Unsupervised Feature Selection via Metric Fusion and Novel Low-Rank Approximation

YIN LONG1, LIANG CHEN1, LINFENG LI1, AND RONG SHI2

1School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China
2School of Applied Technology, Southwest University of Science and Technology, Mianyang 621010, China

Corresponding author: Yin Long (yinnuoloong@163.com)

This work was supported by the National Science Foundation of China under Grant 62101467.

ABSTRACT Unsupervised feature selection aims to derive a compact set of features with desired generalization ability via removing the irrelevant and redundant features, yet challenging due to the unavailability of labels. Works about unsupervised feature selection always need to construct the similarity matrix, which makes the selected features highly depend on the accuracy of similarity measurement. However, existing works usually leverage a single fixed metric to build similarity matrix, which cannot fit various feature types very well and even damage the local manifold structure. To address this problem, we propose an adaptive multi-metric fusion by automatically integrating similarity across different metrics according to the specific data. Besides, to capture the global structure more precisely, a novel low-rank approximation method is proposed, which is relatively insensitive to the rank-norm parameter. Via the proposed novel low-rank approximation method, better tradeoff between the performance and robustness can be provided. Experimental results show that the accuracy performance of the proposed method can be boosted by 2% − 11%, compared with previous methods.

INDEX TERMS Unsupervised feature selection, multi-metric fusion, similarity matrix, low-rank approximation, ADMM.

I. INTRODUCTION Currently, a large amount of unlabeled high-dimensional data has been generated in many fields, such as image data, biology data, and health data. However, high-dimensional data usually contains redundant, uncorrelated, and noisy features which bring high computational costs and dimensional disasters in the processing and analysis. Therefore, the approach of removing irrelevant or redundant features from high-dimensional data was proposed to address this problem, which is called feature selection. Generally, feature selection methods are of two classifications: supervised feature selection and Unsupervised Feature Selection (UFS), according to the availability of labels in the data samples. Among them, UFS is widely used in the realistic scenarios, because the data in the realistic tasks usually is unlabelled.

Moreover, UFS methods are of three types, filters [1], wrappers [2], or embedding [3]. Among them, the performance of filter and wrapper approaches are affected by the search strategy, while the embedding-based method performs the feature selection by learning. Thus, the embedding-based method has a better performance than other methods, thereby attracting more attention. The recent work [4] about embedding-based method tries to implement the feature selection by discovering the global structure of data space. Furthermore, based on the work [4], an improved work [5] is proposed, where the rank constraint is imposed on the self-representation matrix, and better performance is achieved. Besides, since local manifold structure is superior to global structure in some aspects, some recent embedding-based methods turn to leverage the local manifold structure to select features. For example, Structured Optimal Graph Feature Selection (SOGFS) [6] implements the feature selection by using locality preserving projections (LPP) [7]. Feature Selection and Kernel Learning for Local Learning-Based Clustering (LLCFS) [8] proposes a novel feature selection method and a kernel learning method for the local structure learning based clustering. The Dependence
Guided Unsupervised Feature Selection (DGUFS) [9] proposes a projection-free feature selection model based on \(l_2,0\)-norm equality constraints. Nevertheless, in the works mentioned above, only the local manifold structure or global structure is leveraged to reduce dimensionality, thus the local manifold structure and global structure cannot be preserved simultaneously. To overcome the drawback, work [10] proposes to select features by preserving local and global structure both, and better performance is obtained. In [11], to decrease redundant features, the dependence among features is calculated and is applied on the unsupervised feature selection model. In [12], the novel feature selection with the \(l_{2,1}\)-norm regularization as well as simultaneously learns orthogonal representations for each sample to preserve the local structure is proposed. In [13], a virtual label matrix is proposed to boost the selection of features, which is able to guarantee the selected features be more discriminated.

However, the above existing works tend to leverage single fixed metric to build the similarity matrix, which cannot fit various feature types very well and even damage the local manifold structure [6]. Besides, in the above existing works, the global structure is discovered by biased rank approximation, which leads to the performance loss. To address these problems, we propose an adaptive multi-metric fusion by automatically integrating similarity across different metrics according to the specific data, where the global structure is captured by a novel low-rank approximation. Via the proposed novel low-rank approximation, better tradeoff between the performance and robustness can be provided. Experimental results on real datasets validate the effectiveness of the proposed method.

