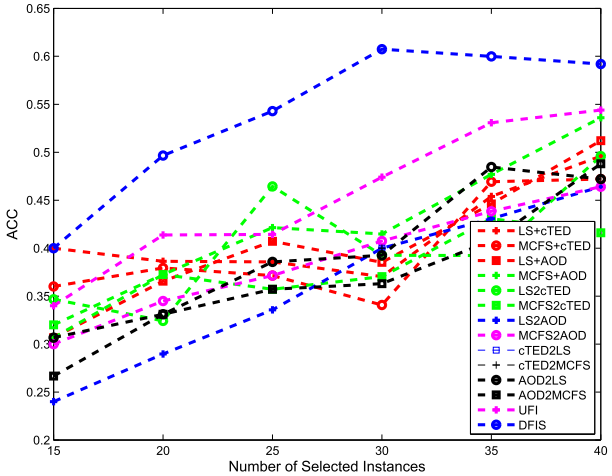
Here, we investigate the effect of dual selection within one unified framework by empirically answering the first question as follows,

- Whether the performance of active learning, i.e., instance selection, can be further improved by the embedding of feature selection?

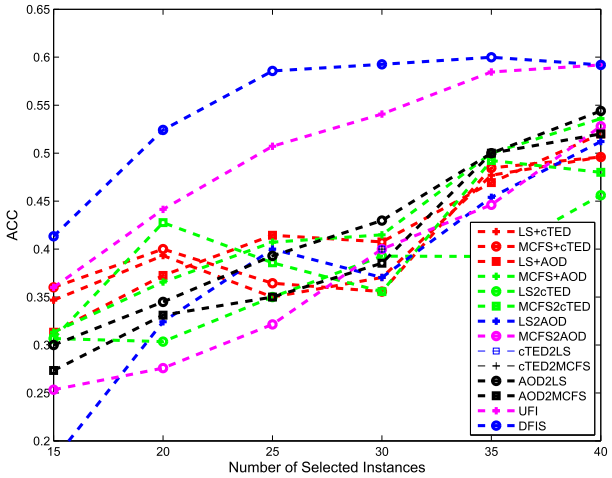
We conduct the following experiment to present qualitative research on the insight of bringing more separable representation for the task of instance selection via the embedding of feature selection. Here, we take a subset of ORL data set with 40 samples, where all these samples can be found in Figure 4(a), as an example for clear illustration. We perform principle component analysis (PCA) on this data set with all 1024 features and project these 40 samples onto the first two principle components, as shown in Figure 4(b). We run DFIS on this data set and select top 300 and 500 features according to the results of DFIS. Then we further perform PCA on these 40 samples with selected top 300 and 500 features. We also show the first two principle components



(a) 300 features



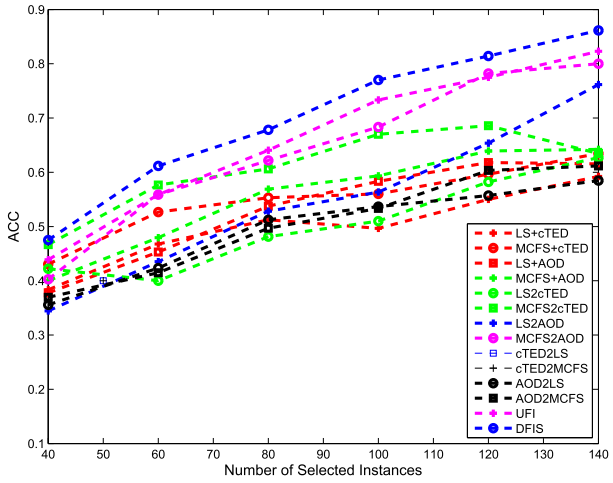
(b) 500 features



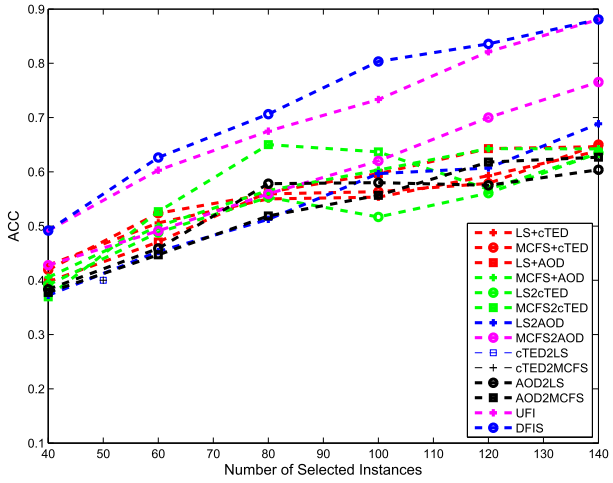
(c) 700 features

FIGURE 1. Comparisons with baselines on the Yale data set.

