Robust Self-Weighted Multi-View Projection Clustering

Beilei Wang,[1] Yun Xiao,[1*] Zhihui Li,[2,3†] Xuanhong Wang,[4] Xiaojiang Chen,[1] Dingyi Fang[1]
[1]School of Information Science and Technology, Northwest University, Xi'an 710127, P.R. China
[2]School of Information Science and Engineering, Shandong Normal University, Jinan 250358, P.R. China
[3]School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia
[4]School of Communications and Information Engineering,
Xi'an University of Posts & Telecommunications, Xi'an 710121, P.R. China
wangbeilei@stumail.nwu.edu.cn, {yxiao, xjchen, dyf}@nwu.edu.cn, zhihuilics@gmail.com, wxh@xupt.edu.cn

Abstract

Many real-world applications involve data collected from different views and with high data dimensionality. Furthermore, multi-view data always has unavoidable noise. Clustering on this kind of high-dimensional and noisy multi-view data remains a challenge due to the curse of dimensionality and ineffective de-noising and integration of multiple views. Aiming at this problem, in this paper, we propose a Robust Self-weighted Multi-view Projection Clustering (RSwMPC) based on $\ell_{2,1}$-norm, which can simultaneously reduce dimensionality, suppress noise and learn local structure graph. Then the obtained optimal graph can be directly used for clustering while no further processing is required. In addition, a new method is introduced to automatically learn the optimal weight of each view with no need to generate additional parameters to adjust the weight. Extensive experimental results on different synthetic datasets and real-world datasets demonstrate that the proposed algorithm outperforms other state-of-the-art methods on clustering performance and robustness.

Introduction

In many real-world applications, instances can be described in many different ways or angles, which produce abundant multi-views data, such as web page classification problems, image processing problems, and the like. Clustering on multi-view data is a fundamental and important topic in data mining, machine learning, pattern recognition and so on. In this era of information explosion, the dimensionality of data is getting higher and higher. Meanwhile, multi-view data always has unavoidable noise. These directly lead to the increase of data storage cost, the further improvement of the complexity of learning algorithm and the decrease of algorithm generalization ability. Clustering on this kind of high-dimensional and noisy multi-view data remains a challenge due to the curse of dimensionality and ineffective de-noising and integration of multiple views.

In the past decades, a number of multi-view clustering approaches have been proposed. In order to maintain the same clustering consistency among all graphs, (Kumar, Rai, and Daume 2011; Cai et al. 2011) proposed co-regularized and multi-modal spectral clustering methods, which can not distinguish the importance of different views and were vulnerable to poor quality views. In order to distinguish the importance of different views for clustering results, (Nie et al. 2016; Nie, Cai, and Li 2017) proposed to automatically assign an ideal weight to each view without introducing redundant hyperparameters. (Nie, Tian, and Li 2018) considered the clustering ability difference of different views and proposed multiview clustering via adaptively weighted procrustes. The above clustering methods successfully solved the clustering problem in low-dimensional data. However, there are still significant challenges to high-dimensional data clustering problems with noise.

In order to process high-dimensional data, (Chen et al. 2018; Xiao et al. 2019) proposed to project the original matrix from the high-dimensional subspace to the low-dimensional subspace by learning the projection matrix, and obtained better clustering results on the single view. For multi-view high-dimensional data processing, (Gao et al. 2015) proposed multi-view subspace clustering (MVSC), which performs clustering on the subspace representation of each view, using a common indicator to ensure consistency between different views. (Wang et al. 2019) proposed two parameter-free weighted projected clustering methods, which can simultaneously perform structural graph learning and dimensionality reduction. However, for multi-view high-dimensional data with noise, these existing algorithms adopt dimensionality reduction method to project data from high-dimensional subspace to low-dimensional subspace, without considering the effect of noise in the datasets.

Since $\ell_{2,1}$-norm has advantages in noise processing and feature selection (Yang et al. 2011), in this paper, we combine it with dimensionality reduction method to solve the clustering problem on high dimensional and noisy multi-view datasets. We propose a Robust Self-weighted Multi-view Projection Clustering (RSwMPC) based on $\ell_{2,1}$-norm, which project the original dataset into the low-dimensional subspace through the projection matrix, and suppress the noise points and outliers through increase the $\ell_{2,1}$-norm penalty of the projection matrix. At the same time, in our proposed RSwMPC, the weight of each view depends on the projection matrix and the similarity matrix, and no redundant parameters are introduced. Different from the conventional post-processing methods (Tang, Lu, and Dhillon

