Information Sciences 554 (2021) 1–14

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Deep random walk of unitary invariance for large-scale data representation

Shiping Wang [a,b], Zhaoliang Chen [a,b], William Zhu [c], Fei-Yue Wang [d,*]

[a] College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China
[b] Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou University, Fuzhou 350116, China
[c] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China
[d] State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences (SKL-MCCS, CASIA), Beijing 100190, China

ARTICLE INFO

Article history:
Received 22 June 2020
Received in revised form 20 November 2020
Accepted 21 November 2020
Available online 1 December 2020

Keywords:
Machine learning
Feature representation
Deep random walk
Unitary invariance
Closed-form solution

ABSTRACT

Data representation aims at learning an efficient low-dimensional representation, which is always a challenging task in machine learning and computer vision. It can largely improve the performance of specific learning tasks. Unsupervised methods are extensively applied to data representation, which considers the internal connection among data. Most of existing unsupervised models usually use a specific norm to favor certain distributions of the input data, leading to an unsustainable encouraging performance for given learning tasks. In this paper, we propose an efficient data representation method to address large-scale feature representation problems, where the deep random walk of unitary invariance is exploited for learning discriminative features. First, the data representation is formulated as deep random walk problems, where unitarily invariant norms are employed to capture diverse beneficial perspectives hidden in the data. It is embedded into a state transition matrix model, where an arbitrary number of transition steps is available for an accurate affinity evaluation. Second, data representation problems are then transformed as high-order matrix factorization tasks with unitary invariance. Third, a closed-form solution is proved for the formulated data representation problem, which may provide a new perspective for solving high-order matrix factorization problems. Finally, extensive comparative experiments are conducted in publicly available real-world data sets. In addition, experimental results demonstrate that the proposed method achieves better performance than other compared state-of-the-art approaches in terms of data clustering.
© 2020 Elsevier Inc. All rights reserved.

1. Introduction

In real-world applications, specific learning tasks are frequently characterized by large-scale data of extremely high dimensions, which requires to efficiently discover beneficial patterns from given massive data. However, most of the extracted features are redundant, irrelevant and inconsistent, which brings about negative effects on learning algorithms and then decreases the performance of learning tasks. Towards this end, how to learn an efficient low-dimensional data representation from massive data is crucial and challenging [20]. A favorable data representation is indicative of powerful discriminative abilities and generalization capabilities of predictive algorithms.

* Corresponding author.
E-mail address: feiyue.wang@ia.ac.cn (F.-Y. Wang).

https://doi.org/10.1016/j.ins.2020.11.039
0020-0255/© 2020 Elsevier Inc. All rights reserved.

Data representation methods tend to be divided as two categories, namely feature selection and feature extraction. The former aims at finding a minimal feature subset from the original feature space based on certain criteria, such as mutual information maximization [30], orthogonal projection maximization [42] and sparse structure preservation [23,43], while the latter is to search a low-dimensional representation through a certain transformation whose typical representatives consist of principal component analysis (PCA) [29], linear discriminant analysis (LDA) [49], locally linear embedding (LLE) [33], and locality preserving projection (LPP) [14]. There have been also a variety of well-known methods for data representation, such as canonical correlation analysis [12] and non-negative matrix factorization [21].

Particularly, matrix factorization has captured much attention in the last decade since it provided new directions for problem formulation and algorithm development in a batch mode [5,6,34]. Most of these data representation methods used a specific distance metric to evaluate the reconstruction error between observed samples and ground truths. For example, Xu et al. employed $\ell_{2,1}$-norm to measure the canonical correlation loss term for data representation [46]. Lu et al. put forward a nuclear norm-based 2-D linear discriminant analysis (HVNN-2DLDA) method for image representation [26]. Lu et al. defined a new tensor nuclear norm and revisited the tensor robust principal component analysis problem [25]. Zhu et al. proposed a robust ℓ_p-norm sparse representation model with adaptive feature weights for biometric image classification [54]. Nevertheless, a specific norm tends to favor certain distributions of the input data and leads to unsatisfying performance in some practical applications. Towards this end, a unified scheme for data representation is constructed, in which unitarily invariant norms are employed to capture multi-view distributions and deep random walk is used to exploit multi-scale nonlinear data representations. In this scheme, any norm minimization problem of unitary invariance is solved by an almost closed-form solution, aiming to provide a new distance metric learning perspective to uncover rewarding patterns from data.

In this paper, a new unified scheme for data representation problems is proposed, where deep random walk is designed to characterize the state transition process of given samples, and unitarily invariant norms are employed to evaluate the residual reconstruction, as demonstrated in Fig. 1. In the first place, the unsupervised data representation is characterized by unitarily invariant norms, where the deep random walk serves as a beneficial pipeline for learning multi-scale nonlinear features. There is one more point, feature representation problems are formulated as high-order matrix factorization problems with unitarily invariant norm minimization. The last but not the least, a closed-form solution is proposed for the formulated optimization problems of unitary invariance. In addition, the proposed method is tested in publicly available real-world data sets including image data, text documents and gene expressions. Comprehensive experiments demonstrate that the proposed method is superior to other compared state-of-the-art feature representation methods in terms of data clustering.

The outline of this paper is arranged as follows. Related work is recalled in Section 2 and a large-scale data representation method based on deep random walk of unitary invariance is proposed in Section 3. In Section 4, comprehensive experiments are conducted to validate the effectiveness and efficiency of the proposed method, and the concluding remark is provided in Section 5.

2. Related work

Data representation is a crucial task in machine learning and computer vision. It aims at learning a low-dimensional representation from given large-scale data that may contain a great deal of noisy or uncorrelated features. By whether the supervised information from class labels is available, data representation can be categorized as supervised, weakly supervised or unsupervised methods. The supervised data representation is to search a low-dimensional subspace under the guidance of class label information [39], while the unsupervised one aims to learn a discriminative subspace using the internal connection among the given data points [18,50]. Weakly supervised data representation falls into an intermediate medium between these two methods, where only a small quantity of exactly labeled training data is available [27].

In the last decade, data representation learning has captured growing attention from researchers in a variety of areas [9,31,45]. As examples, it exerted the effectiveness and efficiency in depth image reconstruction [4], text categorization [19], multi-view learning [7], social image understanding [24], large-scale image data mining [16,47] and video analysis [10]. In particular, Xuan et al. constructed a deep neural network based feature representation, namely variable-wise weighted stacked auto-encoder for soft sensor modeling [52]. Feige introduced an efficient approach to represent data with an invariant and equivariant manner [8]. Wang et al. established a general framework dubbed discerning feature supported encoder, which integrated the auto-encoder and feature selection into a unified model to learn an effective feature representation [40]. Chen et al. put forward a discriminative feature representation method, which benefited both domain alignment and data classification [3]. Wang and Guo came up with an efficient multimodal feature representation algorithm based sparse multi-graph embedding [41]. Ayas and Ekinci introduced a sparse representation based multi-scale feature fusion method to generate a high resolution image [1]. Huynh-The et al. proposed an encoding technique for pose-transition feature learning and constructed an efficient action recognition fine-tuning model [17].

Based on the above analyses, it is suggested that the outlined representative work comes with varying benefits in diverse application scenarios. Nevertheless, these existing works in an unsupervised learning were fed by an affinity matrix that tended to be computed by a specific kernel, which may lead to an inaccurate evaluation. There is one more point that these existing studies took no account of distance metric learning based model construction, problem optimization and algorithm designing.

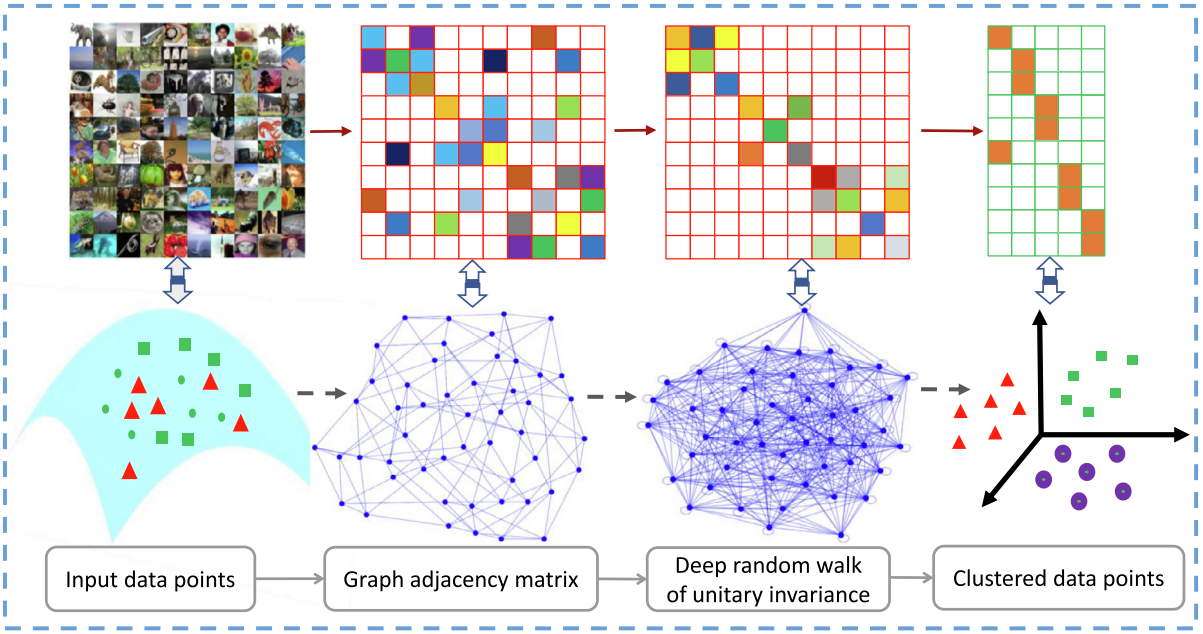

Fig. 1. An illustration of the proposed method. Leveraging deep random walk of unitary invariance, an accurate affinity matrix of samples is constructed and a discriminative low-dimensional representation is learned.