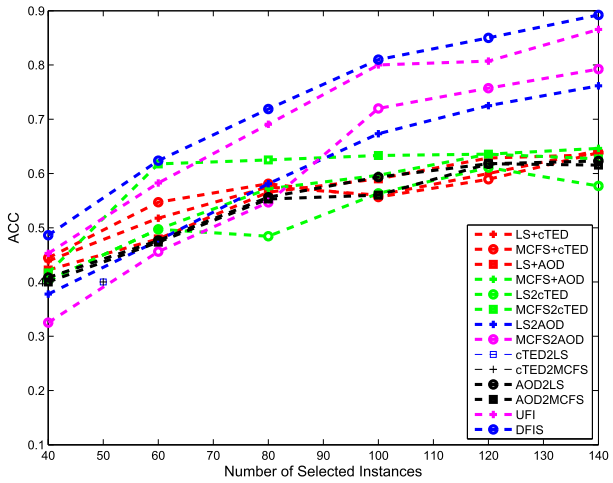
in Figure 4(c) and Figure 4(d). It can be seen that projected samples using top 300 features or top 500 features selected by DFIS, are more separable than that of using all 1024 features. This visualization shows the effectiveness of the dual learning



(a) 300 features

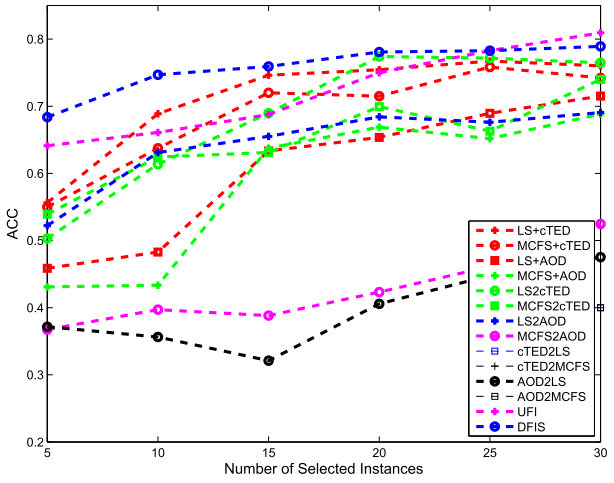


(b) 500 features

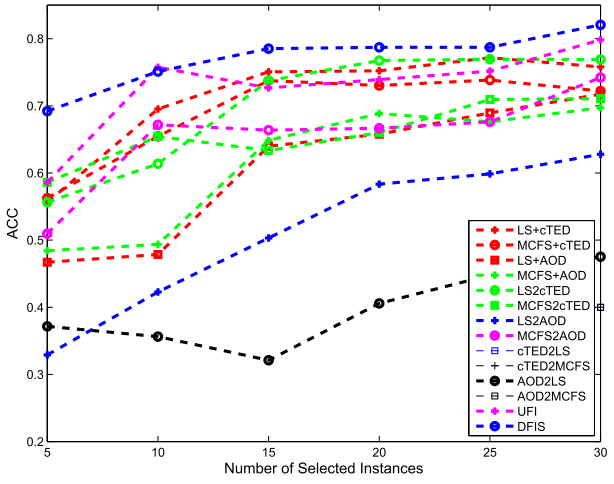


(c) 700 features

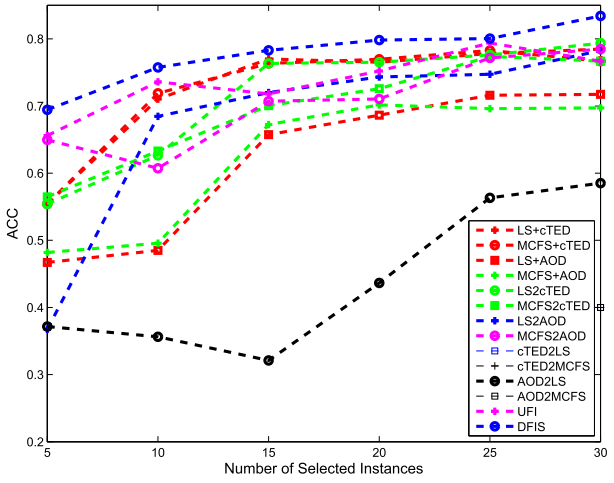
FIGURE 2. Comparisons with baselines on the ORL data set.



(a) 300 features



(b) 500 features



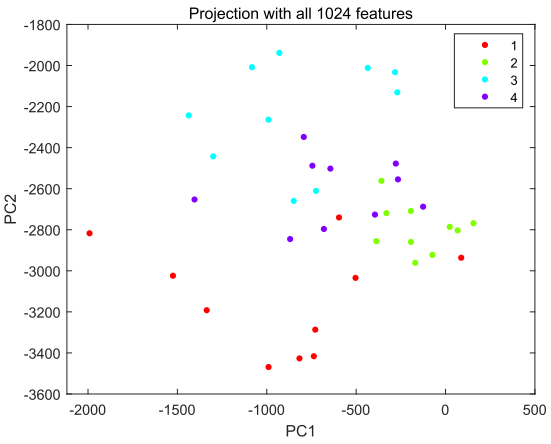
(c) 700 features

FIGURE 3. Comparisons with baselines on the CSTR data set.