II. RELATED WORKS

A. LOW-RANK SELF-REPRESENTATION

Self-representation aims to represent each data as a linear combination of other data in the subspace, where the low-rank hypothesis is used to discover a Low-Rank Representation (LRR) of high-dimensional data, thus the global structure can be obtained. Specifically, LRR problem can be formulated as

\[
\begin{align*}
\min_{Z} & \quad \text{rank} (Z) \\
\text{s.t.} & \quad P = PZ,
\end{align*}
\]

where \(P = [p_1, p_2, \ldots, p_n] \in \mathbb{R}^{d \times n}\) is the original data matrix, \(d\) is the number of features, \(n\) is the number of instances, \(Z \in \mathbb{R}^{n \times d}\) is the low-rank representation matrix, \(\text{rank} (Z)\) is the rank function. Besides, for the sake of description convenience, we denote the data after self-representation by \(X\) where \(X = PZ\).

B. LOCALITY PRESERVING PROJECTIONS

LPP is widely used in subspace learning. Compared with LRR, LPP can better capture the local manifold structure of data. Specifically, we first define a similarity matrix \(S\) whose element \(S_{ij}\) denotes the similarity between the \(i\)-th sample and \(j\)-th sample. Based on the common sense, closer samples are likely to have larger similarity, and the similarity can be evaluated by various metrics. For example, in [6], similarity between the \(i\)-th sample and \(j\)-th sample is measured by the inverse of the square of Euclidean distance between the \(i\)-th sample and \(j\)-th sample, i.e.,

\[
S_{ij} = 1/\|x_i - x_j\|_2^2.
\]

In [8], similarity between the \(i\)-th sample and \(j\)-th sample is measured by the inverse of \(\tau\)-weighted squared euclidean distance between the \(i\)-th sample and \(j\)-th sample, i.e.,

\[
S_{ij} = 1/\|\phi(x_i) - \phi(x_j)\|_\tau^2
\]

where \(\phi()\) is the kernel function. In [9], KNN approach is adopted in the construction of similarity matrix, where

\[
S_{ij} = \begin{cases} 
1 & j \in N_i \text{ or } i \in N_i \\
0 & \text{otherwise}
\end{cases}
\]

where \(N_i\) is the neighbor set of sample \(i\). In [10], the similarity matrix is constructed by the heat kernel, i.e.,

\[
S_{ij} = \begin{cases} 
\exp \left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right) & j \in N_i \text{or } i \in N_i \\
0 & \text{otherwise}
\end{cases}
\]

Furthermore, LPP problem can be described as

\[
\begin{align*}
\min_{W} & \quad \sum_{i,j} \left\|W^T x_i - W^T x_j\right\|_2^2 S_{ij} \\
\text{s.t.} & \quad W^T W = I_k,
\end{align*}
\]

where \(W \in \mathbb{R}^{d \times k}\) is the projection matrix with \(k\) being the dimension of the subspace, and the constraint \(W^T W = I_k\) is to avoid trivial solution [14]. As shown in Equation (6), the similarity between \(x_i\) and \(x_j\) in the original space can be preserved in the subspace. Furthermore, to make the rows of \(W\) sparse, the sparse constraint \(l_{2,1}\)-norm is imposed, then the objective function (6) can be rewritten as

\[
\begin{align*}
\min_{W, Z} & \quad \sum_{i,j} \left\|W^T x_i - W^T x_j\right\|_2^2 S_{ij} + \beta \|W\|_{2,1} \\
\text{s.t.} & \quad W^T W = I_k,
\end{align*}
\]

where \(\beta\) is the balance parameter for the sparse constraint.

C. UFS BASED ON GLOBAL AND LOCAL MANIFOLD STRUCTURE

As shown in (7), when \(k < d\), \(W\) can be used for selecting \(k\) features. Thus, the problem of performing feature selection based on global and local manifold structure can be written as

\[
\begin{align*}
\min_{W, Z} & \quad P - PZ\|_F + \omega \sum_{i,j} \left\|W^T x_i - W^T x_j\right\|_2^2 S_{ij} \\
& \quad + \alpha \text{rank} (Z) + \beta \|W\|_{2,1} \\
\text{s.t.} & \quad W^T W = I_k, \text{ diag} (Z) = 0,
\end{align*}
\]

where \(\alpha\) and \(\omega\) are the balance parameters for the low-rank constraint and LPP, respectively, \(\text{diag} (Z) = 0\) is used to avoid trivial solution.
II. THE PROPOSED APPROACH

To improve the performance of UFS, we propose a novel low-rank approximation to capture the global structure more accurately and propose a scheme of adaptive multi-metric fusion to fit a specific data. The specific details are discussed in the following.