3. Deep random walk of unitary invariance for data representation

For clarity, a number of frequently used notations are recalled. Specifically, the input set of data points is denoted as $\{\mathbf{x}_i\}_{i=1}^n$ where $\mathbf{x}_i$ is a d-dimensional row vector, i.e., $\mathbf{x}_i \in \mathbb{R}^d$ for any $i \in \{1, \cdots, n\}$. The input data matrix is given by $\mathbf{X} = [\mathbf{x}_1; \cdots; \mathbf{x}_n] \in \mathbb{R}^{n \times d}$. And a norm $||\cdot||$ is called unitarily invariant (simply denoted as $||\cdot||_{\text{UI}}$ [51]) if $||\mathbf{M}|| = ||\mathbf{UMV}||$ for any matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ and unitary matrices $\mathbf{U} \in \mathbb{R}^{m \times m}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$. As one of typical representatives of unitarily invariant norms, Schatten p-norm is defined as

$$||\mathbf{M}||_{S_p} = \left(\sum_{i=1}^{\min\{m,n\}} \sigma_i^p(\mathbf{M}) \right)^{1/p}, \tag{1}$$

where $\sigma_i(\mathbf{M})$ is the i-th singular value of $\mathbf{M} \in \mathbb{R}^{m \times n}$.

3.1. Deep random walk for problem formulation

For a given set of data points $\{\mathbf{x}_i\}_{i=1}^n$, its affinity matrix $\mathbf{K} = [\mathbf{K}_{ij}]_{n \times n}$ can be computed by kernel functions, such as Gaussian kernel $\mathbf{K}_{ij} = \exp\left(\frac{-||\mathbf{x}_i - \mathbf{x}_j||_2^2}{2\sigma^2}\right)$, where σ is the kernel width to be predefined. After matrix normalization, $\mathbf{P} = \mathbf{D}^{-1/2}\mathbf{K}\mathbf{D}^{-1/2}$ can be regarded as a probability transition matrix of a random walk, where $\mathbf{D} = [\mathbf{D}_{ij}]_{n \times n}$ is diagonal with $\mathbf{D}_{ii} = \sum_{j=1}^n \mathbf{K}_{ij}$. It is noted that $\mathbf{P}_{ij} \in [0, 1]$ can be interpreted as the probability from the i-th state to the j-th state using only one step.

A main application of deep random walk is to learn a vertex sequence in network discovery [28,32]. For example, DeepWalk method aims at maximizing the overall accessible probability between a target state and its context states within certain transition steps [37,11]. Given a sequence of states $\mathbf{S} = \{\mathbf{s}_1, \cdots, \mathbf{s}_n\}$ and transition step t, the objective function of DeepWalk is represented by the form of

$$\sum_{\mathbf{s} \in \mathbf{S}} \left[\frac{1}{n} \sum_{i=1}^{n} \sum_{-t+i \leqslant j \leqslant t+i, j \neq i} \log P(\mathbf{s}_j | \mathbf{s}_i) \right]. \tag{2}$$

The transition probability $P(\mathbf{s}_j|\mathbf{s}_i)$ is frequently evaluated by the softmax value of the inner product between pairwise samples, defined as

$$P(\mathbf{s}_j|\mathbf{s}_i) = \frac{\exp\left(\mathbf{v}_j^T\mathbf{v}_i\right)}{\sum_{k\in\mathbf{S}}\exp(\mathbf{v}_k^T\mathbf{v}_i)}, \tag{3}$$

where $\mathbf{v}_i$ is the corresponding feature vector of the state $\mathbf{s}_i$. Yang et al. [48] proved that DeepWalk actually is a matrix factorization with the input matrix $\mathbf{A} = [\mathbf{A}_{ij}]_{n\times n}$, defined by

$$\mathbf{A}_{ij} = \log\frac{\left[e\left(\mathbf{P}+\mathbf{P}^2+\cdots+\mathbf{P}^t\right)\right]_{ij}}{t}, \tag{4}$$

where $\left[e\left(\mathbf{P}^k\right)\right]_{ij}$ is the probability that the state i walks to the state j at exactly k steps. Motivated by reinforcement learning [35,38], we view $\mathbf{A}$ as an assignment strategy matrix of the long-run reward. Intuitively, the received rewards in the future will be geometrically discounted over time. Correspondingly, the probability from the i-th state to the j-th state using exactly k steps is computed by $\left[\gamma^k\mathbf{P}^k\right]_{ij}$, in which $\gamma\in[0,1)$ is a decay factor to measure the tendency of a rewarding message to lose its effectiveness over time. In another perspective, γ serves as a discount rate, being indicative of the weaker weight on the total cost function as the number of transition steps increases.

When it comes to the data representation problems, the overall sample affinity matrix $\mathbf{A} = [\mathbf{A}_{ij}]_{n\times n}$ within t transition steps is represented as

$$\mathbf{A}_{ij} = \log\frac{\left[e\left(\gamma\mathbf{P}+[\gamma\mathbf{P}]^2+\cdots+[\gamma\mathbf{P}]^t\right)\right]_{ij}}{t}. \tag{5}$$

The introduced discount factor also provides a mathematical trick to guarantee an upper bound of the infinite sum. Consequently, when taking into account of all possible numbers of transition steps, the affinity matrix $\mathbf{A}$ is computed by

$$\mathbf{A}_{ij} = \log\frac{\left[e(\mathbf{I}-\gamma\mathbf{P})^{-1}\right]_{ij}}{t}. \tag{6}$$

With the constructed affinity matrix $\mathbf{A}$, the objective optimization problem for data representation is then formulated as

$$\min_{\mathbf{Q}\in\mathbb{R}^{n\times k}} ||\mathbf{A}-\mathbf{Q}\mathbf{Q}^T||_{\mathrm{UI}} + \alpha||\mathbf{Q}||_{\mathrm{UI}'}, \tag{7}$$

where $||\cdot||_{\mathrm{UI}}$ and $||\cdot||_{\mathrm{UI}'}$ are two unitarily invariant norms that may differ, $\alpha\geqslant 0$ is a weighted parameter to balance the first fitting term and the second regularization term, and k is the number of reduced dimensions. Herein, $\mathbf{Q}\in\mathbb{R}^{n\times k}$ is a low-dimensional representation of the original data matrix $\mathbf{X}\in\mathbb{R}^{n\times d}$ with $k\ll d$.

3.2. Closed-form optimization algorithm

The singular value decomposition of a matrix $\mathbf{M}\in\mathbb{R}^{m\times n}$ is the factorization of the form $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where both $\mathbf{U}\in\mathbb{R}^{m\times m}$ and $\mathbf{V}\in\mathbb{R}^{n\times n}$ are unitary matrices, and $\mathbf{\Sigma}\in\mathbb{R}^{m\times n}$ is the rectangular diagonal matrix, simply denoted by $(\mathbf{U},\mathbf{\Sigma},\mathbf{V}) = svd(\mathbf{M})$.

In order to explore the closed-form solution to the aforementioned data representation problem, a useful lemma is introduced.

Lemma 1. [53] Let $\mathbf{A},\mathbf{B}\in\mathbb{R}^{m\times n}$, $(\mathbf{U_A},\mathbf{\Sigma_A},\mathbf{V_A}) = svd(\mathbf{A})$, and $(\mathbf{U_B},\mathbf{\Sigma}_B,\mathbf{V_B}) = svd(\mathbf{B})$. Then for any unitarily invariant norm $||\cdot||_{\mathrm{UI}}$, it holds

$$||\mathbf{\Sigma}_A-\mathbf{\Sigma}_B||_{\mathrm{UI}} \leqslant ||\mathbf{A}-\mathbf{B}||_{\mathrm{UI}}. \tag{8}$$

The above lemma reveals that the matrix consisting of all singular values is the minimum stationary point of every unitarily invariant norm. Based on the lemma, the optimal solution to unitarily invariant norm minimization problems is provided in an implicit form.

Theorem 1. *For a given affinity matrix $\mathbf{A}\in\mathbb{R}^{n\times n}$, the minimizer of the objective optimization problem*

$$\min_{\mathbf{Q}\in\mathbb{R}^{n\times k}} ||\mathbf{A}-\mathbf{Q}\mathbf{Q}^T||_{\mathrm{UI}} + \alpha||\mathbf{Q}||_{\mathrm{UI}'} \tag{9}$$

has the form $\mathbf{Q}^* = \mathbf{U}_{\mathbf{A}}^{(k)}\mathbf{\Lambda}_k\mathbf{V}_k^T$ where $\mathbf{\Lambda}_k\in\mathbb{R}^{k\times k}$ is a positive diagonal matrix, $\mathbf{U}_{\mathbf{A}}^{(k)}$ is a k-truncation of $\mathbf{U_A}$ with $(\mathbf{U_A},\mathbf{\Sigma_A},\mathbf{U_A}) = svd(\mathbf{A})$, and $\mathbf{V}_k\in\mathbb{R}^{k\times k}$ is an arbitrary orthogonal matrix.