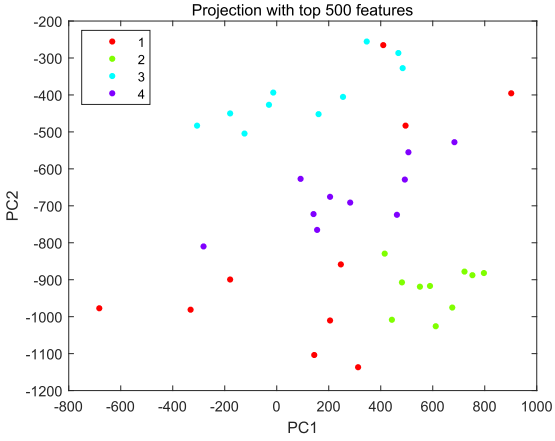
strategy to perform feature selection for the task of instance selection and subsequent classification.

Moreover, we conduct the following experiment to present quantitative research on the improvement of using feature selection technique for the task of instance selection. In this new experiment, we take the cTED method [6] with all features on CSTR data set as the baseline active learning result. Then we provide the results of DFIS with different size of

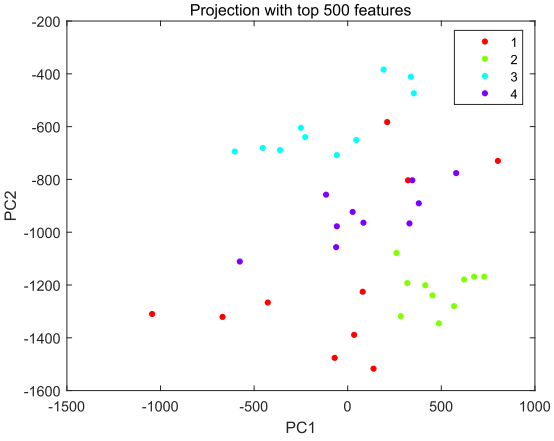
(a) Subset of ORL with 40 samples



(b) Projection on first two principle components using 1024 features



(c) Projection on first two principle components using top 300 features



(d) Projection on first two principle components using top 500 features

FIGURE 4. Subset of ORL dataset and projection on its first two principle components with different feature subset. The horizontal axis is the score of the first principle component, and the vertical axis is the score of the second principle component. Different color marks samples of different classes.

TABLE 1. Classification results on CSTR.

# of instances	5	10	15	20	25	30
cTED (1000)	55.70	72.32	79.18	78.07	78.05	78.48
DFIS (300)	**68.37**	**74.68**	**75.92**	**78.07**	**78.27**	**78.92**
DFIS (500)	**69.21**	**75.11**	**78.52**	**78.73**	**78.71**	**82.06**
DFIS (700)	**69.43**	**75.75**	**78.31**	**79.82**	**80.04**	**83.41**

feature subset, i.e., 300, 500 and 700 features in Table 1. Compared with cTED with all 1000 features, the performance of DFIS is not degenerated with less features. Actually, it can be seen that DFIS with less features achieves better results than cTED with all features. Intuitively, such improvements can be contributed to that the dual selection method can identify the most informative features for instance selection and the subsequent classification task. Mathematically, the main difference between cTED in Eq. (16) and DFIS in Eq. (4) is the joint embedding of feature selection. Thus we can conclude that the dual selection strategy of DFIS is indeed helpful for the instance selection task.

TABLE 2. Classification results on Yale subset.

	Number of selected features					
	300	400	500	600	700	1024
10	20.00	20.00	22.50	20.00	20.00	**22.50**
20	22.50	22.50	25.00	20.00	25.00	**22.50**
30	**25.00**	22.50	**27.50**	**25.00**	25.00	20.00
40	**25.00**	**25.00**	25.00	20.00	**27.50**	**22.50**
50	17.50	17.50	20.00	20.00	20.00	15.00
60	17.50	17.50	20.00	17.50	20.00	17.50
70	20.00	22.50	17.50	17.50	17.50	17.50
80	20.00	20.00	17.50	17.50	17.50	20.00
90	17.50	17.50	17.50	17.50	17.50	15.00
100	20.00	17.50	17.50	17.50	20.00	15.00
110	17.50	20.00	20.00	17.50	20.00	15.00

Now, we aim to further empirically answer the following question.

- Compared with all candidate samples, whether the selection of fewer samples is useful for informative feature selection?

