Unsupervised Feature Selection with Correlation and Individuality Analysis

Xiucai Ye, Kaiyang Ji, and Tetsuya Sakurai

***Abstract*—Feature selection is an important technique for data dimension reduction. Embedded method with sparse regression is wildly used for unsupervised feature selection. The embedded method aims to find a better feature subset by exploiting feature correlation without considering the importance of each feature individually. In this paper, we propose a framework for unsupervised feature selection based on the embedded and spare regression model. Our framework not only exploits the correlation of the features but also analyzes the importance of each individual feature. By using the weight of individual feature to optimize the sparse regression in the process of embedding, the correlation and local structure preserving property of the selected features can be well balanced. We evaluate the proposed framework by using four public datasets. The experimental results demonstrate the superior performance of the proposed framework.**

***Index Terms*—Unsupervised feature selection, embedding method, sparse regression, correlativity, individuality, local preserving.**

I. Introduction

In many applications such as machine learning and data mining, the data samples are often represented by a large amount of features. However, not all the features are important. Usually many of the features are correlated or redundant to each other, and some of them are just noise. Feature selection is one of the dimensionality reduction techniques, which aims to extract the important features and eliminate the noisy ones. Feature selection, bringing the immediate effects for applications such as speeding up a data clustering algorithm and improving the accuracy of the predictive results [1], is proven to be an effective and efficient method to handle high dimensional data [2], [3].

Based on the way of utilizing label information, feature selection methods can be broadly classified into supervised and unsupervised methods. For supervised feature selection, the discriminative information is encoded in the labels. It is able to select features for distinguishing samples from different clusters [4]-[6]. Unsupervised feature selection is considered as a much harder problem, since the definition of relevance of features becomes unclear due to the lack of label information [3], [7], [8]. With the rapid accumulation of high dimensional data, there is usually no shortage of unlabeled data but labels are expensive [9]. Thus, it is of great importance to develop unsupervised feature selection.

Unsupervised feature selection attracts increasing attention in recent years and a large number of methods have been proposed [10]-[13]. The most existing unsupervised feature selection methods are the filter and embedded methods. In unsupervised filter methods, the features are selected one by one based on certain evaluation criteria without involving any learning process. The typical methods include the max variance (MaxVar) method, the Laplacian score (LapScore) method [14] and its extension, i.e., the spectral feature selection (SPEC) method [15]. LapScore is based on the local geometric structure of the data and selects the features with local preserving. However, LapScore only considers the local preserving property of individual feature. A common disadvantage of such unsupervised filter methods is that the correlation among features is neglected [16].

Unsupervised embedded methods are developed to perform feature selection with a learning model simultaneously. Usually, sparse regression is added as a constraint to learn the feature weights correlatively. A number of alternative criteria have been proposed for the learning processes in the unsupervised embedded methods such as data similarity [17], [18], data separability [19], and data discriminative [20]. The method in [17] (i.e., JELSR) uses the similarity via locally linear approximation to construct graph and unifies embedding learning and sparse regression to perform feature selection. The method in [19] (i.e., MCFS) selects the features that can best preserve the multi-cluster structure by manifold learning and ℓ_1 - regularization. The method in [20] (i.e., UDFS) incorporates discriminative analysis and $\ell_{2,1}$ - norm minimization into a joint framework for unsupervised feature selection. Compared with the unsupervised filter methods, the unsupervised embedded methods have been proved to perform better in many cases [21].

In this paper, we propose a novel framework for unsupervised feature selection. Our framework utilizes sparse regression with the embedded method to exploit the feature correlation. Meanwhile, we consider the importance (i.e., the weight) of each individual feature in the framework, which aims to balance the correlation and local preserving property of the selected features. To the best of our knowledge, we are the first to consider feature selection with correlation and individuality analysis simultaneously. In the framework, we use the weight of individual feature to optimize the sparse regression in the process of embedding. Our proposed framework is flexible and extendable, since there are a lot of individual feature measure methods and embedded methods with sparse regression can be incorporated. In this paper, we use the LapScore method [14] to calculate the weight of individual feature, and use the local linear embedding (LEE) method [22] as the embedded method in the framework. We also present an iterative algorithm to efficiently solve the optimization problem in our

Manuscript received October 30, 2015; revised January 17, 2016.