Proof. Since $\mathbf{A}$ is a symmetric matrix, we can write $(\mathbf{U_A}, \mathbf{\Sigma_A}, \mathbf{U_A}) = svd(\mathbf{A})$ with real-valued factors. For any matrix $\mathbf{Q} \in \mathbb{R}^{n\times k}$, let $(\mathbf{U_Q}, \mathbf{\Sigma_Q}, \mathbf{V_Q}) = svd(\mathbf{Q})$. It is noted that $||\mathbf{Q}||_{UI'} = ||\mathbf{\Sigma_Q}||_{UI'}$ and $\mathbf{QQ}^T = \left(\mathbf{U_Q\Sigma_Q V_Q^T}\right)\left(\mathbf{U_Q\Sigma_Q V_Q^T}\right)^T = \mathbf{U_Q\Sigma_Q\Sigma_Q^T U_Q^T}$. According to Lemma 1, it follows

$$\begin{aligned} f(\mathbf{Q}) &\triangleq ||\mathbf{A} - \mathbf{QQ}^T||_{UI} + \alpha||\mathbf{Q}||_{UI'} \\ &= ||\mathbf{U_A\Sigma_A U_A^T} - \mathbf{U_Q\Sigma_Q\Sigma_Q^T U_Q^T}||_{UI} + \alpha||\mathbf{\Sigma_Q}||_{UI'} \\ &\geqslant ||\mathbf{\Sigma_A} - \mathbf{\Sigma_Q\Sigma_Q^T}||_{UI} + \alpha||\mathbf{\Sigma_Q}||_{UI'} \\ &= ||\mathbf{\Sigma_A} - \left[\mathbf{\Sigma_Q^{(n)}}\right]^2||_{UI} + \alpha||\mathbf{\Sigma_Q^{(k)}}||_{UI'} = f\left(\mathbf{U_A U_Q^T Q}\right), \end{aligned} \tag{10}$$

where $\mathbf{\Sigma_Q^{(k)}} \in \mathbb{R}^{k\times k}$ is the k-truncation of $\mathbf{\Sigma_Q}$, i.e., $\mathbf{\Sigma_Q} = \left[\mathbf{\Sigma_Q^{(k)}}; \mathbf{O}_{(n-k)\times k}\right]$ with $\mathbf{O}_{(n-k)\times k}$ being a zero matrix, and $\left[\mathbf{\Sigma_Q^{(n)}}\right]^2 = \begin{bmatrix} \left(\mathbf{\Sigma_Q^{(k)}}\right)^2 & \mathbf{O}_{k\times(n-k)} \\ \mathbf{O}_{(n-k)\times k} & \mathbf{O}_{(n-k)\times(n-k)} \end{bmatrix}$. To sum up, it holds $f(\mathbf{Q}) \geqslant f(\mathbf{U_A U_Q Q})$ for any $\mathbf{Q} \in \mathbb{R}^{n\times k}$, which implies that the minimizer of the above optimization problem has the form of $\mathbf{Q}^* = \mathbf{U_A}\mathbf{\Lambda}_{n\times k}\mathbf{V}_k^T = \mathbf{U_A^{(k)}}\mathbf{\Lambda}_k\mathbf{V}_k^T$.

As mentioned before, unitarily invariant norms can be regarded as an operator for structured learning, whereby the top singular values are preserved. When resorting to special types of unitary invariance, the closed-form solution to the above minimization problem is further explored. Substituting the unitarily invariant norms with Schatten norms, the objective optimization problem has the following explicit analytic solution.

Theorem 2. *For a given affinity matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ with $(\mathbf{U_A}, \mathbf{\Sigma_A}, \mathbf{V_A}) = svd(\mathbf{A})$, the minimizer of the objective optimization problem*

$$\min_{\mathbf{Q}\in\mathbb{R}^{n\times k}} ||\mathbf{A} - \mathbf{QQ}^T||_{S_p} + \alpha||\mathbf{Q}||_{S_q} \tag{11}$$

is attained at $\mathbf{Q}^* = \mathbf{U_A^{(k)}}\mathbf{\Lambda}_k\mathbf{V}_k^T$ where $\mathbf{\Lambda}_k = diag(\Lambda_1, \cdots, \Lambda_k)$ is computed by the following vector optimization problem

$$\min_{\mathbf{\Lambda}_k} \left[\sum_{i=1}^{k} |\sigma_{j_i} - \Lambda_i^2|^p + \sum_{j_i \notin \{j_1,\cdots,j_k\}} \sigma_{j_i}^p\right]^{1/p} + \alpha\left[\sum_{i=1}^{k} |\Lambda_i|^q\right]^{1/q}. \tag{12}$$

Here, $\mathbf{U_A^{(k)}}$ is the corresponding columns of $\mathbf{U_A}$ indexed by $\{j_1, \cdots, j_k\}$, and $\mathbf{V}_k \in \mathbb{R}^{k\times k}$ is an arbitrary orthogonal matrix.

Proof. Since Schatten norm is unitarily invariant, according to Theorem 1, the optimal solution to Problem 11 has the form $\mathbf{Q}^* = \mathbf{U_A^{(k)}}\mathbf{\Lambda}_k\mathbf{V}_k^T$. It is noted that $\mathbf{Q}^*\mathbf{Q}^{*T} = \left[\mathbf{U_A^{(k)}}\mathbf{\Lambda}_k\mathbf{V}_k^T\right]\left[\mathbf{U_A^{(k)}}\mathbf{\Lambda}_k\mathbf{V}_k^T\right]^T = \mathbf{U_A^{(k)}}\mathbf{\Lambda}_k^2\mathbf{U_A^{(k)}}^T$. Without loss of generality, we assume that $\mathbf{U_A^{(k)}}$ is the corresponding columns of $\mathbf{U_A}$ indexed by $\{j_1, \cdots, j_k\}$ and $\widetilde{\mathbf{U}}_\mathbf{A}^{(k)} \in \mathbb{R}^{n\times(n-k)}$ is its complement, then $\mathbf{U_A^{(k)}}\mathbf{\Lambda}_k^2\mathbf{U_A^{(k)}}^T = \left[\mathbf{U_A^{(k)}}, \widetilde{\mathbf{U}}_\mathbf{A}^{(k)}\right]\begin{bmatrix} \mathbf{\Lambda}_k^2 & \mathbf{O}_{k\times(n-k)} \\ \mathbf{O}_{(n-k)\times k} & \mathbf{O}_{(n-k)\times(n-k)} \end{bmatrix}\left[\mathbf{U_A^{(k)}}, \widetilde{\mathbf{U}}_\mathbf{A}^{(k)}\right]^T$. Accordingly, $||\mathbf{A} - \mathbf{QQ}^T||_{S_p} = \left[\sum_{i=1}^{k}|\sigma_{j_i} - \Lambda_i^2|^p + \sum_{j_i\notin\{j_1,\cdots,j_k\}}\sigma_{j_i}^p\right]^{1/p}$, completing the proof.

The above theorem provides a method to solve the optimization problem, however this method is computationally time-consuming. The following lemma is proposed to alleviate the optimization difficulty.

Lemma 2. *Suppose $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_{k+1}, \mathbf{\Lambda} = [\Lambda_1, \cdots, \Lambda_k]$ and $p \geqslant 1, q \geqslant 1, \alpha \geqslant 0$, then it holds*

$$\begin{aligned} &\min_{\mathbf{\Lambda}} \left[\sum_{i=1}^{k} |\sigma_i - \Lambda_i^2|^p + \sigma_{k+1}^p\right]^{\frac{1}{p}} + \alpha\left[\sum_{i=1}^{k}|\Lambda_i|^q\right]^{\frac{1}{q}} \leqslant \\ &\min_{\mathbf{\Lambda}} \left[\sum_{i=2}^{k} |\sigma_i - \Lambda_i^2|^p + |\sigma_{k+1} - \Lambda_1^2|^p + \sigma_1^p\right]^{\frac{1}{p}} + \alpha\left[\sum_{i=1}^{k}|\Lambda_i|^q\right]^{\frac{1}{q}}. \end{aligned} \tag{13}$$

Proof. The inequality is evidently true if $|\sigma_1 - \Lambda_1^2|^p + \sigma_{k+1}^p \leqslant |\sigma_{k+1} - \Lambda_1^2|^p + \sigma_1^p$ for any $\Lambda_1 \leqslant \sigma_1$ with given $\sigma_1 \geqslant \sigma_{k+1}$. Actually, denote $f(\Lambda_1) = |\sigma_{k+1} - \Lambda_1^2|^p + \sigma_1^p - \left(|\sigma_1 - \Lambda_1^2|^p + \sigma_{k+1}^p\right) = \left(\sigma_{k+1} - \Lambda_1^2\right)^p + \sigma_1^p - \left(\sigma_1 - \Lambda_1^2\right)^p - \sigma_{k+1}^p$, then we know $\frac{df(\Lambda_1)}{d\Lambda_1} = 2p\Lambda_1\left[(\sigma_1 - \Lambda_1)^{p-1} - (\sigma_{k+1} - \Lambda_1)^{p-1}\right] \geqslant 0$. Hence, it is evident that $\min_{\Lambda_1\in[0,\sigma_{k+1}]} f(\Lambda_1) \geqslant f(0) = 0$, which implies that the inequality holds, completing the proof.

According to the aforementioned lemmas, the nearly closed-form solution to the objective optimization problem is explored.

Theorem 3. *For a given affinity matrix* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *with* $(\mathbf{U_A}, \mathbf{\Sigma_A}, \mathbf{V_A}) = svd(\mathbf{A})$*, the minimizer of the objective optimization problem*

$$\min_{\mathbf{Q}\in\mathbb{R}^{n\times k}} ||\mathbf{A} - \mathbf{Q}\mathbf{Q}^T||_{S_p} + \alpha||\mathbf{Q}||_{S_q} \tag{14}$$

has a (nearly) closed-form solution $\mathbf{Q}^* = \mathbf{U}_\mathbf{A}^{(k)}\mathbf{\Lambda}_k\mathbf{V}_k^T$ *where* $\mathbf{U}_\mathbf{A}^{(k)}$ *is the corresponding k-truncation of* $\mathbf{U_A}, \mathbf{V}_k \in \mathbb{R}^{k\times k}$ *is an arbitrary orthogonal matrix, and* $\mathbf{\Lambda}_k = diag(\Lambda_1, \cdots, \Lambda_k)$ *is solved by*

$$\min_{\mathbf{\Lambda}_k}\left[\sum_{i=1}^{k}|\sigma_i - \Lambda_i^2|^p + \sum_{j=k+1}^{rank(\mathbf{X})}\sigma_j^p\right]^{1/p} + \alpha\left[\sum_{i=1}^{k}|\Lambda_i|^q\right]^{1/q}. \tag{15}$$

Proof. According to the inequality in Lemma 2, we obtain

$$\begin{aligned}&\min_{\mathbf{\Lambda}_k}\left[\sum_{i=1}^{k}|\sigma_i - \Lambda_i^2|^p + \sum_{i=k+1}^{rank(\mathbf{X})}\sigma_i^p\right]^{1/p} + \alpha\left[\sum_{i=1}^{k}|\Lambda_i|^q\right]^{1/q} \leqslant \\ &\min_{\mathbf{\Lambda}_k}\left[\sum_{i=1}^{k}|\sigma_{j_i} - \Lambda_i^2|^p + \sum_{j_i\notin\{j_1,\cdots j_k\}}\sigma_{j_i}^p\right]^{1/p} + \alpha\left[\sum_{i=1}^{k}|\Lambda_i|^q\right]^{1/q}.\end{aligned} \tag{16}$$

Together with Theorem 2, this completes the proof.