We take the first 110 samples from YALE data set with 10 classes, then split these 110 samples into 70 candidate set and 40 test samples. Then we additionally

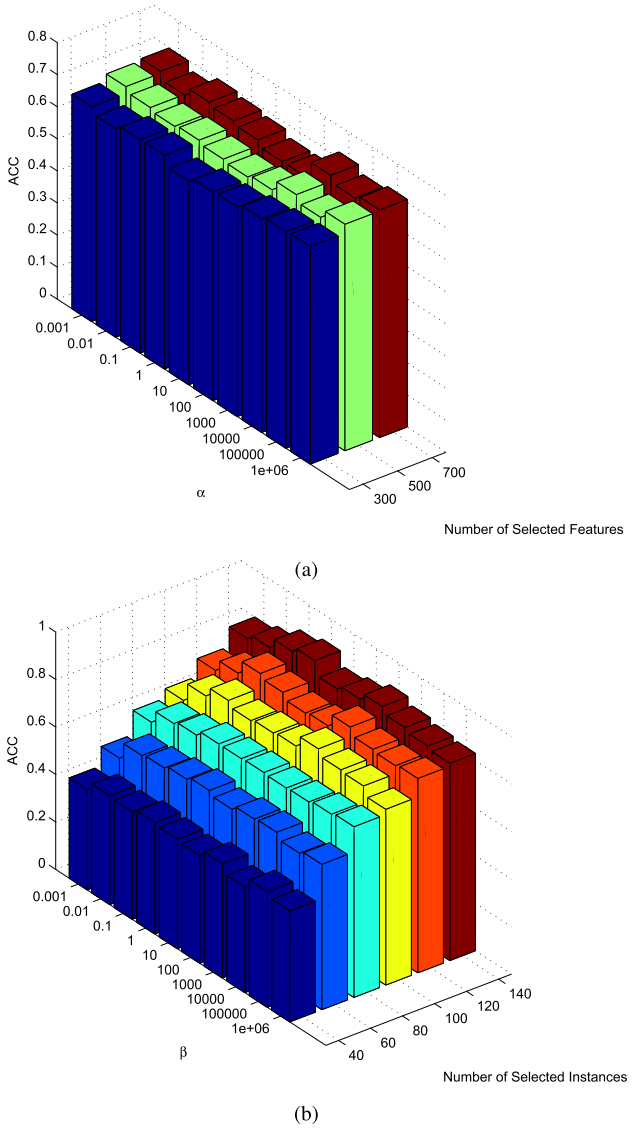


FIGURE 5. **Classification Accuracy with different parameters on ORL data set.**

select 40 samples in the rest 55 samples. To simulate more practical and challenging real scenario, we purposely edit the label of these 40 samples with $[1, 2, \ldots, 10]$, and get 40 noisy samples. Now we conduct dual learning via DFIS on the combined data set with 70 candidate samples and 40 noisy samples. With different size of selected samples and features, we evaluate the performance of DFIS on the rest 40 test samples. The classification accuracy via SVM is present in Table 2. It can be seen that for the different size of selected features, such as 300 selected features in the first column, the best results are achieved by select 30 or 40 samples. That is to say, the usage of all 110 noisy samples does not improve the classification results. It can also been find that, all the best results in each column are achieved with few labeled samples, not all 110 samples. In summary, we conclude that the feature selection procedure can also be improved by using selected informative samples. Such results well justify the motivation and the correctness of the proposed method.