The authors are with the Computer Science, University of Tsukuba, Tsukuba, Japan (e-mail: yexiucai@mma.cs.tsukuba.ac.jp, jikaiyang@mma.cs.tsukuba.ac.jp, sakurai@cs.tsukuba.ac.jp).

framework. Many experimental results are provided for demonstration.

The rest of this paper is organized as follows. The relative methods are presented in Section II. We present the proposed framework and provide an efficient solution algorithm in Section III. Section IV shows some comparing results on the real world datasets. The conclusion is presented in Section V

II. Related Methods

A. Notation

In this paper, we use $x_1, \ldots, x_n$ to denote the *n* unlabeled data samples, $x_i \in R^m$ and $X = [x_1, \ldots, x_n]$ is the data matrix. We use $f_1, \ldots, f_m$ to denote the feature vectors of the data samples, and the data matrix is also denoted as $X = [f_1, \ldots, f_m]^T$. We want to select *d* features form $f_1, \ldots, f_m$ to represent the original data, where $d < m$. For a matrix $A \in R^{u \times v}$, its Frobenius norm is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^{u}\sum_{j=1}^{v} a_{ij}^2}. \quad (1)$$

And, the $\ell_{2,1}$- norm of *A* is defined as

$$\|A\|_{2,1} = \sum_{i=1}^{u}\sqrt{\sum_{j=1}^{v} a_{ij}^2}. \quad (2)$$

For the *n* data samples, the pairwise similarity among them can be represented as a symmetric matrix $S = (s_{ij})_{n \times n}$, where s_{ij} is the pairwise similarity between data x_i and x_j. We construct an undirected *k*-nearest neighbor (k-NN) graph with the *n* data samples. In the k-NN graph, data x_i is connected to x_j if x_i is among the *k* nearest neighbors of x_j or x_j is among the *k* nearest neighbors of x_i. According to the k-NN graph, the pairwise similarity is s_{ij} ($s_{ij} \neq 0$) if x_i is connected to x_j, otherwise, $s_{ij} = 0$.

B. Laplacian Score

Laplacian Score is proposed to select features that preserve sample locality specified by the similarity matrix *S*. The pairwise similarity is calculated as

$$s_{ij} = \begin{cases} \exp\left(-\|x_i - x_j\|^2/\sigma^2\right), & x_i \text{ is connected to } x_j, \\ 0, & \text{otherwises}, \end{cases} \quad (3)$$

where $\|x_i - x_j\|$ is the Euclidean distance between x_i and x_j, and σ is the kernel parameter. For the similarity matrix *S*, the corresponding degree matrix *D* is a $n \times n$ diagonal matrix with $d_i = \sum_{j=1}^{n} s_{ij}$ on the diagonal. The corresponding Laplacian matrix of *S* is $L = D - S$. The Laplacian Score of feature f_i is calculated as

$$L(f_i) = \frac{\tilde{f}_i^T L \tilde{f}_i}{\tilde{f}_i^T D \tilde{f}_i}, \quad (4)$$

where $\tilde{f}_i = f_i - \frac{f_i^T D \vec{1}}{\vec{1}^T D \vec{1}}\vec{1}$ and $\vec{1} = [1, \ldots, 1]^T$. In Laplacian Score, the features are evaluated independently. Feature selection is performed by selecting the top *d* features which have the minimal scores.