In order to obtain an explicit closed-form solution to the optimization problem, the following corollary is proposed to solve the decoupled optimization.

Corollary 1. *For a given affinity matrix* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *with* $(\mathbf{U_A}, \mathbf{\Sigma_A}, \mathbf{V_A}) = svd(\mathbf{A})$*, the minimizer of the objective optimization problem*

$$\min_{\mathbf{Q}\in\mathbb{R}^{n\times k}} \frac{1}{p}||\mathbf{A} - \mathbf{Q}\mathbf{Q}^T||_{S_p}^p + \frac{2\alpha}{q}||\mathbf{Q}||_{S_q}^q \tag{17}$$

has a (nearly) closed-form solution $\mathbf{Q}^* = \mathbf{U}_\mathbf{A}^{(k)}\mathbf{\Lambda}_k\mathbf{V}_k^T$ *where* $\mathbf{U}_\mathbf{A}^{(k)}$ *is the corresponding k-truncation of* $\mathbf{U_A}, \mathbf{V}_k \in \mathbb{R}^{k\times k}$ *is an arbitrary orthogonal matrix, and* $\mathbf{\Lambda}_k = diag(\Lambda_1, \cdots, \Lambda_k)$ *is computed by*

$$\min_{\mathbf{\Lambda}_k} \frac{1}{p}\sum_{i=1}^{k}|\sigma_i - \Lambda_i^2|^p + \frac{2\alpha}{q}\sum_{i=1}^{k}|\Lambda_i|^q. \tag{18}$$

With the aforementioned analyses in details, closed-form solutions to the data representation problem with specific values of p and q are provided in Table 1. Besides, the procedure for the problem is summarized in Algorithm 1.

Algorithm 1: DRWDR: Deep Random Walk based Data Representation.

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and the number k of reduced dimensions.
Output: A low-dimensional representation $\mathbf{Q} \in \mathbb{R}^{n\times k}$.
1: Initialize hyper-parameters including step number t, decay factor γ, regularization coefficient α and Schatten norm indicators p and q;
2: Construct the probability transition matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ by kernel evaluation and matrix normalization;
3: Compute the overall affinity matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ with deep random walk using Eq. (5);
4: Solve the diagonal matrix $\mathbf{\Lambda}_k \in \mathbb{R}^{k\times k}$ in Problem (18) with a vector optimization method;
5: Reconstruct the low-dimensional data representation $\mathbf{Q} \in \mathbb{R}^{n\times k}$ using the closed-form solution given in Theorem (3);
6: **return**Data representation $\mathbf{Q}$.

The computational complexity of the probability transition matrix construction is $O\left(dn^2\right)$, and that of overall affinity matrix evaluation needs $O(n^3)$. Besides, the complexities of the vector optimization and low-dimensional reconstruction require $O(n^3)$. Accordingly, the overall computational complexity of the proposed method is $O(n^2(n+d))$.

4. Experimental analysis

In this section, extensive experiments are conducted to verify the effectiveness and efficiency of the proposed method. Here, the test data sets range from face and digit images to gene expressions and text documents, which provides a reliable test platform for algorithm comparisons.

Table 1
Closed-form solutions to data representation problems of decoupled forms with some specific p and q. Denote $(\mathbf{U}_\mathbf{A}, \mathbf{\Sigma}_\mathbf{A}, \mathbf{V}_\mathbf{A}) = svd(\mathbf{A})$.

Schatten norm p and q	Optimal solution $\mathbf{Q}^* = \mathbf{U}_\mathbf{A}^{(k)}\mathbf{\Lambda}_k$
$p = 1, q = 2$	$\mathbf{\Lambda}_k = \left[\mathbf{\Sigma}_\mathbf{A}^{(k)}\right]^{1/2}$ if $\alpha \in (0, 1]$, and $\mathbf{\Lambda}_k = \mathbf{0}$ if $\alpha \in (1, \infty)$
$p = 2, q = 2$	$\mathbf{\Lambda}_k = \left[\left(\mathbf{\Sigma}_\mathbf{A}^{(k)} - \alpha\mathbf{I}\right)_+\right]^{1/2}$
$p \geqslant 2, q = 2$	$\mathbf{\Lambda}_k = \left[\left(\mathbf{\Sigma}_\mathbf{A}^{(k)} - \alpha^{\frac{1}{p-1}}\mathbf{I}\right)_+\right]^{1/2}$
$p = 1, q = 2m > 2$	$\mathbf{\Lambda}_k = \alpha^{-1/(q-2)}\mathbf{I}$
$p = m, q = 2m \geqslant 2$	$\mathbf{\Lambda}_k = \left[\frac{1}{1+\alpha^{1/(m-1)}}\mathbf{\Sigma}_\mathbf{A}^{(k)}\right]^{1/2}$

4.1. Data sets

All test data sets are derived from publicly available machine learning repositories, aiming at providing a fair test bed for the compared data representation methods. Each data set is characterized in more details, shown as follows.

ORL consists of 400 face images from 40 distinct individuals [1]. These images were taken with varying facial expressions as the lightening and background differ. Each sample is a 4,096-dimensional feature vector, reshaped by a 64×64 gray-scale image.

FEI is a Brazilian face database that was taken against a white homogenous background [2]. There are 200 characters for each of 14 images, a total of 2,800 images. Each image was cropped to the size of 60×80×3 pixels, forming a 14,400-dimensional feature vector.

GISETTE is a digit recognition data set that aims to make a distinction between two highly confusible digits '4' and '9' [3]. Each digit image contains 5,000 features, where a large number of irrelevant and noisy features without predictive power were created and only a small percentage of features are discriminative.

MNIST is a subset of the distinguished handwritten digit recognition database [4]. The digits with a total of 4,000 images were size-normalized and centered in a fixed-size gray level of 28×28 pixels. This database contains 10 classes with varying scenes of skewing, noisy and blurring images.

USPS is a popular subset consisting of 9,298 handwritten digit images in total, each corresponding to a gray-scale level of 16×16 pixels which constructs a 256-dimensional feature vector [5]. Due to low resolutions, it is a widely used challenging image recognition database as well.

CLLSUB is a microarray data set to identify molecular correlates of genetically and clinically distinct subgroups of B-cell chronic lymphocytic leukemia [6]. Gene expression profiling is used to gain frequent genomic aberrations for 11,340 features in total.

CNAE is a database containing 1,080 free text documents from business descriptions of Brazilian companies [7]. Each document was represented by an 856-dimensional feature vector, where each feature value was the frequency of one word occurred in the document.

TDT2 is a subset of the well-known TDT corpus comprising 2 newswires, 2 radio programs and 2 television programs [8]. In this data set, only the largest 30 categories were preserved, a total of 9,394 documents. Each document was represented as a feature vector, corresponding to the frequencies of terms.

With the aforementioned detailed characterization to all test data sets, a summarization is provided as well, outlined in Table 2.

4.2. Parameter settings

In order to validate the effectiveness and efficiency of the proposed method DRWDR, a number of state-of-the-art data representation methods are compared. The compared methods include baseline method, principal component analysis (PCA), locally linear embedding (LLE) [33], locality preserving projection (LPP) [14], neighborhood preserving embedding (NPE) [13], Laplacian Eigenmaps (LE) [2], auto-encoder (AE) [15], isometric feature mapping (ISOMAP) [36], discriminative unsupervised dimensionality reduction (DUDR) [44], robust structured subspace learning (RSSL) [22], and sparse multi-graph embedding (SMGE) [41].

[1] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html
[2] https://fei.edu.br/ cet/facedatabase.html
[3] http://archive.ics.uci.edu/ml/datasets/Gisette
[4] http://yann.lecun.com/exdb/mnist/
[5] http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html
[6] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = GSE2466
[7] http://archive.ics.uci.edu/ml/datasets/CNAE-9
[8] http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html

Table 2
A brief description of the test data sets.

ID	Data sets	# Samples	# Features	# Classes	Data types
1	ORL	400	4,096	40	Face image
2	FEI	2,800	14,400	200	Face image
3	GISETTE	7,000	5,000	2	Digit image
4	MNIST	4,000	784	10	Digit image
5	USPS	9,298	256	10	Digit image
6	CLLSUB	111	11,340	3	Gene expression
7	CNAE	1,080	856	9	Text document
8	TDT2	9,394	36,771	30	Text document

For all compared methods, their default settings are preserved if feasible. As an example, the numbers of the nearest neighbors for LLE, LPP and NPE are fixed as 5, and the sample affinity matrices are evaluated by Gaussian kernel. It is noted that the number of the nearest neighbors will be modified as a large value when the corresponding method comes to badly singular matrices. For RSSL, the initial label prediction matrix $\mathbf{F}$, linear transformation matrix $\mathbf{Q}$ and projection matrix $\mathbf{P}$ are generated randomly. For DUDR, the predefined number of clusters is set as that of classes. Since RSSL and DUDR are quite time-consuming, the top 4,000 features are selected for algorithm comparisons on the large-scale data sets FEI and TDT2. As to the proposed method DRWDR, the regularization parameter is tuned as $\alpha = 0.1$.

In order to provide a fair comparison for all tested data representation methods, aside from the aforementioned default settings, feature numbers range in $\{10, 15, \cdots, 100\}$. The low-dimensional data representation by any compared method is evaluated using its K-means clustering performance, in which three evaluation metrics, including clustering accuracy (ACC), normalized mutual information (NMI) and adjusted rand index (ARI), are jointly employed for a comprehensive assessment. Because the K-means clustering is sensitive to initial inputs, it performs repeatedly 20 times, and we report their mean values and standard deviations.