A. NOVEL LOW-RANK APPROXIMATION

| TABLE 1. Summary of some previous rank norm. |
|---------------------------------------------|
| Rank norm                      | Advantage             | Disadvantage           |
| Nuclear norm                   | Low complexity        | Large bias             |
| Schatten norm                  | Flexible              | Sensitive to norm parameter p |
| Gamma norm                     | Small bias            | Sensitive to rank-norm parameter γ |

As shown in Equation (8), the optimization problem of UFS involves a rank minimization problem, which is known to be NP-hard. For the NP-hard problem, the nuclear norm is widely leveraged as a convex surrogate of the rank function. However, the nuclear norm and its enhanced methods are the low-rank approximation to capture the global structure more accurately enough. This enables us to use non-convex relaxations as much as possible, such as the Gamma norm [15]. The Gamma norm is an extension of the min-max concave plus function, which provides the unbiased estimation for the nuclear norm and its enhanced methods cannot approximate the rank function accurately enough. This enables us to use non-convex relaxations as much as possible, such as the Gamma norm [15].

The Gamma norm is an extension of the min-max concave plus function, which provides the unbiased estimation for the nuclear norm and its enhanced methods cannot approximate the rank function accurately enough. This enables us to use non-convex relaxations as much as possible, such as the Gamma norm [15].

As shown in Equation (8), the optimization problem of UFS involves a rank minimization problem, which is known to be NP-hard. For the NP-hard problem, the nuclear norm is widely leveraged as a convex surrogate of the rank function. However, the nuclear norm and its enhanced methods are the low-rank approximation to capture the global structure more accurately enough. This enables us to use non-convex relaxations as much as possible, such as the Gamma norm [15].

The Gamma norm is an extension of the min-max concave plus function, which provides the unbiased estimation for the nuclear norm and its enhanced methods cannot approximate the rank function accurately enough. This enables us to use non-convex relaxations as much as possible, such as the Gamma norm [15].

As shown in Equation (8), the optimization problem of UFS involves a rank minimization problem, which is known to be NP-hard. For the NP-hard problem, the nuclear norm is widely leveraged as a convex surrogate of the rank function. However, the nuclear norm and its enhanced methods are the low-rank approximation to capture the global structure more accurately enough. This enables us to use non-convex relaxations as much as possible, such as the Gamma norm [15].

The Gamma norm is an extension of the min-max concave plus function, which provides the unbiased estimation for the nuclear norm and its enhanced methods cannot approximate the rank function accurately enough. This enables us to use non-convex relaxations as much as possible, such as the Gamma norm [15].

As shown in Equation (8), the optimization problem of UFS involves a rank minimization problem, which is known to be NP-hard. For the NP-hard problem, the nuclear norm is widely leveraged as a convex surrogate of the rank function. However, the nuclear norm and its enhanced methods are the low-rank approximation to capture the global structure more accurately enough. This enables us to use non-convex relaxations as much as possible, such as the Gamma norm [15].

The Gamma norm is an extension of the min-max concave plus function, which provides the unbiased estimation for the nuclear norm and its enhanced methods cannot approximate the rank function accurately enough. This enables us to use non-convex relaxations as much as possible, such as the Gamma norm [15].

As shown in Equation (8), the optimization problem of UFS involves a rank minimization problem, which is known to be NP-hard. For the NP-hard problem, the nuclear norm is widely leveraged as a convex surrogate of the rank function. However, the nuclear norm and its enhanced methods are the low-rank approximation to capture the global structure more accurately enough. This enables us to use non-convex relaxations as much as possible, such as the Gamma norm [15].

The Gamma norm is an extension of the min-max concave plus function, which provides the unbiased estimation for the nuclear norm and its enhanced methods cannot approximate the rank function accurately enough. This enables us to use non-convex relaxations as much as possible, such as the Gamma norm [15].
we leverage the weighting scheme by
\[
h_t = 1/(2 \|m_t(X) - S_t\|_F^2).
\] (16)
In this way, the different similarity metric can be fused adaptively to fit a specific data. Therefore, by considering the adaptive multi-metric fusion, the optimization problem (14) can be reformulated as
\[
\min_{W, Z, F, S} \sum_{t=1}^{q} h_t \|m_t(X) - S_t\|_F^2
\]
+ \omega \sum_{i,j} \left\| W^T x_i - W^T x_j \right\|_2^2 S_{ij} + \theta \|P - PZ\|_F^2
+ \alpha \varphi_e(Z) + \beta \|W\|_{2,1} + \lambda Tr(F^T L_x F)
\]
\text{s.t. } \text{diag}(Z) = 0, S^T 1 = 1, S \succeq 0, F^T F = I_c, (17)
where \(\min Tr(F^T L_x F)\) guarantees the similarity matrix \(S\) containing \(c\) components [6], \(L_x\) is the Laplace matrix of \(S\), i.e., \(L_x = D - (S + S^T)/2\) with degree matrix \(D\) whose diagonal element \(D_{ii} = \sum_{j \neq i} S_{ij}\), \(\lambda\) and \(\theta\) represent the balance parameters.