C. Embedded Methods with Sparse Regression

Different from the Laplacian Score, many methods evaluate the features by considering their correlation. In the embedded unsupervised feature selection methods, the original data x_i is embedded in a low dimensional space by a transformation matrix *W*. In the low dimensional space, data x_i is represented by $y_i \in R^s$, where *s* is the dimensionality of embedding. We use $Y = [y_1, \ldots, y_n]$ to denote the embedding data matrix of *X*. By embedding, the most valuable information of the original data is retained and the feature redundancies are eliminated. The problem becomes to solve the following optimization problem to obtain the transformation matrix *W* and the embedding data matrix *Y*.

$$\min_{W,\,Y} \|W^T X - Y\|_F^2 + \alpha\|W\|_{2,1} + \beta \mathrm{Loc}(Y), \quad (5)$$

where α and β are the balance parameters, and Loc(*Y*) is a promoting regularization term to satisfy that the structure of the original data is retained in the low dimensional embedding data. Note that the second term in (5) is the $\ell_{2,1}$-norm of the transformation matrix *W* to promote row sparsity, which has an effect of feature selection and help to avoid selecting redundant features. Denote w_i as the i^{th} row of the $m \times s$ matrix *W*, i.e., $W = [w_1, \ldots, w_m]^T$. After *W* is obtained, each feature f_i is ranked according to $\|w_i\|_2$ in descending order and the top rank *d* features are selected.

III. The Proposed Framework

A. Formulations

In order to balance the correlation and local preserving property of the selected features, we propose a novel unsupervised feature selection framework as

$$\min_{W,\,Y} \|W^T X - Y\|_F^2 + \alpha\|\mathrm{diag}(L(f_i))W\|_{2,1} + \beta \mathrm{Loc}(Y). \quad (6)$$

In (6), we use the Laplacian Score $L(f_i)$ as the penalty parameter for the $\ell_{2,1}$- norm of *W*. The weight of each row in *W* is adjusted by $L(f_i)$. In (5), when $\|w_i\|_2 = \|w_j\|_2$, the importance of features f_i and f_j cannot be distinguished. The case of $\|w_i\|_2 = \|w_j\|_2$ in (5) is just the case of $L(f_i)\|w_i\|_2 = L(f_j)\|w_j\|_2$ in (6). In (6), when $L(f_i)\|w_i\|_2 = L(f_j)\|w_j\|_2$, we can distinguish f_i and f_j according to $L(f_i)$ and $L(f_j)$. If $L(f_i) > L(f_j)$, we can obtain $\|w_i\|_2 < \|w_j\|_2$, that is, feature f_j is more important than feature f_i. Note that, $L(f_i) > L(f_j)$ denotes that f_j

has higher local preserving level than f_i. Thus, by adjusting the $\ell_{2,1}$- norm of W with Laplacian Score, we can select the features with less redundancy and higher local preserving property.

In this paper, we use the local linear embedding (LLE) [22] method to calculate Loc(Y) in (6). The pairwise similarity s_{ij} is calculated by the following locally linear approximation.

$$\min_{S,\ \sum_j s_{ij}=1} \sum_{i=1}^{n} \left\| x_i - \sum_{x_j \in N(x_i)} s_{ij} x_j \right\|_2^2, \tag{7}$$

where $N(x_i)$ is the set of neighbors of x_i in the k-NN graph. If x_i is not connected to x_j, $s_{ij} = 0$. Since the structure of the original data x_i should be retained in the low dimensional embedding data y_i, we calculate Loc(Y) as

$$\mathrm{Loc}(Y) = \min_{YY^T = I_{s\times s}} \sum_{i=1}^{n} \left\| y_i - \sum_{j=1}^{n} s_{ij} y_j \right\|_2^2. \tag{8}$$

Let $M = (I_{n\times n} - S)^T (I_{n\times n} - S)$, (8) is transformed as

$$\mathrm{Loc}(Y) = tr(YMY^T). \tag{9}$$

According to (9), the objective function of our proposed framework in (6) is formulated as

$$\begin{aligned} \Phi(W,Y) = \min_{W,\ YY^T = I_{s\times s}} & \left\| W^T X - Y \right\|_F^2 + \alpha \left\| \mathrm{diag}(L(f_i))W \right\|_{2,1} \\ & + \beta tr(YMY^T) \end{aligned}. \tag{10}$$

After obtaining W, we select the features by ranking each feature f_i according to $\|w_i\|_2$ in descending order and selecting the top rank d features.