4.3. Experimental results

The best clustering performance of all compared data representation approaches is reported in this subsection, where Tables 3–5 are presented for ACC, NMI and ARI, respectively. From these three tables, we can draw the following observations. In the first place, data representation methods are efficient in the performance boost of learning tasks. Overall, an encouraging low-dimensional representation achieves a significant increase in terms of the clustering performance, which can be observed from almost entire test data sets. Compared with the baseline method, there is a rise of nearly 50% with regard to clustering accuracy and adjusted rand index for quite a few data sets (e.g. FEI, CNAE and TDT2). There is one more point, the proposed method DRWDR gains the best clustering performance in the test data sets. In addition, it is superior to the second best method by a great deal in some certain data sets (e.g. GISETTE and CNAE). One potential reason is that the proposed method DRWDR learns a discriminative compact low-dimensional representation by the unitary invariance minimization. Simultaneously, deep random walk evaluates a more accurate sample affinity matrix. The last but not the least, varying data representation methods come with respective strengths in different data sets. As an example, DUDR yields the best clustering performance in the data set CLLSUB, though it comes with undesired performance in most of other data sets. Besides, it is suggested that SMGE attains promising clustering performance in a slice of data sets (e.g. ORL, USPS and TDT2) while ISOMAP works favorably in the data sets GISETTE and USPS.

The time complexities and actual running times of all compared data representation methods are shown in Tables 6,7. It is noted that PCA, LLE, LPP and NPE are also solved with closed-form solutions. Overall, the data representation methods having closed-form solutions come with less running times on the test data sets. Certainly, the proposed method DRWDR performs with an acceptable running time since it is provided with an almost closed-form solution. On the contrary, AE, DUDR, and RSSL consume more running times to obtain optimal solutions. This may account for the reason that matrix closed-form solutions are easy to be optimized and accelerated with some parallel computing techniques.

4.4. Parameter sensitivities

The proposed data representation method DRWDR is composed of several algorithmic parameters, including feature numbers of the reduced subspace, step numbers of the deep random walk, decay factor γ, Schatten norm indicators p and q, and regularization coefficient α. In this subsection, it is examined whether DRWDR is sensitive to these hyper-parameters.

The learning performance of the proposed method DRWDR is reported in Figs. 2–4 for ACC, NMI and ARI of entire data sets. In the three figures, algorithmic parameters are constrained to feature number $k \in \{10, 20, \cdots, 90\}$ and step number $t \in \{1, 2, \cdots, 9, \infty\}$, while fixing regularization coefficient $\alpha = 0.1$, decay factor $\gamma = 0.8$ and Schatten norm indicator $p = 5$ and $q = 2$. Here, ∞ represents the infinite number of probability transition steps. From these three figures, we can draw a number of beneficial observations. The first and foremost, a large step number significantly improves the clustering performance of DRWDR, which can be observed from the fact that DRWDR comes with pleasurable results with $t \geqslant 2$. In contrast,

Table 3
Clustering accuracy of compared data representation methods (mean% ± std%). The best results are marked in bold.

Method/ Data	ORL	FEI	GISETTE	MNIST	USPS	CLLSUB	CNAE	TDT2
Baseline	49.2 ± 2.5	31.2 ± 1.1	70.4 ± 0.2	51.7 ± 3.8	67.4 ± 6.7	44.5 ± 9.0	40.9 ± 7.2	33.0 ± 1.0
PCA	51.7 ± 2.5	36.4 ± 0.5	57.4 ± 0.1	52.1 ± 2.1	68.9 ± 5.8	48.1 ± 8.8	48.0 ± 7.4	48.1 ± 4.6
LLE	60.0 ± 3.7	36.9 ± 0.5	72.1 ± 9.9	38.3 ± 6.5	55.4 ± 6.9	46.9 ± 0.6	14.7 ± 1.1	19.9 ± 0.1
LPP	48.9 ± 1.8	49.1 ± 0.6	72.4 ± 9.8	51.9 ± 3.7	66.9 ± 3.3	47.0 ± 0.6	51.2 ± 3.6	20.7 ± 0.1
NPE	61.5 ± 1.4	44.5 ± 0.6	62.0 ± 9.6	50.9 ± 2.3	65.5 ± 1.1	51.4 ± 2.0	50.9 ± 6.2	26.0 ± 5.1
LE	53.3 ± 1.7	42.1 ± 0.5	71.9 ± 6.9	56.6 ± 3.7	66.7 ± 7.1	47.7 ± 2.8	55.9 ± 5.8	51.9 ± 3.1
AE	45.9 ± 2.0	31.3 ± 0.6	71.7 ± 6.6	53.4 ± 1.0	69.4 ± 2.9	47.8 ± 0.0	34.6 ± 1.0	50.9 ± 0.6
ISOMAP	49.3 ± 2.2	32.5 ± 0.5	82.5 ± 0.2	54.8 ± 3.6	69.0 ± 4.0	52.3 ± 0.0	53.2 ± 5.6	33.5 ± 1.6
DUDR	54.3 ± 0.0	43.3 ± 0.0	78.5 ± 0.0	26.4 ± 0.6	37.5 ± 1.4	**55.0 ± 0.0**	37.8 ± 2.2	42.0 ± 4.3
RSSL	54.1 ± 2.1	34.0 ± 0.4	68.8 ± 9.2	55.0 ± 5.9	63.7 ± 1.8	42.0 ± 1.7	51.8 ± 7.3	41.0 ± 4.0
SMGE	63.5 ± 1.9	36.6 ± 0.8	70.8 ± 9.1	59.8 ± 5.2	68.2 ± 6.8	49.2 ± 7.7	52.8 ± 9.4	61.3 ± 9.8
DRWDR	**64.3 ± 1.9**	**49.8 ± 0.7**	**86.2 ± 9.9**	**64.2 ± 3.8**	**69.6 ± 6.0**	52.6 ± 0.4	**58.6 ± 6.0**	**65.2 ± 2.4**

Table 4
Normalized mutual information of compared data representation methods (mean% ± std%). The best results are marked in bold.

Method/ Data	ORL	FEI	GISETTE	MNIST	USPS	CLLSUB	CNAE	TDT2
Baseline	67.0 ± 1.5	68.9 ± 0.6	13.3 ± 0.2	45.6 ± 2.4	65.4 ± 2.1	11.8 ± 5.9	32.9 ± 6.3	30.2 ± 1.6
PCA	71.5 ± 0.9	71.1 ± 0.4	1.57 ± 0.1	48.3 ± 1.6	65.9 ± 1.7	16.0 ± 9.1	42.1 ± 4.5	63.0 ± 2.0
LLE	78.8 ± 1.7	67.8 ± 0.7	23.5 ± 2.0	42.2 ± 6.5	49.5 ± 7.8	8.64 ± 4.8	13.9 ± 1.1	10.0 ± 1.1
LPP	68.8 ± 1.1	77.5 ± 0.4	26.4 ± 9.5	46.8 ± 2.0	64.4 ± 1.9	2.30 ± 0.7	48.2 ± 4.5	12.1 ± 1.0
NPE	78.1 ± 1.1	74.1 ± 0.3	22.9 ± 9.0	43.1 ± 2.4	65.2 ± 1.5	16.1 ± 4.1	46.2 ± 4.6	18.6 ± 4.3
LE	72.5 ± 1.2	73.0 ± 0.2	25.4 ± 9.5	56.4 ± 1.9	67.2 ± 4.9	4.39 ± 0.4	50.1 ± 3.7	67.4 ± 0.7
AE	68.4 ± 1.2	68.2 ± 0.3	16.4 ± 5.7	45.4 ± 1.1	66.6 ± 1.7	1.81 ± 0.0	27.1 ± 1.5	44.2 ± 0.5
ISOMAP	70.6 ± 0.7	67.8 ± 0.4	37.9 ± 0.4	50.6 ± 2.0	73.1 ± 1.7	18.0 ± 0.0	48.4 ± 3.8	34.9 ± 1.2
DUDR	70.0 ± 0.0	60.9 ± 0.0	24.8 ± 0.0	20.8 ± 0.2	40.7 ± 0.1	**26.3 ± 0.0**	23.2 ± 1.6	36.8 ± 2.7
RSSL	70.7 ± 1.8	69.5 ± 0.4	19.7 ± 9.9	58.5 ± 5.8	57.5 ± 0.7	2.47 ± 0.2	48.9 ± 7.8	36.3 ± 3.6
SMGE	70.6 ± 0.7	68.7 ± 0.6	25.0 ± 9.5	62.3 ± 2.7	68.6 ± 2.5	12.3 ± 5.0	53.6 ± 8.7	51.7 ± 9.6
DRWDR	**80.4 ± 1.0**	**77.0 ± 0.3**	**53.4 ± 9.4**	**63.4 ± 2.4**	**81.9 ± 3.2**	16.2 ± 1.5	**54.4 ± 4.7**	**65.8 ± 1.1**

Table 5
Adjusted rand index of compared data representation methods (mean% ± std%). The best results are marked in bold.