[*] Corresponding author

[†] Corresponding author

2009; Kumar and Daumé 2011; Huang, Nie, and Huang 2013; Nie et al. 2016), which need similar k-means to obtain a cluster label, the Laplacian matrix rank constraint is also introduced, so that the learned affinity matrix has a display clustering structure, and the clustering result can be obtained directly. Unlike SwMPC (Wang et al. 2019), we suppress noise by increasing the $\ell_{2,1}$-norm penalty term of the projection matrix, and select the dimensionality corresponding to the best clustering result in each view as our projection dimensionality. Extensive experimental results on different synthetic datasets and real-world datasets demonstrate that the proposed algorithm outperforms other state-of-the-art methods on clustering performance and robustness.

The contributions of this paper are as follows:

- In our proposed RSwMPC, subspace learning is performed by adding a projection matrix. Meanwhile, feature selection and noise suppression are achieved by introducing the $\ell_{2,1}$-norm penalty term of the projection matrix. Therefore, RSwMPC is robust and effective.
- A parameter-free self-weighted strategy which combines the effective information of different views is proposed and used in the RSwMPC. In addition, the direct search method is adopted for dimensionality reduction, thus the optimal reduced dimensionality for each view is obtained.
- The clustering results can be obtained directly in the process of constructing affinity matrix S without further processing.
- We have verified the effectiveness and robustness of our method by conducting extensive experiments on synthetic datasets and real-world datasets.

The Proposed RSwMPC Methodology

Notation

Throughout the paper, all matrices are capitalized. For matrix $M \in \mathbb{R}^{d\times n}$, m_i represents the i-th column vector of the matrix M and its j-th element is represented by m_{ij}. In addition, we also define the rank, trace, transposition and inverse of the matrix M as $rank(M)$, $Tr(M)$, M^T and M^{-1}, respectively. The Frobenius norm and the $\ell_{2,1}$-norm of matrix M are $\|M\|_F$, $\|M\|_{2,1}$. $\|v\|_2$ represents the ℓ_2-norm of vector v. $\mathbf{1}$ and I represent a unit column vector and an identity matrix in the paper, respectively.

Multi-view Initial Affinity Matrix Learning

In multi-view clustering, denote $X_1, X_2, ..., X_v$ represent the data matrix for each view. $X_v \in \mathbb{R}^{d_v\times n}$ represents the v-th view, where n is the number of data points and d_v is the feature dimensionality. Most of the existing graph-based multi-view clustering algorithms need a pre-constructed graph, and the clustering performance of these algorithms depends on the quality of the constructed graph. One of the simple ways is that the features of all views are connected in series, and then the single-view clustering is performed based on the series-connected features. In this case, however, a view containing a larger amount of information is treated with other views that contain less information. Thus the final solution is not optimal. Therefore, we use a method of increasing weight to learn the similarity matrix. The optimization problem of learning unified similarity matrix can be described as follows (Nie, Cai, and Li 2017):

$$\begin{aligned} \min_{S,\alpha_v} \quad & \sum_v (\alpha_v \sum_{i,j=1}^{n} \|x_i^v - x_j^v\|_2^2 s_{ij} + \lambda\|\alpha_v\|_2^2) + \beta\|S\|_F^2, \\ s.t. \quad & s_i^T\mathbf{1} = 1, 0 \le s_{ij} \le 1, rank(L_S) = n - c, \\ & \alpha_v^T\mathbf{1} = 1, 0 \le \alpha_v \le 1. \end{aligned} \tag{1}$$

where $s_i \in \mathbb{R}^{n\times 1}$ is the i-th vector of similarity matrix $S \in \mathbb{R}^{n\times n}$ and its j-th element is s_{ij}. In spectral analysis, $L_S = D_S - (S^T + S)/2$ is a Laplace matrix, where the degree matrix D_S is the diagonal matrix about S whose i-th diagonal element d_{ii} is $\sum_j (s_{ij} + s_{ji})/2$. c is the number of 0 eigenvalues of L_S matrix. β is the tuning parameter, $\beta\|S\|_F^2$ is used to avoid assigning a similarity of 1 only to the nearest point of the data point x_i, while the similarity of the other points is 0. λ is the non-negative value, which is used to smooth the weight distribution. Because artificial weight is subjective, and assigning the same weight to each graph ignores the difference between different features, it is easy to be interfered by poor quality features, which leads to poor clustering accuracy.