B. Solutions

The optimization problem in (10) is not convex when both W and Y are optimized simultaneously. Moreover, the $\ell_{2,1}$- norm of W makes the problem non-smooth. Inspired by [17] and [23], we solve this problem in an alternative way. For convenience, we denote $\Psi(W) = \left\| \mathrm{diag}(L(f_i))W \right\|_{2,1}$.

The derivative of $\Psi(W)$ with respect to W is

$$\frac{\partial \Psi(W)}{\partial W} = 2U\,\mathrm{diag}(L(f_i))W, \tag{11}$$

where $U \in R^{m\times m}$ is a diagonal matrix with the i^{th} diagonal element as

$$U_{ii} = \frac{1}{2\|w_i\|_2}. \tag{12}$$

When U is fixed, the derivative in (10) can also be regarded as the derivative of (13). Thus, we try to solve the problem in (13) to approximate the solution to (10).

$$\begin{aligned} \Phi(W,U,Y) = \min_{W,\ YY^T = I_{s\times s}} & \left\| W^T X - Y \right\|_F^2 \\ & + \alpha tr(W^T U\,\mathrm{diag}(L(f_i))W) + \beta tr(YMY^T) \end{aligned}. \tag{13}$$

We take the derivative of $\Phi(W)$ with respect to W as the following equation.

$$\begin{aligned} & \frac{\partial \Phi(W,U,Y)}{\partial W} \\ & = 2XX^T W - 2XY^T + 2\alpha U\,\mathrm{diag}(L(f_i))W \end{aligned}. \tag{14}$$

By setting (15) to equal to 0, we can obtain

$$W = (XX^T + \alpha U\,\mathrm{diag}(L(f_i)))^{-1} XY^T. \tag{15}$$

Note that (13) can be transformed as

$$\begin{aligned} \Phi(W,U,Y) = & \min_{W,\ YY^T = I_{s\times s}} tr(W^T XX^T W) - 2tr(W^T XY^T) + tr(YY^T) \\ & + \alpha tr(W^T U\,\mathrm{diag}(L(f_i))W) + \beta tr(YMY^T) \end{aligned} \tag{16}$$

Then, by substituting (15) into (16), we have

$$\begin{aligned} \Phi(W,U,Y) = & \min_{W,\ YY^T = I_{s\times s}} tr(YY^T) + \beta tr(YMY^T) \\ & - tr(W^T (XX^T + \alpha U\,\mathrm{diag}(L(f_i)))W) \end{aligned}. \tag{17}$$

Let $B = XX^T + \alpha U\,\mathrm{diag}(L(f_i))$, according to (15), (17) becomes

$$\begin{aligned} & \Phi(W,U,Y) \\ & = \min_{YY^T = I_{s\times s}} tr(YY^T) + \beta tr(YMY^T) - tr(YX^T B^{-1} XY^T) \\ & = \min_{YY^T = I_{s\times s}} tr(Y(\beta M + I_{n\times n} - X^T B^{-1} X)Y^T) \end{aligned}. \tag{18}$$

Since M and B are fixed, the solution of (18) can be obtained by solving the following eigenvalue problem.

$$(\beta M + I_{n\times n} - X^T B^{-1} X) y_i = \lambda y_i. \tag{19}$$

The matrix Y, containing the eigenvectors corresponding to the s smallest eigenvalues as the row vectors, is the solution of (18).

In summary, we solve the optimization problem in (10) in an alternative way. When W is fixed, U can be updated according to (12). When U is fixed, Y can be updated according to (19). Then, (15) can be used to update W. After that, U can be updated again according to W as defined in (12). Similar to the results in [17] and [23], the objective function in (10) is convergence. We do not show the convergence analysis in this paper due to the limited space. The updating process will be done until the objective function in (10) converges.