**IV. OPTIMIZATION ALGORITHM**

By resorting to alternating direction method of multipliers (ADMM), the optimization problem (17) can be divided into several subproblems.

**A. UPDATE Z**

Firstly, one may observe that, \(m_t(X)\) in (17) involves the optimization variable \(Z\). However, \(m_t\) is an element-wise matrix function whose optimization is still an unsolved challenge in the optimization field. Hence, to solve the problem (17), similar to [16], we first decouple \(Z\) and \(X\) in \(m_t(X)\) by considering \(X\) in \(m_t(X)\) as a constant denoted by \(X^*\). Thus, with fixed \(W, F, S\), problem (17) reduces to
\[
\min_Z \theta \|P - PZ\|_F^2 + \alpha \varphi_e(Z)
+ \omega \sum_{i,j} \left\| W^T x_i - W^T x_j \right\|_2^2 S_{ij}
\]
\text{s.t. } \text{diag}(Z) = 0. (18)
By taking the self-representation into it, problem (18) can be further written as
\[
\min_Z \theta \|P - PZ\|_F^2 + \alpha \varphi_e(Z)
+ \omega Tr \left( W^T Z^T P^T L_c P Z W \right)
\]
\text{s.t. } \text{diag}(Z) = 0. (19)
Then, by introducing the auxiliary variables \(A, Z_1\) and \(Z_2\), the optimization problem can be rewritten as
\[
\min_{A, Z_1, Z_2} \theta \|P - PA\|_F^2 + \alpha \varphi_e(Z_1)
+ \omega Tr \left( W^T Z_1^T L'_c Z_2 W \right)
\]
\text{s.t. } A = Z_2 - \text{diag}(Z_2), A = Z_1, (20)
where \(L'_c = P^T L_c P\). Then, the augmented Lagrange of Equation (20) is
\[
\mathcal{L}_1(A, Z_1, Z_2, A_1, A_2)
= \theta \|P - PA\|_F^2 + \alpha \varphi_e(Z_1) + \omega Tr \left( W^T Z_1^T L'_c Z_2 W \right)
+ \omega Tr \left( A_1^T (A - Z_2 + \text{diag}(Z_2)) \right) + \omega Tr \left( A_2^T (A - Z_1) \right)
+ \frac{\mu}{2} \|A - Z_2 + \text{diag}(Z_2)\|_F^2 + \frac{\mu}{2} \|A - Z_1\|_F^2. (21)
\]
where \(A_1\) and \(A_2\) are two Lagrange multipliers, \(\mu\) is a positive penalty parameter.

\[
\varphi_e(Z) = \sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_i^2 + \epsilon} = Tr \left( \begin{pmatrix} \text{diag} \left( \sigma^2(Z) \right) + \epsilon I_r & 0 \\ 0 & \epsilon I_{n-r} \end{pmatrix} \right)^{-1} \begin{pmatrix} \text{diag} \left( \sigma^2(Z) \right) \\ 0 \end{pmatrix}
= Tr \left( U \Sigma \begin{pmatrix} \text{diag} \left( \sigma^2(Z) \right) + \epsilon I_r & 0 \\ 0 & \epsilon I_{n-r} \end{pmatrix} \right)^{-1} \Sigma^T U^T
= Tr \left( U \Sigma V^T \begin{pmatrix} \text{diag} \left( \sigma^2(Z) \right) + \epsilon I_r & 0 \\ 0 & \epsilon I_{n-r} \end{pmatrix} \right)^{-1} V^T \Sigma V^T U^T
= Tr \left( Z V \begin{pmatrix} \text{diag} \left( \sigma^2(Z) \right) + \epsilon I_r & 0 \\ 0 & \epsilon I_{n-r} \end{pmatrix} \right)^{-1} V^T Z^T
= Tr \left( Z^T \Sigma V^T + \epsilon I_n \right)^{-1} Z^T
= Tr \left( Z^T Z + \epsilon I_n \right)^{-1} Z^T. (9)
\]
1) UPDATE A