Robust Self-weighted Multi-view Projection Clustering

Based on the superiority of subspace learning in the processing of high dimensional data, we introduce the projection matrix W into our method. Define the projection transformation matrix $W_v \in \mathbb{R}^{d_v\times d'_v}, d'_v \le d_v$, which projects the original dataset X of the v-th view into a low dimensional subspace. Here d'_v is the characteristic dimension of the low dimensional subspace. This low dimensional subspace can be represented as $W_v^T X_v$, which not only preserves the effective information of the data, but also alleviates the dimension disaster problem. In the following, we give the optimization goal based on subspace learning:

$$\begin{aligned} \min_{S,W_v,\alpha_v} \quad & \sum_v (\alpha_v \sum_{i,j} \|W_v^T x_i^v - W_v^T x_j^v\|_2^2 s_{ij} + \lambda\|\alpha_v\|_2^2 \\ & + \gamma\|W_v\|_{2,1}) + \beta\|S\|_F^2, \\ s.t. \quad & s_i^T\mathbf{1} = 1, 0 \le s_{ij} \le 1, rank(L_S) = n - c, \\ & W_v^T X_v X_v^T W_v = I, \alpha_v^T\mathbf{1} = 1, 0 \le \alpha_v \le 1. \end{aligned} \tag{2}$$

where γ is the tuning parameters. $\|W_v\|_{2,1}$ is the $\ell_{2,1}$-norm of projection matrix W, which is used to suppress noise and remove redundant features. The application of the orthogonal constraint to the scattering matrix $W_v^T X_v X_v^T W_v$ is actually used for intrinsic subspace learning, and the d_v-dimensionality feature space on the original dataset X_v is converted into a statistically non-relevant d'_v-dimensionality intrinsic subspace.

In order to solve the problem of weight distribution, we propose a Robust Self-weighted Multi-view Projection Clustering, which is as follows:

$$\begin{aligned} \min_{S,W_v} \quad & \sum_v ((\sum_{i,j} \|W_v^T x_i^v - W_v^T x_j^v\|_2^2 s_{ij})^{1/2} + \gamma \|W_v\|_{2,1}) \\ & + \beta \|S\|_F^2, \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1, rank(L_S) = n - c, \\ & W_v^T X_v X_v^T W_v = I. \end{aligned} \tag{3}$$

The Lagrange function of problem (3) can be written as

$$\begin{aligned} & \sum_v ((\sum_{i,j} \|W_v^T x_i^v - W_v^T x_j^v\|_2^2 s_{ij})^{1/2} + \gamma \|W_v\|_{2,1}) \\ & + \beta \|S\|_F^2 + \mathcal{G}(\Lambda_S, S) + \mathcal{G}(\Lambda_W, W_v), \end{aligned} \tag{4}$$

where Λ_S, Λ_W are the Lagrange multiplier, $\mathcal{G}(\Lambda_S, S)$, $\mathcal{G}(\Lambda_W, W_v)$ are the formalized term derived from constraints. Take the derivative of problem (4) with respect to S and make it equal to 0, we have

$$\begin{aligned} \sum_v \alpha_v \frac{\partial \sum_{i,j} \|W_v^T x_i^v - W_v^T x_j^v\|_2^2 s_{ij}}{\partial S} + \frac{\partial \beta \|S\|_F^2}{\partial S} \\ + \frac{\partial \mathcal{G}(\Lambda_S, S)}{\partial S} = 0, \end{aligned} \tag{5}$$

where

$$\alpha_v = \frac{1}{2(\sum_{i,j} \|W_v^T x_v^v - W_v^T x_j^v\|_2^2 s_{ij})^{1/2}} \tag{6}$$

Since α_v is dependent on the target variable S and W_v, so problem (5) cannot be directly solved. But if α_v is set to be stationary, problem (5) can be considered accounting for following problem:

$$\begin{aligned} \min_{S,W_v} \quad & \sum_v (\alpha_v \sum_{i,j} \|W_v^T x_i^v - W_v^T x_j^v\|_2^2 s_{ij} + \gamma \|W_v\|_{2,1}) \\ & + \beta \|S\|_F^2, \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1, rank(L_S) = n - c, \\ & W_v^T X_v X_v^T W_v = I. \end{aligned} \tag{7}$$

We update each W_v and S through problem (7). The weight α_v of each view depends on W_v and S, so α_v can update at the same time. This solves the problem of linearly combining different view weights without introducing extra parameters. That is, solving the problem (7) is equivalent to solving the problem (3). By solving the problem (7), the similarity matrix S can be learned directly for clustering.

Optimization Algorithm

For the solution of problem (7), we first optimize it by (Fan 1949), and iteratively update the algorithm by updating the parameters until converge.