We summarize the procedure of the proposed framework by using LLE to calculate Loc(Y) in Algorithm 1.

Algorithm 1 The proposed framework using LLE to calculate Loc(Y)

Input: Data matrix X; Balance parameter α and β; Neighborhood size k; Dimensionality of embedding s; Selected feature numbers d.
Output: d selected features.
1: Construct the k-NN graph;
2: Compute the Laplacian Score $L(f_i)$ for each feature by using (4);
3: Compute the similarity matrix S according to (7) and compute M as $M = (I_{n\times n} - S)^T (I_{n\times n} - S)$;
4: Initialize $U_0 = I_{m\times m}$ and set $t = 0$;
5: Repeat
6: Compute Y_t by solving the eigenvalue problem in (19);
7: Compute W_t by using (15);
8: Compute U_{t+1} by using (12);
9: $t = t+1$;
10: Until convergence;
11: Sort each feature f_i according to $\|w_i\|_2$ in descending order and select the top d ranked ones.

IV. Experiments

In this section, we test the performance of the proposed framework for unsupervised feature selection. Similar to that considered in [14] and [19], we test the performance in terms of clustering. After selecting the features, clustering is performed by using only the selected features.

A. Experiment Setup

In our experiment, we use a diversity of four public datasets to compare the performance of different unsupervised feature selection methods. The datasets include one object dataset, i.e., COIL20 [19], one face image dataset, i.e., UMIST [17], one handwritten digit dataset, i.e., USPS [23], and one spoken letter recognition data, i.e., Isolet1 [24]. The properties of the four datasets are summarized in Table I.

We compare the proposed framework with other existing unsupervised feature selection methods. Since our framework selects features by considering individuality and correlation, we use ICFS to denote the method in Algorithm 1 in the rest of this paper. To show that our framework is also efficient when the sparse constrain of W is in ℓ_1- norm, we also use Laplacian Score to optimize the embedded process of the MCFS method [19] in the experiment.

TABLE I: Properties of the Datasets

Dataset	Size	# of features	# of classes
COIL20	1440	1024	20
UMIST	575	400	20
USPS	9298	256	10
Isolet	1560	617	26

Including the proposed ICFS method, the compared unsupervised feature selection methods are summarized as follow.

1) LapScore [14], which selects the features that have high level of locality preserving ability.
2) MCFS [19], which selects the features by using spectral regression with ℓ_1- norm regularization.
3) JELSR [17], which performs feature selection by jointing embedding learning with sparse regression.
4) IC-MCFS, which uses the Laplacian Score $L(f_i)$ as the penalty parameter for the ℓ_1- norm of W in [19].

There are some parameters that should be set in advance. We set the number of nearest neighbors as k = 5 for all the compared methods. To fairly compare different unsupervised feature selection method, we tune the parameters from $\{10^{-6},10^{-4}, ..., 10^{6}\}$. The number of selected features is ranged from 20 to 200. We report the best result of all the methods by using different parameters. Each feature selection algorithm is first performed to select features. Then K-means is performed based on the selected features. We repeat K-means 20 times with random initializations and report the average results.

We apply two widely used evaluation metrics, i.e., Accuracy (ACC) and Normalized Mutual Information (NMI), to evaluate the clustering results. Given a clustering result C, c clusters and the ground truth label G, g clusters. Denote c_i as the index of the clustering result of x_i and g_i as the ground truth label of x_i. ACC is defined as