By setting the partial derivative about $A$ on Equation (21) as zero, $A$ can be updated as

$$A = \left(2\theta B^T P + \eta A \right)^{-1} \times \left(2\theta B^T P + \mu (Z_1 + Z_2) - \Lambda_1 - \Lambda_2 \right). \quad (22)$$

2) UPDATE $Z_1$

$Z_1$ can be updated iteratively as

$$Z_1^{m+1} = Z_1^m - \eta_1 \nabla Z_1,$$  \quad (23)

where $m$ represents the number of iterations, $\eta_1$ denotes the learning rate, and the gradient $\nabla Z_1$ is

$$\nabla Z_1 = \mu Z_1 - \mu A - \Lambda_2^T + \left(2\alpha Z_1 \left(Z_1^T Z_1 + \epsilon I_n \right)^{-1} \times \left(I_n - Z_1^T Z_1 (\epsilon I_n + Z_1^T Z_1)^{-1} \right) \right).$$ \quad (24)

3) UPDATE $Z_2$

The update rule of $Z_2$ can be written as

$$Z_2^{(m+1)} = Z_2^{(m+1)} - \text{diag} \left(Z_2^{(m+1)} \right),$$ \quad (25)

where $Z_2^{(m+1)} = Z_2^m - \eta_2 \nabla Z_2$, and the gradient $\nabla Z_2$ is

$$\nabla Z_2 = \omega (L_q)^T Z_2 W W^T + \omega L_q^T Z_2 W W^T + \mu Z_2 - \Lambda_1 - \mu A.$$ \quad (26)

Besides, given $A$, $Z_1$ and $Z_2$, the dual variables $\Lambda_1$ and $\Lambda_2$ can be updated by the rules for the dual variables in [17].

When the self-representation matrix $Z$ is updated, the constant $X'$ can be updated by the latest $Z$, and be used in the following updates.

B. UPDATE $W$

With fixed $F$, $S$, $Z$, by dropping other unrelated terms, problem (17) reduces to

$$\min_F \omega \text{Tr} \left(W^T X^T L_s^T X W \right) + \beta \|W\|_{2,1}$$

$$\text{s.t. } W^T W = I_k.$$ \quad (27)

The optimal solution $W$ can be derived by resorting to Algorithm 1 in [6].

C. UPDATE $F$

With fixed $W$, $S$, $Z$, by dropping other unrelated terms, problem (17) reduces to

$$\min_F \lambda \text{Tr} \left(F^T L_s F \right)$$

$$\text{s.t. } F \in \mathbb{R}^{n \times c}, \quad F^T F = L.$$ \quad (28)

Obviously, the optimal solution of $F$ is the matrix consisting of the eigenvectors corresponding to the $c$ smallest eigenvalues of the Laplacian matrix $L_s$.

D. UPDATE $S$

With fixed $W$, $F$, $Z$, by dropping unrelated terms, problem (17) reduces to

$$\min_S \sum_{t=1}^q h_t \|m_t (X') - S\|_F^2 + \omega \sum_{i,j} \|W^T x_i - W^T x_j\|_2^2 S_{ij}$$

$$+ \lambda \sum_{i,j} \|F_i - F_j\|_2^2 S_{ij}. \quad (29)$$

To simplify the optimization problem, let

$$M_{ij} \triangleq \|W^T x_i - W^T x_j\|_2^2,$$$$

$$N_{ij} \triangleq \|F_i - F_j\|_2^2,$$$$

$$Q \triangleq \sum_{t=1}^q h_t m_t (X') - \omega M - \lambda N. \quad (30)$$

Then, formula (29) can be reformulated as

$$\min_{S^{t=1} \in S \geq 0} \text{Tr} \left(S^T S \right) - \text{Tr} \left(S^T Q \right). \quad (31)$$

Moreover, in the view of each column of $S$, we decompose formula (31) into $n$ terms, then the Lagrangian equation of the $i$-th ($i = 1, 2, \ldots, n$) term is

$$L_2 (s_i, v, u_i) = s_i^T s_i - s_i^T q_i - v \left(s_i^T 1 - 1 \right) - u_i s_i,$$ \quad (32)

where $s_i$ and $q_i$ are the $i$-th column of $S$ and $Q$, respectively, $v$ is the multiplier for the equality constraint, and $u_i$ is the multiplier for the inequality constraint. For any $j$, the KKT condition of Equation (32) is as follows

$$\nabla S_{ij} L_2 = S_{ij} - Q_{ij} - v - u_{ij} = 0,$$$$

$$S_{ij} u_{ij} = 0,$$$$

$$u_{ij} \geq 0. \quad (33)$$

Thus, based on the above KKT conditions, each entry in $S$ can be derived by

$$S_{ij} = \begin{cases} Q_{ij} + v + u_{ij}, & \text{if } Q_{ij} + v + u_{ij} \geq 0 \\ 0, & \text{if } Q_{ij} + v + u_{ij} < 0. \end{cases} \quad (34)$$