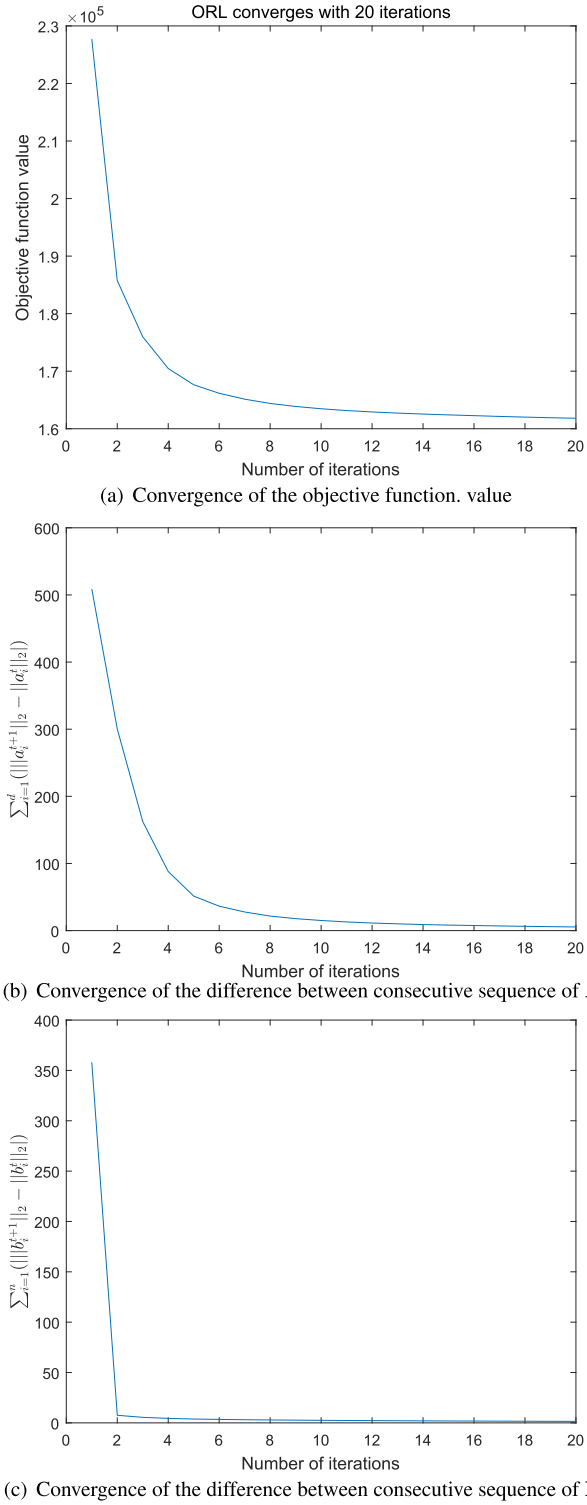


FIGURE 6. **Convergence behavior of DFIS on ORL data set.**

E. SENSITIVE ANALYSIS AND CONVERGENCE ANALYSIS

We now study the sensitiveness of parameters. Due to space limit, we only report the results on ORL data set. The performance with respect to the parameter α and different number of selected features is provided in Figure 5(a), where the number of selected instances is fixed as 60. We can see that, the

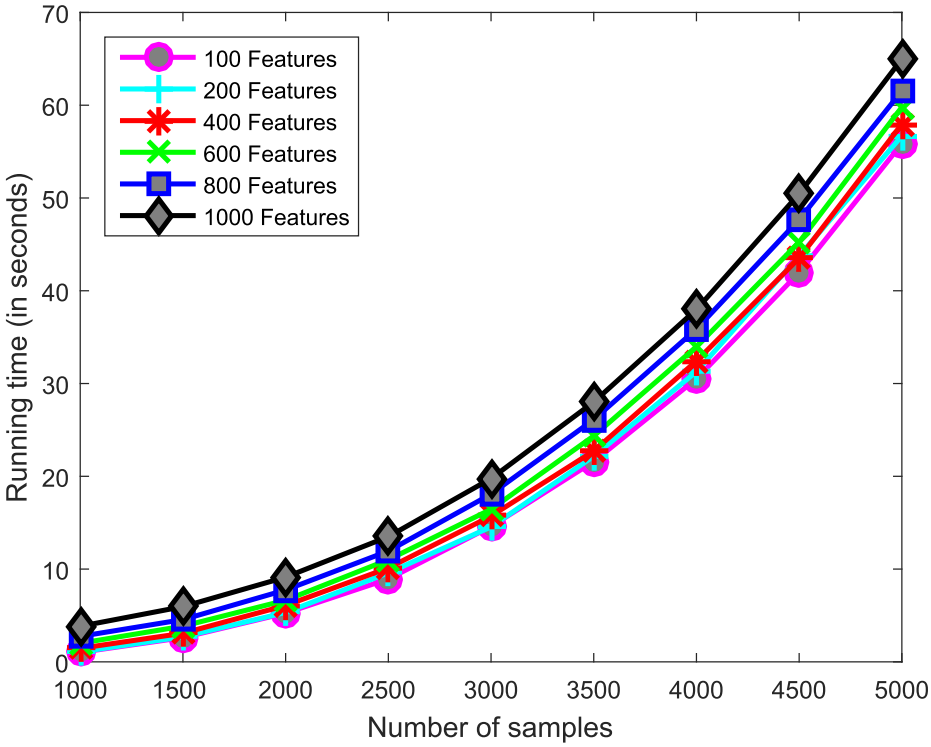


FIGURE 7. Computational cost on synthetic data with different size of samples and features.

performance of our method is not sensitive to α and increases as the number of selected feature increases. The performance with respect to the parameter β and selected instance number can be found in Figure 5(b), where the number of selected features is fixed as 700. We see that, the performance of our method is not sensitive to β and increases as the number of selected instances increases.