Method/ Data	ORL	FEI	GISETTE	MNIST	USPS	CLLSUB	CNAE	TDT2
Baseline	31.8 ± 2.7	20.6 ± 0.9	16.6 ± 0.3	33.1 ± 3.5	57.1 ± 3.5	3.25 ± 0.6	19.0 ± 4.7	2.11 ± 1.0
PCA	36.0 ± 1.4	23.6 ± 0.6	2.15 ± 0.1	35.2 ± 2.0	57.7 ± 2.9	6.06 ± 0.7	26.4 ± 5.9	34.9 ± 5.5
LLE	48.4 ± 3.7	16.7 ± 1.0	25.0 ± 9.9	27.3 ± 8.6	44.2 ± 8.3	3.01 ± 0.7	1.20 ± 0.1	5.30 ± 0.1
LPP	31.1 ± 1.7	33.9 ± 1.2	26.6 ± 9.4	33.9 ± 2.6	53.2 ± 3.2	1.10 ± 0.1	28.1 ± 6.0	6.00 ± 1.0
NPE	46.8 ± 1.7	25.9 ± 0.9	21.9 ± 9.3	29.5 ± 1.7	56.1 ± 1.9	7.67 ± 0.4	24.1 ± 5.6	2.96 ± 0.4
LE	38.2 ± 2.8	27.4 ± 0.4	21.0 ± 9.3	42.6 ± 1.9	58.2 ± 9.1	2.85 ± 0.5	31.4 ± 5.2	40.6 ± 2.9
AE	30.7 ± 2.4	18.2 ± 1.3	20.3 ± 7.1	33.5 ± 1.0	50.6 ± 5.5	1.19 ± 0.0	18.2 ± 7.2	36.3 ± 0.5
ISOMAP	34.3 ± 1.5	18.3 ± 0.8	42.2 ± 0.5	37.8 ± 2.8	65.7 ± 3.1	8.58 ± 0.0	30.6 ± 2.8	16.4 ± 1.0
DUDR	19.4 ± 0.0	29.7 ± 0.0	32.4 ± 0.0	18.4 ± 0.1	25.1 ± 1.3	**12.4 ± 0.0**	14.1 ± 1.0	13.4 ± 2.2
RSSL	28.5 ± 4.5	20.9 ± 0.4	16.1 ± 3.5	41.8 ± 4.8	49.6 ± 1.1	3.66 ± 0.4	25.7 ± 5.5	13.4 ± 2.1
SMGE	44.4 ± 2.6	20.3 ± 0.9	21.0 ± 9.3	47.5 ± 4.8	58.1 ± 5.9	4.48 ± 0.7	30.1 ± 2.4	42.1 ± 9.4
DRWDR	**47.5 ± 3.7**	**34.4 ± 1.0**	**58.3 ± 9.6**	**50.0 ± 4.0**	**67.7 ± 6.3**	8.29 ± 0.3	**34.9 ± 6.2**	**53.7 ± 2.5**

Table 6
Computational complexities of compared data representation methods.

Method	Baseline	PCA	LLE	LPP	NPE	LE
Complexity	$O(nd)$	$O(n^2(n+d))$	$O(n^3)$	$O\left(d^2(n+d)\right)$	$O\left(n^3+d^3\right)$	$O(n^3)$
Method	AE	ISOMAP	DUDR	RSSL	SMGE	DRWDR
Complexity	$O(nd)$	$O(n^2d)$	$O\left(d^2+n^2\right)$	$O\left(d^2(n+d)\right)$	$O(n^2(n+d))$	$O(n^2(n+d))$

DRWDR works unacceptably by setting $t = 1$, which may be indicative of how to accurately evaluate an affinity matrix of given data points. There is one more point, excessively large step number may leave a negative effect on the clustering performance, which can be validated by the fact that the step number $t = \infty$ gives rise to an undesired performance in the data sets GISETTE, MNIST, USPS and CNAE. The last but not the least, the proposed method DRWDR is inclined to attaining the

Table 7
Running times of all compared data representation methods (in seconds).

Method/ Data	ORL	FEI	GISETTE	MNIST	USPS	CLLSUB	CNAE	TDT2
Baseline	0.21	24.23	7.79	0.61	1.17	0.05	0.17	1.47
PCA	1.10	4.53	7.21	0.07	0.05	0.01	0.14	51.63
LLE	336.91	8.45	125.25	24.21	212.51	0.21	56.02	255.69
LPP	22.31	18.16	130.57	0.44	7.28	0.02	0.25	383.40
NPE	282.73	57.96	29.05	1.79	34.10	0.10	0.61	349.37
LE	1.55	4.64	40.21	6.00	18.99	0.22	0.54	53.51
AE	239.37	989.31	373.72	104.62	47.45	260.03	23.09	6040.78
ISOMAP	0.31	23.28	301.66	56.39	534.66	0.04	2.14	574.64
DUDR	327.53	700.39	6605.70	49.27	372.3	3844.30	8.00	8026.60
RSSL	1284.88	2523.43	2743.2	62.88	161.32	1114.42	25.77	2682.07
SMGR	18.10	124.30	316.43	173.38	404.68	5.28	47.68	440.89
DRWDR	0.09	7.62	120.38	22.31	282.69	0.03	0.42	292.27

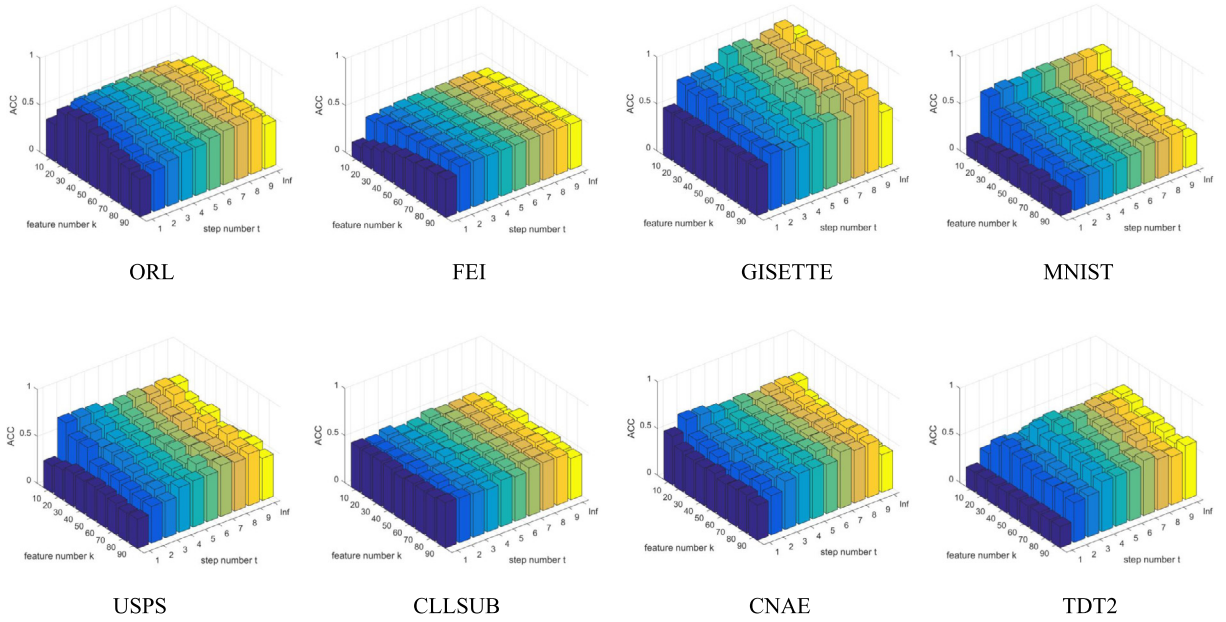

Fig. 2. The relations among clustering accuracy, feature number ranging in $\{10, 20, \cdots, 90\}$ and step number in $\{1, 2, \cdots, 9, \infty\}$ of the proposed method DRWDR.

best clustering performance with a smaller number of dimensions of the learned feature subspace. This is owing to the observation that DRWDR works best in the data sets MNIST, USPS and CNAE while the feature number is tuned as $k = 10$.

The clustering performance of the proposed method DRWDR is presented in Fig. 5 while the decay factor γ ranges in $\{0.05, 0.1, \cdots, 0.9, 0.95\}$ with keeping regularization coefficient $\alpha = 0.1$ and step number $t = 5$. For convenience, the figure is a collection of the best performance as the number of reduced features varies in $\{5, 10, \cdots, 95, 100\}$. It is observed from this figure that DRWDR behaves unfavorably with $0 < \gamma < 0.5$. In another word, a small value of decay factor γ leads to a trivial performance, which is straightforward to be explained in that γ is indicative of the tendency of a promotional message to lose its effectiveness over time, and a small γ suggests a low preservation of current indispensable information. Overall, DRWDR results in a robust acceptable clustering performance as parameter γ ranges in $\{0.5, 0.55, \cdots, 0.9, 0.95\}$.

As the feature number fluctuates in $\{5, 10, \cdots, 95, 100\}$, the relation between clustering performance of the proposed method DRWDR and Schatten norm indicator $p \in \{1, 2, \cdots, 9\}$ is illustrated in Fig. 6, while fixing $q = 2$, decay factor $\gamma = 0.8$, and regularization parameter $\alpha = 0.1$. It is observed from this figure that DRWDR brings about varying clustering performance with different values of p. As an example, DRWDR works best with $p = 6$ in the data set GISETTE and with $p = 8$ in USPS. But overall, it performs favorably in the selected wide range, which is suggestive of the robustness of the proposed method in some sense.

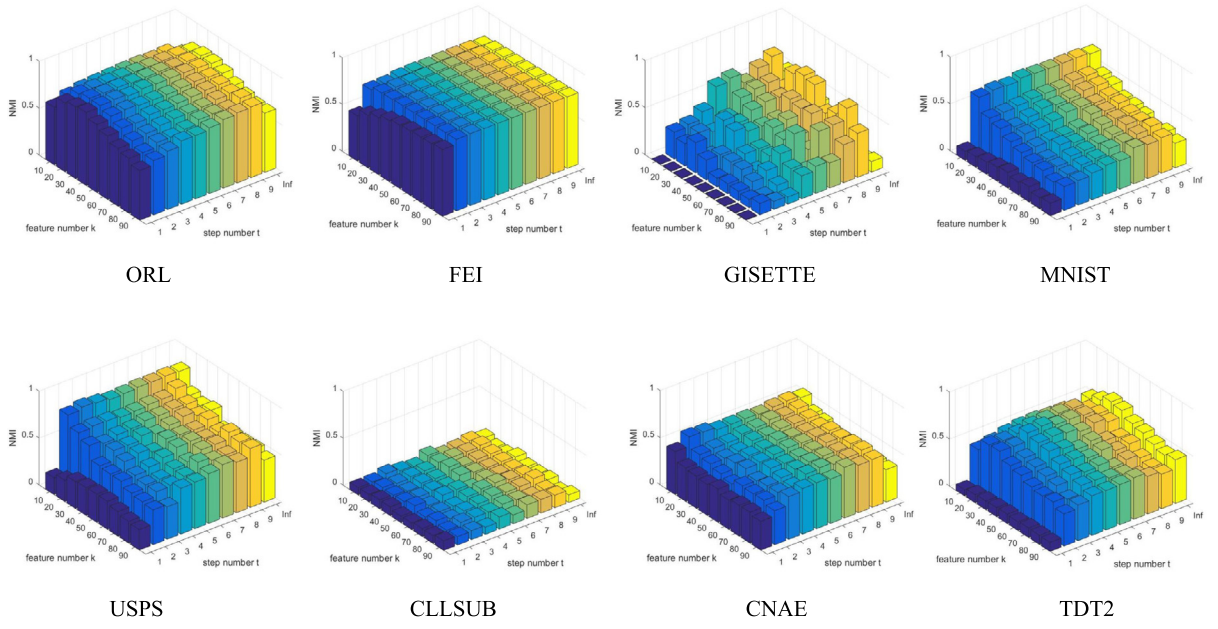

Fig. 3. The relations among normalized mutual information, feature number ranging in $\{10, 20, \cdots, 90\}$ and step number in $\{1, 2, \cdots, 9, \infty\}$ of the proposed method DRWDR.