So far, the update rules about all variables are provided above, these rules are then implemented repeatedly until the convergence or the achievement of maximum number of iteration. Furthermore, the update rules described above are summarized in the following algorithm.

E. COMPLEXITY ANALYSIS

The main computation cost of the optimization algorithm is the update of $Z$ and $F$. Specifically, the complexity of updating $Z$ is $O(n^3)$, since the matrix inversion and multiplication. The complexity of updating $F$ is $O(n^2)$, due to the implementation of SVD.

V. EXPERIMENTS

In this section, we perform validation experiments in clustering tasks.
Algorithm 1

Input: $X, m_t (t = 1, \cdots, q), \theta, \alpha, \beta, \omega, \eta, \epsilon, \mu, \lambda, c$

Output: Projection matrix $W$

Initialize: $\{Z_t\}_{t=1}^q, A, m_t(X)(t = 1, \cdots, q)$

while not converged do

Fix others and update $A$ by solving (22)
Fix others and update $Z_1$ by solving (23)
Fix others and update $Z_2$ by solving (25)
Fix others and update $W$ by solving (27)
Fix others and update $F$ by solving (28)
Fix others and update $S$ by solving (34)

end while

Obtain projection matrix $W$

---

Table 2: Datasets information ("Class" denotes the number of classes, "Instance" denotes the number of instances, and "Dim" denotes the feature dimensionality).

| Dataset | Class | Instance | Dim |
|---------|-------|----------|-----|
| COIL20 | 20    | 1440     | 1024|
| Yale    | 15    | 165      | 3456|
| MSRA25  | 12    | 1799     | 1024|
| Lung    | 5     | 203      | 3312|
| ORL     | 40    | 400      | 644 |
| USPS    | 10    | 256      | 9298|

A. DATASETS

In experiments, we use artificial dataset and public datasets to validate the effectiveness of our method, respectively. For the artificial data, the three-dimensional ring is used here, where three rings are in a three-dimensional space simultaneously and each ring has different radius or height to others. Specifically, the artificial dataset can be shown in Fig. 1. Besides, we use six public datasets, including object image (i.e. COIL201), human face (i.e. MSRA25 [18], ORL2, Yale [19]), biology (i.e. Lung [20]), handwritten digit (i.e. USPS [21]). Detailed information of the datasets are shown in Table 2.

B. EXPERIMENT SETTING

To evaluate the results of the experiments, the proposed approach is compared with the state-of-the-art UFS algorithms, such as SOGFS [6], LLCFS [8], LGSP [10], DGUFS [9], PDUFS [11], JSCFS [12], and GVLR [13]. And 8-fold cross validation is implemented. For the sake of fairness, the corresponding similarity metrics used in the above four works are fused in the proposed method, including weight squared Euclidean distance, KNN, squared Euclidean distance and heat kernel. In addition, we also verify the effectiveness of the proposed novel low-rank approximation. For the sake of fairness, the same UFS framework under different rank-norm relaxations are compared, such as self-adaptive multi-metric unsupervised feature selection with the nuclear norm (SMUN), the Gamma norm (SMUG) and the proposed low-rank approximation (Ours). Especially in SMUN, to compare with the state-of-the-art method, the weighted nuclear norm [22], the latest nuclear norm based method, is adopted here. Moreover, all parameters of existing approaches mentioned above are based on the respective parameter-searching strategy provided in the open source code or paper. Besides, all balance parameters in the proposed method are tuned from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 5, 10\}$. In addition, the parameter for the proposed rank-norm approximation $\epsilon$ is tuned from $\{50^{-3}, 10^{-3}, 50^{-2}, 10^{-2}, 50^{-1}\}$, while the parameter for the Gamma norm $\gamma$ is varied from $50^{-3}$ to $10^{-1}$ with step $50^{-3}$. Apart from that, the learning rate is set as 0.05, and the maximum number of iteration is 300. To evaluate the performance of these algorithms, we use clustering accurate (ACC) as evaluation metrics in this paper.