Optimization Objective Function

In the problem (7), the i-th smallest eigenvalue of L_S is represented by $\sigma_i(L_S)$. Because the Laplace matrix L_S is positive semidefinite, $\sigma_i(L_S)$ is non-negative, that is $\sum_{i=1}^{c} \sigma_i(L_S) \ge 0$, which ensures the establishment of rank constraint $rank(L_S) = n - c$. Given a sufficiently large η, problem (7) can be written as

$$\begin{aligned} \min_{S,W_v} \quad & \sum_v (\alpha_v \sum_{i,j} \|W_v^T x_i^v - W_v^T x_j^v\|_2^2 s_{ij} + \gamma \|W_v\|_{2,1}) \\ & + \beta \|S\|_F^2 + 2\eta \sum_{i=1}^{c} \sigma_i(L_S), \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1, W_v^T X_v X_v^T W_v = I. \end{aligned} \tag{8}$$

According to Ky Fan's Theory, the following equation is true:

$$\sum_{i=1}^{c} \sigma_i(L_S) = \min_{F \in \mathbb{R}^{n \times c}, F^T F = I} Tr(F^T L_S F). \tag{9}$$

Based on the problem (9), the problem (8) can be further equivalent to

$$\begin{aligned} \min_{S,W_v} \quad & \sum_v (\alpha_v \sum_{i,j} \|W_v^T x_i^v - W_v^T x_j^v\|_2^2 s_{ij} + \gamma \|W_v\|_{2,1}) \\ & + \beta \|S\|_F^2 + 2\eta Tr(F^T L_S F), \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \le s_{ij} \le 1, F \in \mathbb{R}^{n \times c}, F^T F = I, \\ & W_v^T X_v X_v^T W_v = I. \end{aligned} \tag{10}$$

Using an efficient iterative algorithm, problem (10) can be optimized iteratively.

i. Update F

When S, W and α_v are fixed, problem (10) is equivalent to solving the following problem:

$$\min_{F \in \mathbb{R}^{n \times c}, F^T F = I} Tr(F^T L_S F). \tag{11}$$

So the optimal solution of F is the c eigenvectors corresponding to the first c minimum eigenvalues of L_S.

ii. Update W

When S, F and α_v are fixed, problem (10) is equivalent to solving the following problem:

$$\begin{aligned} \min_{W_v} \quad & \sum_v (\alpha_v \sum_{i,j} \|W_v^T x_i^v - W_v^T x_j^v\|_2^2 s_{ij} + \gamma \|W_v\|_{2,1}), \\ s.t. \quad & W_v \in \mathbb{R}^{d_v \times d'_v}, W_v^T X_v X_v^T W_v = I. \end{aligned} \tag{12}$$

Since for different v, the problem (12) is independent, it is equivalent to solving the following problems for each view

$$\begin{aligned} \min_{W} \quad & \sum_{i,j} \alpha \|W^T x_i - W^T x_j\|_2^2 s_{ij} + \gamma \|W\|_{2,1}, \\ s.t. \quad & W \in \mathbb{R}^{d \times d'}, W^T X X^T W = I. \end{aligned} \tag{13}$$

```
Algorithm 1 Optimization in problem (7)
Input: $X = \{X_1, X_2, ..., X_v\}, X_v \in \mathbb{R}^{d_v \times n}$, clustering number $k$, parameter $\beta$, $\eta$ and $\gamma$.
Output: affinity matrix $S \in \mathbb{R}^{n \times n}$ with exact $c$ connected components, where $c = k$; projection matrix $W = \{W_1, W_2, ..., W_v\}, W \in \mathbb{R}^{d_v \times d'_v}$.
    Initialize each vector $s_i$ of $S$ by the optimal solution to the following problem, where the initial value of $\alpha_v$ is $\frac{1}{v}$:
    $\min_{s_i^T \mathbf{1}=1, 0 \leq s_{ij} \leq 1} \sum_{j=1}^{n} (\alpha_v \sum_v \|x_i^v - x_j^v\|_2^2 s_{ij} + \beta s_{ij}^2)$
    Initialize $W_v$ by the optimal solution to the following problem:
    $\min_{W_v} Tr(W_v^T X_v L_S X_v^T W_v), s.t. \quad W_v^T X_v X_v^T W_v = I$
    repeat
        i. update $\alpha_v$ by solving the problem (6),
        ii. update $F$ by solving the problem (11),
        iii. update the projection matrix $W_v$ for each view by solving the problem (18),
        iv. update each row vector of $S$ by solving the problem (23).
    until converge
```