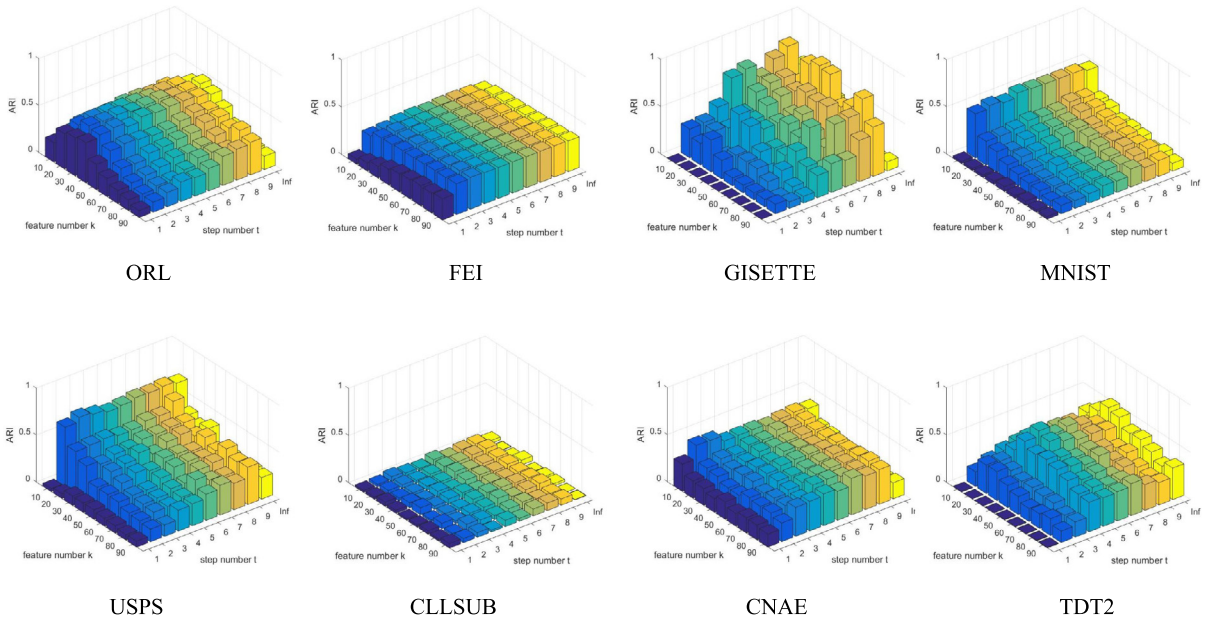

Fig. 4. Relations among adjusted rand index, feature number in $\{10, 20, \cdots, 90\}$ and step number in $\{1, 2, \cdots, 9, \infty\}$ of the proposed method DRWDR.

5. Conclusion and future work

In this paper, we proposed a matrix factorization method based on deep random walk for data representation problems. In the first place, unitarily invariant norm minimization was embedded into the formulated problem framework, which was suggestive of how to learn an adaptive distance metric. There was one more point that the deep random walk was used to

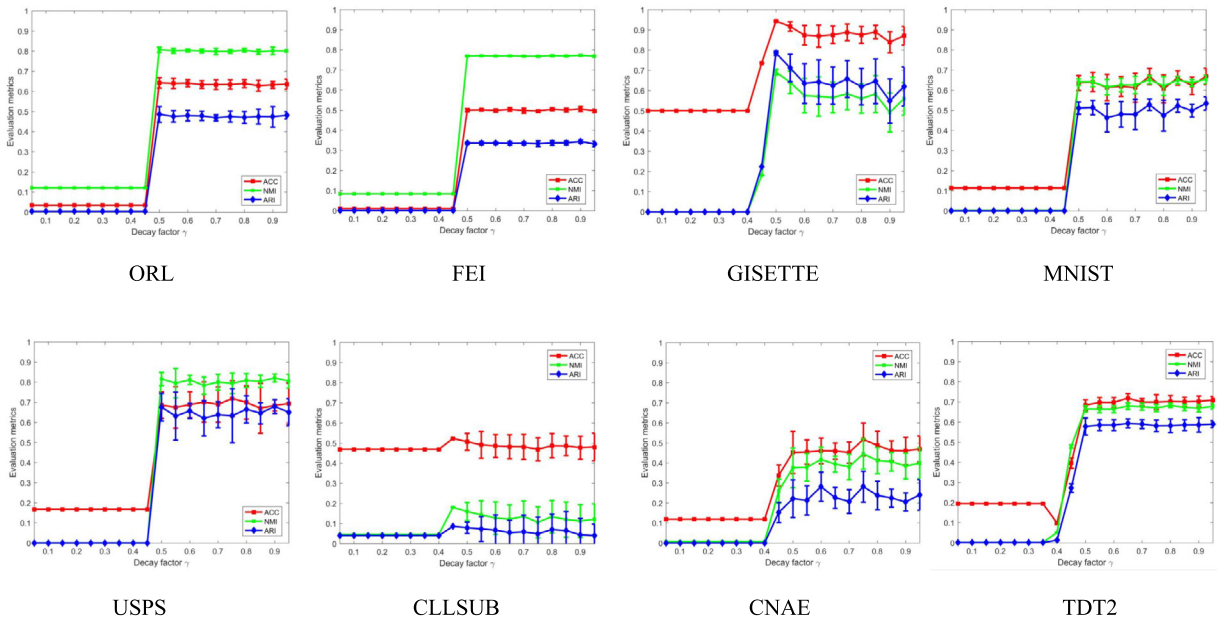

Fig. 5. The connection between clustering performance and decay factor γ in $\{0.05, 0.1, \cdots, 0.90, 0.95\}$ of the proposed method DRWDR.

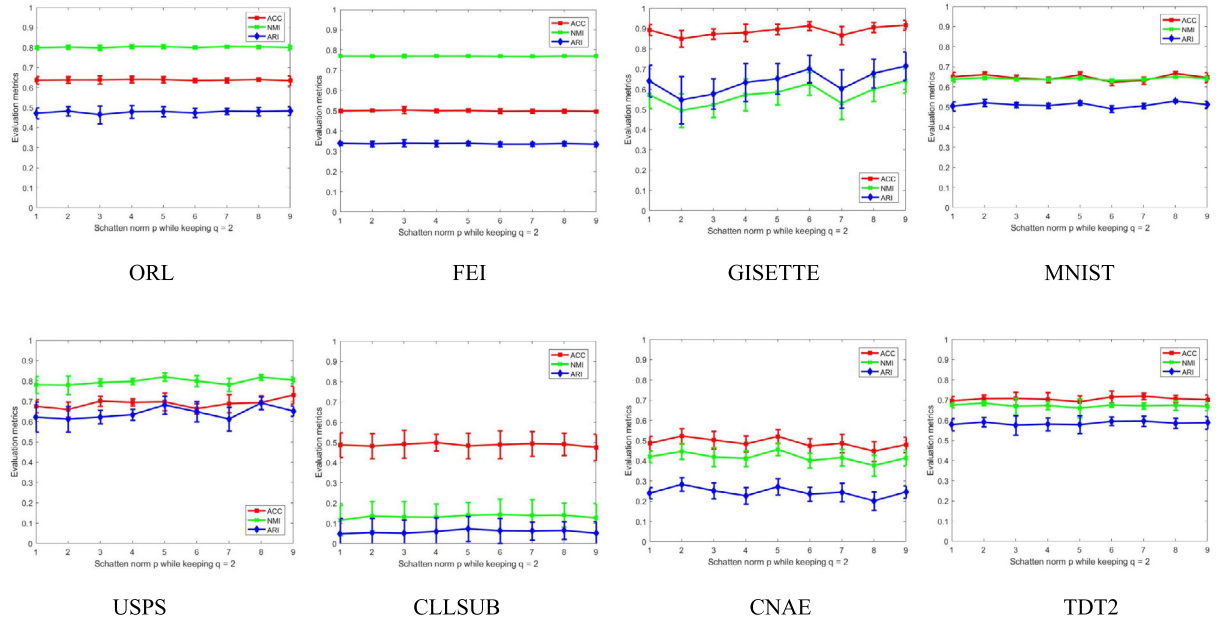

Fig. 6. The connection between clustering performance and Schatten norm p in $\{1, 2, \cdots, 9\}$ of the proposed method DRWDR while fixing $q = 2$.

evaluate a more accurate affinity matrix and learn a nonlinear multi-scale data representation. The last but not the least, a new optimization algorithm with guaranteed convergence was put forward. The proposed method was solved by a nearly closed-form solution and succeeded in addressing data representation problems by capturing multi-view distributions hidden in data. Moreover, the model was applied to clustering tasks, and experimental results demonstrated that it was superior to both traditional and state-of-the-art approaches. Nevertheless, a major problem of the proposed method to be further improved is that the singular value decomposition requires relatively intensive computational complexity, which is the most time-consuming module when handling large-scale data representation problems.