$$ACC = \frac{\sum_{i=1}^{n} \delta(g_i, map(c_i))}{n}, \quad (20)$$

where $\delta(a,b)=1$ if $a = b$ and $\delta(a,b)=0$ otherwise, $map(c_i)$ is the best mapping function that permutes clustering labels to match the ground truth labels using the Kuhn-Munkres algorithm. A larger value of ACC denotes a better clustering result. Denote n_l as the number of data in the cluster C_l ($1 \le l \le c$). Denote $n_h^{'}$ as the number of data in the ground truth cluster G_h ($1 \le h \le g$). Let $n_{l,h}$ denote the number of data that are in the intersection between the clusters C_l and G_h. According to [25], NMI is defined as

$$NMI = \frac{\sum_{l=1}^{c}\sum_{h=1}^{g} n_{l,h}\log(n \cdot n_{l,h}/(n_l \cdot n_h^{'}))}{\sqrt{(\sum_{l=1}^{c} n_l \log(n_l/n))(\sum_{h=1}^{g} n_h^{'}\log(n_h^{'}/n))}}. \quad (21)$$

A larger value of NMI denotes better performance. The largest value of NMI is 1, which occurs when all of the data points are assigned to their correct clusters.

B. Experimental Results

TABLE II: Clustering Results (NMI% ± std) of Different Feature Selection Methods

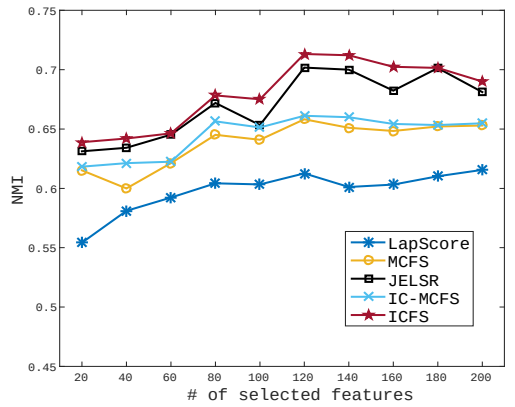
Dataset	LapScore	MCFS	JELSR	IC-MCFS	ICFS
COIL20	61.56 ±4.21	65.82 ±5.86	70.18 ±5.23	66.12 ±5.72	**71.22** ±5.68
UMIST	56.76 ±3.28	64.83 ±4.32	70.98 ±3.14	65.14 ±4.29	**71.33** ±3.20
USPS	62.57 ±3.21	64.19 ±5.12	64.87 ±4.78	64.42 ±5.14	**64.95** ±4.80
Isolet	71.38 ±2.01	74.42 ±1.94	75.10 ±2.35	74.83 ±2.36	**75.84** ±2.42

TABLE III: Clustering Results (ACC% ± std) of Different Feature Selection Methods

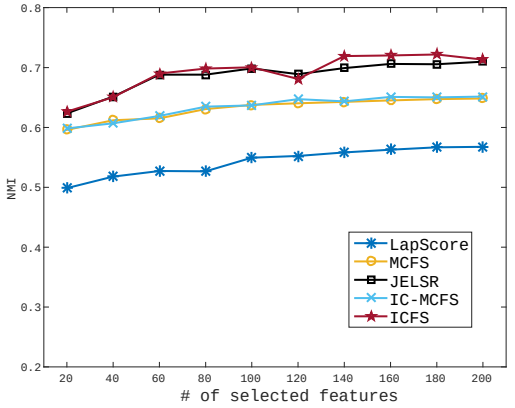
Dataset	LapScore	MCFS	JELSR	IC-MCFS	ICFS
COIL20	45.78 ±6.26	50.13 ±5.22	56.67 ±4.28	50.85 ±5.31	**57.23** ±4.12
UMIST	40.17 ±2.15	43.47 ±3.39	52.43 ±2.15	43.86 ±3.42	**53.14** ±2.35
USPS	60.47 ±2.57	61.01 ±1.98	61.27 ±2.03	61.08 ±2.03	**61.92** ±2.26
Isolet	56.46 ±3.18	60.83 ±4.60	61.52 ±4.25	61.16 ±4.26	**62.02** ±4.50