C. RESULTS

Table 2 shows the ACC performance on artificial dataset where the feature space of 10% and 30% original samples corrupted by noise arranging from -0.3 to 0.3. As shown in Table 2, we can see that, as the ratio of corrupted data increases, the ACC performance of each method decreases obviously. But, one may notice that, the ACC performance of the proposed method beats other methods, and the proposed method has slower deceasing speed of ACC performance with the increase of the ratio of corrupted data. This is because that, the noise problem can be alleviated by self-representation, and an adaptive multi-metric fusion by automatically integrating similarity across different metrics according to the specific data. On the other side, in the existing works usually leverage a single fixed metric to build similarity matrix, which cannot fit various feature types very well and even damage the local manifold structure.

Table 4 and Table 5 show the ACC performance on public datasets by selecting 50% and 30% features, respectively. As shown in Table 4 and Table 5, compared with SOGFS, LLCFS, DGUFS, which consider the local manifold structure only, the ACC performance of LGSP, SMUN, SMUG and Ours are obviously better. This is because that, making use of local manifold and global structure simultaneously can achieve higher accuracy than the method based on local or global structure preserving only. Furthermore, one may observe that, the ACC performance of SMUN, SMUG and Ours generally tend to outperform the ACC performance of LGSP. Hence, the effectiveness of the proposed adaptive multi-metric fusion method can be verified. Besides, it can be seen that, the ACC performance of SMUG and Ours is better than that of SMUN. This is because that, compared with the nuclear norm, the proposed low-rank approximation and the Gamma norm can be nearly unbiased to approximate the rank function. In addition, one may find that, the ACC performance of Ours, which is based on the proposed low-rank approximation, is almost about the same as it of SMUG. Therefore, the proposed low-rank approximation can achieve almost the same performance of the rank function approximation via the
Gamma norm. However, although the same performance is achieved, compared to the parameter \( \epsilon \) in the proposed low-rank approximation, the method based on the Gamma norm may require a finer-grained grid search for the parameter \( \gamma \). This is because that, better robustness can be derived by the proposed low-rank approximation.

Table 6 shows the effectiveness of similarity metric fusion, where the highest ACC under the different number of similarity metrics for fusion is demonstrated. As shown in Table 6, the ACC performance is inferior to its counterpart of existing methods, such as JSCFS and PDUFS, when only single fixed metric is leveraged to build similarity matrix. Besides, one
may observe that, as the different number of similarity metrics for fusion increases, an adaptive multi-metric fusion can be used on more and more metrics, thus the ACC performance of the proposed method enhances gradually.

To further study the robustness on the parameter of the Gamma norm and the proposed method, Table 7 shows how the parameters change of low-rank approximation methods (i.e., $\epsilon$ in the proposed low-rank approximation or $\gamma$ in the Gamma norm) impacts on ACC performance. Due to the length limit, the robustness is compared on the basis of the result of Table 4 only, namely, results of selecting the 50% features is considered only. In Table 7, the change of the parameters of rank-norm approximation methods is denoted by $f$ which means the percentage change of the parameters compared with the value chosen in Table 4. For example, $f = 10\%$ means that, the parameters are increased and decreased by 10% from the value in Table 4, and the corresponding ACC in Table 7 is the highest one among the two ACC values. As can be seen, the impact of the variation of $f$ on the ACC performance of Ours is smaller than its counterpart on the ACC performance of SMUG. Therefore, compared with the Gamma norm, the proposed method is less sensitive to the parameter $\epsilon$.

In Fig. 2 and 3, the curves of adaptive weights as a function of the number of iterations for the COIL20 dataset and the Lung dataset is shown. It can be seen that, the convergence rate of the adaptive metric weights in the proposed method is very fast. And all the similarity metrics used in this paper, by some iterations, they tend to converge to a fixed value. As can be seen from Fig 2, in the dataset COIL20, the KNN similarity metric is much more adaptable to the data than other methods, while the heat kernel method has the lowest adaptability. On the other side, for the dataset Lung, the heat kernel metric method has the highest fitness, followed by the squared Euclidean metric method. Conversely, the KNN metric method, which performs well in the dataset COIL20, has lower weights. It can be seen that under different datasets, the proposed method has a preference for the choice of measurement method.
The analysis of parameter sensitivity is shown in Fig. 4, where different combinations of balance parameter $\theta$ and $\omega$ are conducted the analysis and the ACC is taken as the performance metric. As shown in Fig. 4, under different parameter perturbations, the proposed method can obtain stable performance.