In the future work, we will view the unitary invariance as a surrogate of adaptive distance metrics to learn a compact data representation. On the other hand, we will attempt to find more efficient alternatives to decrease the computational complexity of the proposed method. As an example, constructing differentiable blocks for the optimization variables makes them learnable in a neural network. Besides, we will further exploit the relationship between unitarily invariant norm minimization and sparse regularizer embedding.

CRediT authorship contribution statement

Shiping Wang: Conceptualization, Formal analysis, Methodology, Writing - original draft. **Zhaoliang Chen:** Conceptualization, Formal analysis, Methodology, Writing - review & editing. **William Zhu:** Supervision, Validation, Visualization. **Fei-Yue Wang:** Funding acquisition, Writing - review & editing.

Declaration of Competing Interest

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

Acknowledgments

This work is partly supported by the National Natural Science Foundation of China under Grant Nos. U1705262 and 61672159, the Fujian Collaborative Innovation Center for Big Data Application in Governments, and the Technology Innovation Platform Project of Fujian Province under Grant No. 2014H2005.

References

[1] S. Ayas, M. Ekinci, Single image super resolution using dictionary learning and sparse coding with multi-scale and multi-directional gabor feature representation, Inf. Sci. 512 (2020) 1264–1278.
[2] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Advances in Neural Information Processing Systems, 2002, pp. 585–591.
[3] C. Chen, Z. Chen, B. Jiang, X. Jin, Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation, in: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, 2019, pp. 3296–3303.
[4] P.-H. Chen, H.-C. Yang, K.-W. Chen, Y.-S. Chen, Mvsnet++: learning depth-based attention pyramid features for multi-view stereo, IEEE Trans. Image Process. 29 (2020) 7261–7273.
[5] X. Dai, X. Su, W. Zhang, F. Xue, H. Li, Robust manhattan non-negative matrix factorization for image recovery and representation, Inf. Sci. 527 (2020) 70–87.
[6] J.A. Duersch, M. Gu, Randomized projection for rank-revealing matrix factorizations and low-rank approximations, SIAM Rev. 62 (2020) 661–682.
[7] M. Federici, A. Dutta, P. Forré, N. Kushman, Z. Akata, Learning robust representations via multi-view information bottleneck, in: Proceedings of the International Conference on Learning Representations, 2020, pp. 1–26.
[8] I. Feige, Invariant-equivariant representation learning for multi-class data, in: Proceedings of the Thirty-Sixth International Conference on Machine Learning, 2019, pp. 1882–1891.
[9] X. Feng, S. Wu, Robust sparse coding via self-paced learning for data representation, Inf. Sci. 546 (2020) 448–468.
[10] C. Gong, H. Shi, J. Yang, J. Yang, Multi-manifold positive and unlabeled learning for visual analysis, IEEE Trans. Circuits Syst. Video Technol. 30 (2020) 1396–1409.
[11] L. Guo, X. Cai, H. Qin, Y. Guo, F. Li, G. Tian, Citation recommendation with a content-sensitive deepwalk based approach, in: Proceedings of the Nineteenth IEEE International Conference on Data Mining Workshop, 2019, pp. 538–543.
[12] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, Neural Comput. 16 (2004) 2639–2664.
[13] X. He, D. Cai, S. Yan, H.-J. Zhang, Neighborhood preserving embedding, in: IEEE International Conference on Computer Vision, 2005, pp. 1208–1213.
[14] X. He, P. Niyogi, Locality preserving projections. In Advances in Neural Information Processing Systems, 2004. pp. 153–160..
[15] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (2006) 504–507.
[16] C.-C. Hsu, C.-W. Lin, Cnn-based joint clustering and representation learning with feature drift compensation for large-scale image data, IEEE Trans. Multimedia 20 (2018) 421–429.
[17] T. Huynh-The, C.-H. Hua, T.-T. Ngo, D.-S. Kim, Image representation of pose-transition feature for 3d skeleton-based action recognition, Inf. Sci. 513 (2020) 112–126.
[18] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, S. Ermon, Tile2vec: Unsupervised representation learning for spatially distributed data, Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 3967–3974.
[19] S. Jin, S. Wiseman, K. Stratos, K. Livescu, Discrete latent variable representations for low-resource text classification, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4831–4842.
[20] P. Jing, Y. Su, Z. Li, L. Nie, Learning robust affinity graph representation for multi-view clustering, Inf. Sci. 544 (2020) 155–167.
[21] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.
[22] Z. Li, J. Liu, J. Tang, H. Lu, Robust structured subspace learning for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 2085–2098.
[23] Z. Li, J. Liu, Y. Yang, X. Zhou, H. Lu, Clustering-guided sparse structural learning for unsupervised feature selection, IEEE Trans. Knowl. Data Eng. 26 (2014) 2138–2150.
[24] Z. Li, J. Tang, T. Mei, Deep collaborative embedding for social image understanding, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 2070–2083.
[25] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, S. Yan, Tensor robust principal component analysis with a new tensor nuclear norm, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020) 925–938.
[26] Y. Lu, C. Yuan, Z. Lai, X. Li, D. Zhang, W.K. Wong, Horizontal and vertical nuclear norm-based 2dlda for image representation, IEEE Trans. Circuits Syst. Video Technol. 29 (2019) 941–955.
[27] Y. Meng, R. Shang, F. Shang, L. Jiao, S. Yang, R. Stolkin, Semi-supervised graph regularized deep nmf with bi-orthogonal constraints for data representation, IEEE Trans. Neural Networks Learn. Syst. 31 (2020) 3245–3258.
[28] T. Mikolov, I. Sutskever, K. Chen, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.

[29] F. Nie, J. Yuan, H. Huang, Optimal mean robust principal component analysis, in: Proceedings of the 31st International Conference on Machine Learning, 2014, pp. 1062–1070.
[30] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238.
[31] P. Perera, V.I. Morariu, R. Jain, V. Manjunatha, C. Wigington, V. Ordonez, V.M. Patel, Generative-discriminative feature representations for open-set recognition, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition, 2020, pp. 11811–11820.
[32] B. Perozzi, R. Al-Rfou, S. Skiena, Max-margin deepwalk: Discriminative learning of network representation, in: Proceedings of the Twentieth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 701–710.
[33] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2002) 2323–2326.
[34] Z. Shu, X.-J. Wu, C.-Z. You, Z. Liu, P. Li, H. Fan, F. Ye, Rank-constrained nonnegative matrix factorization for data representation, Inf. Sci. 528 (2020) 133–146.
[35] R.S. Sutton, A.G. Barto, Reinforcement learning: An introduction, IEEE Trans. Neural Networks 9 (1998), 1054–1054.
[36] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[37] C. Tu, W. Zhang, Z. Liu, M. Sun, Max-margin deepwalk: Discriminative learning of network representation, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016, pp. 3889–3895.
[38] J.X. Wang, Z. Kurth-Nelson, D. Kumaran, D. Tirumala, Prefrontal cortex as a meta-reinforcement learning system, Nat. Neurosci. 21 (2018) 860–868.
[39] Q. Wang, J. Lai, L. Claesen, Z. Yang, L. Lei, W. Liu, A novel feature representation: Aggregating convolution kernels for image retrieval, Neural Networks 130 (2020) 1–10.
[40] S. Wang, Z. Ding, Y. Fu, Discerning feature supported encoder for image representation, IEEE Trans. Image Process. 28 (2019) 3728–3738.
[41] S. Wang, W. Guo, Sparse multigraph embedding for multimodal feature representation, IEEE Trans. Multimedia 19 (2017) 1454–1466.
[42] S. Wang, W. Pedrycz, Q. Zhu, W. Zhu, Subspace learning for unsupervised feature selection via matrix factorization, Pattern Recogn. 48 (2015) 10–19.
[43] S. Wang, W. Zhu, Sparse graph embedding unsupervised feature selection, IEEE Trans. Syst., Man, Cybernetics: Syst. 48 (2018) 329–341.
[44] X. Wang, Y. Liu, F. Nie, H. Huang, Discriminative unsupervised dimensionality reduction, in: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015, pp. 3925–3931.
[45] Y. Xiao, R. Li, X. Lu, Y. Liu, Link prediction based on feature representation and fusion, Inf. Sci. 548 (2020) 1–17.
[46] M. Xu, Z. Zhu, X. Zhang, Y. Zhao, X. Li, Canonical correlation analysis with $\ell_{2,1}$-norm for multiview data representation, IEEE Trans. Cybern (2019) 1–11.
[47] W. Yan, Q. Sun, H. Sun, Y. Li, Joint dimensionality reduction and metric learning for image set classification, Inf. Sci. 516 (2020) 109–124.
[48] C. Yang, Z. Liu, D. Zhao, M. Sun, E.Y. Chang, Network representation learning with rich text information, in: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015, pp. 2111–2117.
[49] J. Yang, A.F. Frangi, J.-Y. Yang, D. Zhang, Z. Jin, Kpca plus lda: a complete kernel fisher discriminant framework for feature extraction and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 230–244.
[50] Z. Yang, Q. Li, L. Wenyin, J. Lv, Shared multi-view data representation for multi-domain event detection, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020) 1243–1256.
[51] Y.-L. Yu, D. Schuurmans, Rank/norm regularization with closed-form solutions: Application to subspace clustering, in: Proceedings of International Joint Conference on Artificial Intelligence, 2011, pp. 778–785.
[52] X. Yuan, B. Huang, Y. Wang, C. Yang, W. Gui, Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted sae, IEEE Trans. Industr. Inf. 14 (2018) 3235–3243.
[53] X. Zhan, Matrix theory, American Mathematical Society, Providence, 2013.
[54] Q. Zhu, N. Xu, S. Huang, J. Qian, D. Zhang, Adaptive feature weighting for robust ℓ_p-norm sparse representation with application to biometric image classification, Int. J. Mach. Learn. Cybern. 11 (2020) 463–474.