In the next, we conduct experiment to demonstrate the convergence behavior of DFIS on ORL data set. We plot the decreasing curve of the objective function value in Figure 6(a). Since the purpose of DFIS is to select the top features and top samples, we further compute the changes between consecutive sequences of $\{\mathbf{A}^t\}$ and $\{\mathbf{B}^t\}$. We use the following criteria to measure the convergence of the feature score and sample score among iterations

$$\text{feaScoreDiff} = \sum_{i=1}^{d}(|||\mathbf{a}_i^{t+1}||_2 - ||\mathbf{a}_i^t||_2|), \tag{20}$$

$$\text{smpScoreDiff} = \sum_{i=1}^{n}(|||\mathbf{b}_i^{t+1}||_2 - ||\mathbf{b}_i^t||_2|). \tag{21}$$

The results of feature score difference and sample score difference between two consecutive iterations are present in Figure 6(b) and Figure 6(c). These three figures well demonstrate the convergence behavior of the proposed optimization schema for DFIS. It can also be seen that DFIS often converges in few iterations.

Although the complexity of DFIS has been provided in SectionIV-C, we take additional experiment to show the computation cost of DFIS. Here, we randomly generate a data set with different size of samples and features. The size of samples changes from [1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000] and the size of features comes from [100, 200, 400, 600, 800, 1000]. Then we run DFIS on this toy data set with different size of samples and features. We record and plot the computational time for DFIS on these different combination of feature and sample in Figure 7. The code of DFIS is implemented by Matlab 2015b. The experiment is conducted on a 3.6-GHz Windows machine. It can be seen that DFIS takes more time for the increasing of samples and features.

VI. CONCLUSION

In this paper, we propose a novel method, which performs unsupervised feature selection and instance selection within an unified dual selection framework. It is expected that the whole features can be well reconstructed by the selected features and all the instances can also be approximated by the selected instances. The dual selection procedures are achieved by dual sparse regularization on both feature side and instance side. The whole dual selection model can be solved by the coordinate decent algorithm. The experimental results show that our proposed method can achieve better performance when compared with the comparing methods.

Compared with existing works on feature selection and instance selection, the main theoretical contributions of this paper are to design the dual reconstruction model and integrate these two separated and connected process within one unified framework. Compared with most closely related work, i.e., UFI, we provide non-greedy learning algorithm to solve the newly developed dual selection model. We also provide the convergence analysis and the complexity analysis of the proposed method.

Although it is a good attempt to unify two separated processes, DFIS may be further improved in several different ways. It is better to eliminate the additional tuning parameters by replacing the sparse regularization with ℓ_0 constraints to make the algorithm more practical. Due to the low quality of data, it is better to improve the robustness of the selection procedure by taking more robust loss functions. It is also important to improve the diversity of the selected samples or features, where less redundant information may be preserved.

ACKNOWLEDGMENT

The authors would like to thank the reviewers and Dr. L. Shi for their helpful suggestions to improve this article.

REFERENCES

[1] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2010, pp. 333–342.

[2] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. Neural Inf. Process. Syst.*, vol. 18, 2006, pp. 507–514.

[3] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "$l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, vol. 22, no. 1, p. 1589.

[4] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 1151–1157.

[5] A. C. Atkinson, A. N. Donev, and R. D. Tobias, *Optimum Experimental Designs, With SAS*, vol. 34. London, U.K.: Oxford Univ. Press, 2007.

[6] K. Yu, S. Zhu, W. Xu, and Y. Gong, "Non-greedy active learning for text categorization using convex ansductive experimental design," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2008, pp. 635–642.

[7] F. Nie, H. Wang, H. Huang, and C. Ding, "Early active learning via robust representation and structured sparsity," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1572–1578.

[8] H. Liu, H. Motoda, and L. Yu, "A selective sampling approach to active feature selection," *Artif. Intell.*, vol. 159, nos. 1–2, pp. 49–74, Nov. 2004.

[9] L. Zhang, C. Chen, J. Bu, and X. He, "A unified feature and instance selection framework using optimum experimental design," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2379–2388, May 2012.

[10] Y. Zhang, H.-G. Li, Q. Wang, and C. Peng, "A filter-based bare-bone particle swarm optimization algorithm for unsupervised feature selection," *Int. J. Speech Technol.*, vol. 49, no. 8, pp. 2889–2898, Aug. 2019.

[11] Y. Zhang, S. Cheng, Y. Shi, D.-W. Gong, and X. Zhao, "Cost-sensitive feature selection using two-archive multi-objective artificial bee colony algorithm," *Expert Syst. Appl.*, vol. 137, pp. 46–58, Dec. 2019.

[12] X.-F. Song, Y. Zhang, Y.-N. Guo, X.-Y. Sun, and Y.-L. Wang, "Variable-size cooperative coevolutionary particle swarm optimization for feature selection on high-dimensional data," *IEEE Trans. Evol. Comput.*, early access, Jan. 2, 2020, doi: 10.1109/TEVC.2020.2968743.

[13] Y. Zhang, D.-W. Gong, X.-Z. Gao, T. Tian, and X.-Y. Sun, "Binary differential evolution with self-learning for multi-objective feature selection," *Inf. Sci.*, vol. 507, pp. 67–85, Jan. 2020.