We first compare the performance of different unsupervised feature selection methods. The experimental results in terms of NMI and ACC evaluation metrics are shown in Table II and Table III, respectively. We can see from the two tables that the proposed ICFS method performs better than the other methods. By the proposed framework, the selected features are with less redundancy and higher local preserving property. Note that JELSR also applies the LLE [22] method. The main difference between JELSR and ICFS is that ICFS has the Laplacian Score $L(f_i)$ as the penalty parameter for the $\ell_{2,1}$- norm of W while JELSR has not. We compare JELSR with ICFS to show the efficiency of the proposed framework. To show that our framework is also

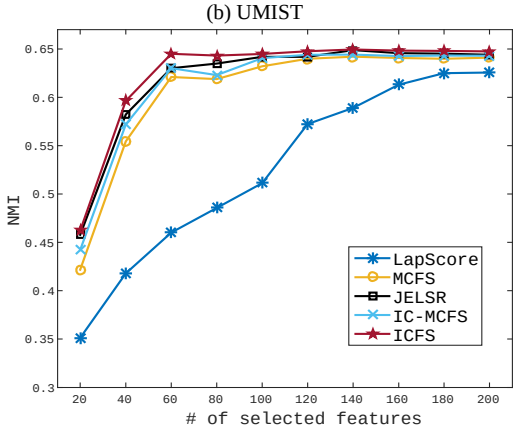
efficient when the sparse constrain of W is in ℓ_1- norm, we also compare ICFS with IC-MCFS. From Table II and Table III we can see that the IC-MCFS method performs better than the MCFS method. The performance of MCFS is improved by applying the proposed framework.



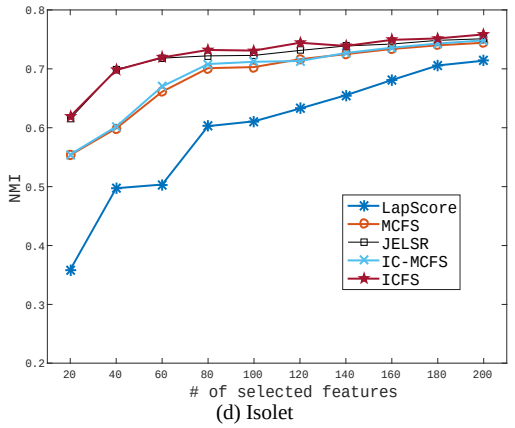
(a) COIL20



(b) UMIST



(c) USPS



(d) Isolet

Fig. 1. NMI of the clustering results on four datasets by varying the number of selected features.

Fig. 1 shows the performance of the clustering results on the four datasets by varying the number of selected features. We can see from Fig. 1 that the proposed ICFS method performs better than other methods in most cases. Also, the IC-MCFS method performs better than the MCFS method. In Fig. 1, we show the performance of the clustering results by NMI evaluation metric. We do not show the performance of clustering results by ACC evaluation metric, since the trends of the performance by using the two evaluation metrics are very similar. The proposed ICFS method also performs better than other methods when ACC evaluation metric is applied.

V. Conclusion

In this paper, we propose a novel framework for unsupervised feature selection to balance the correlation and local preserving property of the selected features. Based on the framework, we use LapScore to calculate the weight of the individual feature, and use the LEE method as the embedded method, by which the selected features are with less redundancy and higher local preserving property. Many experimental results are provided to demonstrate the superior performance of the proposed framework.

References

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, March 2003.

[2] G. John, R. Kohavi, and K. Pfleger, "Irrelevant feature and the subset selection problem," *International Proceedingsof Machine Learning*, pp. 121-129, 1994.

[3] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *Journal of Machine Learning Research*, vol. 5, pp. 845–889, Dec 2004.

[4] R. Duda, P. Hart, D. Stork *et al*., *Pattern Classification*, Wiley New York, 2012, ch. 1, pp. 11-12.

[5] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389-422, January 2002.

[6] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-elevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226-1238, June 2005.