VI. CONCLUSION

In this paper, to address the problem that the fixed metric cannot fit various feature types well, we propose an adaptive multi-metric fusion by automatically fusing refined similarity across different metrics according to the specific data. Besides, to obtain better performance, we preform the local structure learning and global structure learning simultaneously by the proposed novel low-rank approximation. Via the proposed novel low-rank approximation, better tradeoff between the performance and robustness can be provided. Experimental results show that the accuracy performance of the proposed method can be enhanced by 2% – 11%, compared with the state-of-the-art methods.

REFERENCES

[1] M. Dash, K. Choi, P. Scheuermann, and H. Liu, “Feature selection for clustering—A filter solution,” in *Proc. IEEE Int. Conf. Data Mining*, Macau City, Japan, Dec. 2002, pp. 115–122.

[2] S. Tabakhi, P. Moradi, and F. Akhlaghian, “An unsupervised feature selection algorithm based on ant colony optimization,” *Eng. Appl. Artif. Intell.*, vol. 32, pp. 112–123, Jun. 2014.

[3] 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.

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

[5] W. Li and L. Wei, “Unsupervised feature selection based on low-rank regularized self-representation,” *Open Access Library J.*, vol. 7, no. 4, pp. 1–12, 2020.

[6] F. Nie, Z. Wei, and X. Li, “Unsupervised feature selection with structured graph optimization,” in *Proc. AAAI*, Phoenix, AZ, USA, Feb. 2016, pp. 1302–1308.

[7] X. He and P. Niyogi, “Locality preserving projections,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16, Dec. 2004, pp. 153–160.

[8] H. Zeng and Y.-M. Cheung, “Feature selection and kernel learning for local learning-based clustering,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1532–1547, Aug. 2011.

[9] J. Guo and W. Zuo, “Dependence guided unsupervised feature selection,” in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Apr. 2018, pp. 2232–2239.

[10] X. Zuo, S. Zhang, R. Hu, Y. Zuo, 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.

[11] H. Lim and D.-W. Kim, “Pairwise dependence-based unsupervised feature selection,” *Pattern Recogn.*, vol. 111, pp. 107663–107674, Mar. 2021.

[12] J. Zhu, J. Jang-Jaccard, T. Liu, and J. Zhou, “Joint spectral clustering based on optimal graph and feature selection,” *Neural Process. Lett.*, vol. 53, no. 1, pp. 257–273, Feb. 2021.

[13] C. Sheng and P. Song, “Graph regularized virtual label regression for unsupervised feature selection,” *Digit. Signal Process.*, vol. 123, pp. 103393–103402, Apr. 2022.

[14] D. Wang, F. Nie, and H. Huang, “Unsupervised feature selection via unified trace ratio formulation and K-means clustering (track),” in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Nancy, France, Jan. 2014, pp. 306–321.

[15] H. Zhang, J. Yang, J. Xie, J. Qian, and B. Zhang, “Weighted sparse coding regularized nonconvex matrix regression for robust face recognition,” *Inf. Sci.*, vol. 394, pp. 1–17, Jul. 2017.

[16] W.-B. Xie, B. Liu, L.-W. Bu, Y.-L. Wang, and J. Zhang, “A decoupling approach for observer-based controller design of T–S fuzzy system with unknown premise variables,” *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 9, pp. 2714–2725, Sep. 2021.

[17] M. Brbic and J. Kopriva, “Multi-view low-rank sparse subspace clustering,” *Pattern Recognit.*, vol. 73, pp. 247–258, Jan. 2018.

[18] H. Li, M. Wang, and X.-S. Hua, “MSRA-MM 2.0: A large-scale web multimedia dataset,” in *Proc. IEEE Int. Conf. Data Mining Workshops*, Miami, FL, USA, Dec. 2009, pp. 164–169.

[19] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, “Eigenfaces vs. Fish-erfaces: Recognition using class specific linear projection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[20] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, and P. Tamayo, “Gene expression correlates of clinical prostate cancer behavior,” *Cancer Cell*, vol. 1, no. 2, pp. 203–209, Mar. 2002.

[21] J. J. Hull, “A database for handwritten text recognition research,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, May 1994.

[22] S. Gu, Q. Xie, D. Meng, W. Zuo, X. Feng, and L. Zhang, “Weighted nuclear norm minimization and its applications to low level vision,” *Int. J. Comput. Vis.*, vol. 121, no. 2, pp. 183–208, Jan. 2017.