[14] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2012, pp. 1026–1032.

[15] N. Gu, M. Fan, L. Du, and D. Ren, "Efficient sequential feature selection based on adaptive eigenspace model," *Neurocomputing*, vol. 161, pp. 199–209, Aug. 2015.

[16] M. Fan, X. Chang, X. Zhang, D. Wang, and L. Du, "Top-k supervise feature selection via ADMM for integer programming," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1646–1653.

[17] 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.

[18] P. Zhu, Q. Xu, Q. Hu, and C. Zhang, "Co-regularized unsupervised feature selection," *Neurocomputing*, vol. 275, pp. 2855–2863, Jan. 2018.

[19] L. Du and Y.-D. Shen, "Joint clustering and feature selection," in *Proc. Int. Conf. Web-Age Inf. Manage.*, 2013, pp. 241–252.

[20] P. Zhou, L. Du, M. Fan, and Y.-D. Shen, "An LLE based heterogeneous metric learning for cross-media retrieval," in *Proc. SIAM Conf. Data Mining*, 2015, pp. 64–72.

[21] Y. Zhang, Q. Wang, D.-W. Gong, and X.-F. Song, "Nonnegative Laplacian embedding guided subspace learning for unsupervised feature selection," *Pattern Recognit.*, vol. 93, pp. 337–352, Sep. 2019.

[22] P. Zhou, J. Chen, M. Fan, L. Du, Y.-D. Shen, and X. Li, "Unsupervised feature selection for balanced clustering," *Knowl.-Based Syst.*, vol. 193, Apr. 2020, Art. no. 105417.

[23] P. Zhu, W. Zuo, L. Zhang, Q. Hu, and S. C. K. Shiu, "Unsupervised feature selection by regularized self-representation," *Pattern Recognit.*, vol. 48, no. 2, pp. 438–446, Feb. 2015.

[24] P. Zhu, W. Zhu, Q. Hu, C. Zhang, and W. Zuo, "Subspace clustering guided unsupervised feature selection," *Pattern Recognit.*, vol. 66, pp. 364–374, Jun. 2017.

[25] Y. Li, C. Lei, Y. Fang, R. Hu, Y. Li, and S. Zhang, "Unsupervised feature selection by combining subspace learning with feature self-representation," *Pattern Recognit. Lett.*, vol. 109, pp. 35–43, Jul. 2018.

[26] S. Wang, W. Pedrycz, Q. Zhu, and W. Zhu, "Subspace learning for unsupervised feature selection via matrix factorization," *Pattern Recognit.*, vol. 48, no. 1, pp. 10–19, Jan. 2015.

[27] S. Wang, J. Tang, and H. Liu, "Embedded unsupervised feature selection," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 470–476.

[28] X. Zhu, S. Zhang, R. Hu, Y. Zhu, and J. Song, "Local and global structure preservation for robust unsupervised spectral feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 3, pp. 517–529, Mar. 2018.

[29] P. Zhu, Q. Hu, C. Zhang, and W. Zuo, "Coupled dictionary learning for unsupervised feature selection," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2422–2428.

[30] N. Zhou, Y. Xu, H. Cheng, J. Fang, and W. Pedrycz, "Global and local structure preserving sparse subspace learning: An iterative approach to unsupervised feature selection," *Pattern Recognit.*, vol. 53, pp. 87–101, May 2016.

[31] N. Zhou, H. Cheng, W. Pedrycz, Y. Zhang, and H. Liu, "Discriminative sparse subspace learning and its application to unsupervised feature selection," *ISA Trans.*, vol. 61, pp. 104–118, Mar. 2016.

[32] W. Zhou, C. Wu, Y. Yi, and G. Luo, "Structure preserving non-negative feature self-representation for unsupervised feature selection," *IEEE Access*, vol. 5, pp. 8792–8803, 2017.

[33] C. Tang, X. Zhu, J. Chen, P. Wang, X. Liu, and J. Tian, "Robust graph regularized unsupervised feature selection," *Expert Syst. Appl.*, vol. 96, pp. 64–76, Apr. 2018.

[34] L. Du, Z. Shen, X. Li, P. Zhou, and Y.-D. Shen, "Local and global discriminative learning for unsupervised feature selection," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dec. 2013, pp. 131–140.

[35] L. Du and Y.-D. Shen, "Unsupervised feature selection with adaptive structure learning," in *Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2015, pp. 209–218.

[36] P. Zhou, L. Du, X. Li, Y.-D. Shen, and Y. Qian, "Unsupervised feature selection with adaptive multiple graph learning," *Pattern Recognit.*, vol. 105, Sep. 2020, Art. no. 107375.

[37] H. Liu, M. Shao, and Y. Fu, "Consensus guided unsupervised feature selection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 2016, pp. 1874–1880.