[7] J. G. Dy, C. E. Brodley, A. C. Kak, L. S. Broderick, and A. M. Aisen, "Unsupervised feature selection applied to content- based retrieval of lung images," *IEEE Transactions on Pattern Analysis and Machine Intelligenc*, vol. 25, pp. 373-378, March, 2003.

[8] J. G. Dy and C. E. Brodley, "Visualization and interactive feature selection for unsupervised data," in *Proc. the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 360-364.

[9] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu, "Clustering-guided sparse structural learning for unsupervised feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, pp. 2138-2150, April 2014.

[10] L. Wolf and A. Shashua, "Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weighted-based approach," *Journal of Machine Learning Research*, vol. 6, pp. 1855-1887, 2005.

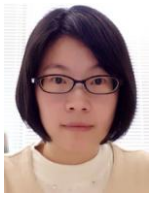
[11] Lei. Shi, L. Du, and Y. Shen, "Robust spectral learning for unsupervised feature selection," in *Proc. IEEE International Conference on Data Mining*, 2014, pp. 977-982.

[12] C. Constantinopoulos, M. Titsias, and A. Likas, "Bayesian feature and model selection for gaussian mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligenc*, vol. 28, pp. 1013-1018, June, 2006.

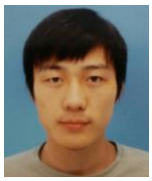
[13] P. Zhu, W. Zuo, L. Zhang, Q. Hu, and S. C. Shiu, "Unsupervised feature selection by regularized self-representation," *Pattern Recognition*, vol. 48, pp. 438–446, 2015.

[14] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," *Advances in Neural Information Processing Systems*, pp. 507-514, 2006.

[15] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. ACM International Conference on Machine Learning*, 2007, pp. 1151–1157.
[16] S. Alelyani, J. Tang, and H. Liu, "Feature selection for clustering: A review," *Data Clustering: Algorithms and Applications*, ch. 2, pp. 29-36, 2013.
[17] C. Hou, F. Nie, D. Yi, and Y. Wu, "Feature selection via joint embedding learning and sparse regression," in *Proc. International Joint Conference on Artificial Intelligence*, 2011, pp. 1324–1329.
[18] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, pp. 619-632, 2013.
[19] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 333-342.
[20] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou, "L2, 1-norm regularized discriminative feature selection for unsupervised learning," in *Proc. International Joint Conference on Artificial Intelligence*, 2011, pp. 1589–1594.
[21] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *Proc. the Twenty-4th AAAI Conference on Artificial Intelligence*, 2010, pp. 1-6.
[22] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding, " *Science*, pp. 2323–2326, 2000.
[23] J. J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 550–554, 1994.
[24] M. A. Fanty and R. Cole, "Spoken letter recognition," in *Proc. the Third DARPA Speech and Natural Language Workshop*, 1990, pp. 385-390.
[25] A. Strehl and J. Ghosh, "Cluster ensembles–a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2002.



Xiucai Ye received her PhD in computer science from the University of Tsukuba, Tsukuba Science City, Japan, in 2014. She is currently working as a postdoctoral research fellow at the Department of Computer Science, University of Tsukuba, Tsukuba Science City, Japan. Her current research interests include clustering, feature selection, machine learning and its application fields.



Kaiyang Ji is currently pursuing his master degree at the Department of Computer Science, University of Tsukuba, Tsukuba Science City, Japan. His current research interests include feature selection, clustering, and intelligent algorithms.

Tetsuya Sakurai is a professor of computer science, dean of the College of Information Science, and Director of the Office of Strategies of R&D at University of Tsukuba. He received the Ph.D. degree in computer engineering from Nagoya University, Japan, in 1992. His research interests include parallel algorithms for linear algebraic computations, spectral methods for data and image analysis, and mathematical software. He is currently the director of the Japan Society for Industrial and Applied Mathematics (JSIAM) and the chief editor of the online journal JSIAM Letters.