[38] L. Du, C. Ren, X. Lv, Y. Chen, P. Zhou, and Z. Hu, "Local graph reconstruction for parameter free unsupervised feature selection," *IEEE Access*, vol. 7, pp. 102921–102930, 2019.

[39] X. Zhang, M. Fan, D. Wang, P. Zhou, and D. Tao, "Top-k feature selection framework using robust 0-1 integer programming," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 31, 2020, doi: 10.1109/TNNLS.2020.3009209.

[40] H. Wang, L. Du, P. Zhou, L. Shi, and Y.-D. Shen, "Convex batch mode active sampling via α-relative pearson divergence," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3045–3051.

[41] K. Yu, J. Bi, and V. Tresp, "Active learning via transductive experimental design," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 1081–1088.
[42] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *Int. J. Comput. Vis.*, vol. 113, no. 2, pp. 113–127, Jun. 2015.
[43] W. Liu, X. Chang, L. Chen, D. Phung, X. Zhang, Y. Yang, and A. G. Hauptmann, "Pair-based uncertainty and diversity promoting early active learning for person re-identification," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 2, pp. 21:1–21:15, 2020.
[44] D. Cai and X. He, "Manifold adaptive experimental design for text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 4, pp. 707–719, Apr. 2012.
[45] Y. Feng, J. Xiao, Z. Zha, H. Zhang, and Y. Yang, "Active learning for social image retrieval using locally regressive optimal design," *Neurocomputing*, vol. 95, pp. 54–59, Oct. 2012.
[46] L. Shi and Y. Shen, "Diversifying convex transductive experimental design for active learning," in *Proc. IJCAI*, S. Kambhampati, Ed., 2016, pp. 1997–2003.
[47] Q. He, Z. Xie, Q. Hu, and C. Wu, "Neighborhood based sample and feature selection for SVM classification learning," *Neurocomputing*, vol. 74, no. 10, pp. 1585–1594, May 2011.
[48] N. García-Pedrajas, A. de Haro-García, and J. Pérez-Rodríguez, "A scalable memetic algorithm for simultaneous instance and feature selection," *Evol. Comput.*, vol. 22, no. 1, pp. 1–45, 2014.
[49] Y.-H. Shao, C.-N. Li, L.-W. Huang, Z. Wang, N.-Y. Deng, and Y. Xu, "Joint sample and feature selection via sparse primal and dual LSSVM," *Knowl.-Based Syst.*, vol. 185, Dec. 2019, Art. no. 104915.
[50] S. Rathee and J. Ahuja, "A proposal for dual data selection using parallel genetic algorithm," in *Proc. Decis. Anal. Appl. Ind.*, 2020, pp. 217–223.
[51] R. Makkhongkaew, K. Benabdeslem, and H. Elghazel, "Semi-supervised co-selection: Features and instances by a weighting approach," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 3477–3484.
[52] R. Makkhongkaew and K. Benabdeslem, "Semi-supervised similarity preserving co-selection," in *Proc. IEEE 16th Int. Conf. Data Mining Workshops*, Dec. 2016, pp. 756–761.
[53] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint ℓ_2, 1-norms minimization," in *Proc. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.
[54] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou, "$\ell_2$1-norm regularized discriminative feature selection for unsupervised learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1589–1594.
[55] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Proc. AAAI*, 2012, pp. 1026–1032.
[56] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

LIANG DU (Member, IEEE) received the B.E. degree in software engineering from Wuhan University, in 2007, and the Ph.D. degree in computer science from the Institute of Software, University of Chinese Academy of Sciences, in 2013. From July 2013 to July 2014, he was a Software Engineer with Alibaba Group. He was also an Assistant Researcher with the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences. He is currently a Lecturer with Shanxi University. He has published more than 40 papers in top conferences and journals, including KDD, IJCAI, AAAI, ICDM, TKDE, SDM, and CIKM. His research interests include clustering with noise and heterogeneous data, ranking for feature selection, active learning, and document summarization.

XIN REN is currently pursuing the M.S. degree with Shanxi University, Taiyuan, China. Her research interests include data mining and machine learning algorithms.

PENG ZHOU (Member, IEEE) received the B.E. degree in computer science and technology from the University of Science and Technology of China, in 2011, and the Ph.D. degree in computer science from the Institute of Software, University of Chinese Academy of Sciences, in 2017. He is currently a Lecturer with Anhui University. His research interests include machine learning, data mining, and artificial intelligence.

ZHIGUO HU received the B.S. degree in command and control engineering from the Artillery College, Shenyang, China, in 2001, the M.S. degree in command and control engineering from the Artillery College, Hefei, China, in 2006, and the Ph.D. degree in computer science from Tongji University, China, in 2012. In 2015, he joined Shanxi University, Taiyuan, where he is currently an Associate Professor with the School of Computer and Information Technology. His research interests include network measurement, data mining, and machine leaning.