If the function value of $W^T x_i \in \mathbb{R}^{c \times 1}$ is regarded as the value of a node i, the following equation holds:

$$\sum_{i,j} \alpha \|W^T x_i - W^T x_j\|_2^2 s_{ij} = 2Tr(W^T(\alpha X L_S X^T)W). \tag{14}$$

So the problem (13) can be written as follows:

$$\begin{aligned} \min_{W} \quad & Tr(W^T(\alpha X L_S X^T)W) + \frac{\gamma}{2}\|W\|_{2,1}, \\ s.t. \quad & W \in \mathbb{R}^{d \times d'}, W^T X X^T W = I. \end{aligned} \tag{15}$$

Due to the introduction of $\ell_{2,1}$-norm, it is difficult to solve the problem (15). In this paper, we use Lagrangian multiplier method to solve the problem (15). The Lagrangian function of the problem (15) is

$$\begin{aligned} L(W) = & Tr(W^T \alpha X L_S X^T W) + \frac{\gamma}{2}\|W\|_{2,1} \\ & - Tr(\Lambda(W^T X X^T W - I)), \end{aligned} \tag{16}$$

where Λ is a diagonal matrix to enforce the constraint on problem (16). The next step is to derive $L(W)$ and make it equal to 0, that is

$$\frac{\partial L(W)}{\partial W} = (A + A^T)W + 2\gamma DW - 2BW\Lambda = 0, \tag{17}$$

where $D = diag(\frac{1}{4\|\tilde{W}(1,:)\|_2}, ..., \frac{1}{4\|\tilde{W}(d,:)\|_2})$, $\tilde{W}$ represents the current solution, $A = \alpha X L_S X^T$, $B = XX^T$. Because of $A = A^T$, so problem (17) can be simplified as follows:

$$\frac{A + \gamma D}{B} W = W\Lambda. \tag{18}$$

Let Q denote $\frac{A+\gamma D}{B}$, so the problem (18) can be abbreviated to $QW = W\Lambda$. Since Λ is a diagonal matrix, and the characteristic equation is $Q\omega_i = \lambda_i \omega_i (i = 1, 2, ..., d')$, then W is composed of eigenvectors corresponding to c smallest eigenvalues divided by 0. Here λ_i is the i-th smallest eigenvalue except the zero eigenvalues in the eigen-equation, and ω_i is the eigenvector corresponding to the i-th eigenvalue.

iii. Update S

When W, F, and α_v are fixed, problem (10) is equivalent to solving the following problem:

$$\begin{aligned} \min_{S} \quad & \sum_v \alpha_v \sum_{i,j} \|W_v^T x_i^v - W_v^T x_j^v\|_2^2 s_{ij} + \beta\|S\|_F^2 \\ & + 2\eta Tr(F^T L_S F), \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \leq s_{ij} \leq 1. \end{aligned} \tag{19}$$

Let $Z_v = X_v^T W_v$, the problem (19) can be further rewritten as

$$\begin{aligned} \min_{S} & \sum_{i,j} (\sum_v \alpha_v \|z_i^v - z_j^v\|_2^2 s_{ij} + \beta s_{ij}^2 + \eta\|f_i - f_j\|_2^2 s_{ij}), \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \leq s_{ij} \leq 1. \end{aligned} \tag{20}$$

where z_i^v is the i-th vector of Z_v, f_i is the i-th vector of F. Denote

$$d_{ij}^z = \sum_v \alpha_v \|z_i^v - z_j^v\|_2^2, \quad d_{ij}^f = \|f_i - f_j\|_2^2. \tag{21}$$

Note that the problem (20) is independent with respect to each i, we are equivalent to solving the following problem for each i:

$$\begin{aligned} \min_{s_i} \quad & \sum_{j=1}^{n} (d_{ij}^z s_{ij} + \beta s_{ij}^2 + \eta d_{ij}^f s_{ij}), \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \leq s_{ij} \leq 1. \end{aligned} \tag{22}$$

Denote $d_i \in \mathbb{R}^{c \times 1}$ as a vector with the j-th element as $d_{ij} = d_{ij}^z + \eta d_{ij}^f$, then the above problem can be written as follow:

$$\min_{s_i^T \mathbf{1}=1, 0 \leq s_{ij} \leq 1} \|s_i + \frac{1}{2\beta} d_i\|_2^2. \tag{23}$$

The solution of the problem (23) has been given by (Nie, Wang, and Huang 2014), we can get the matrix S with k strong connected subgraphs by updating s_i.

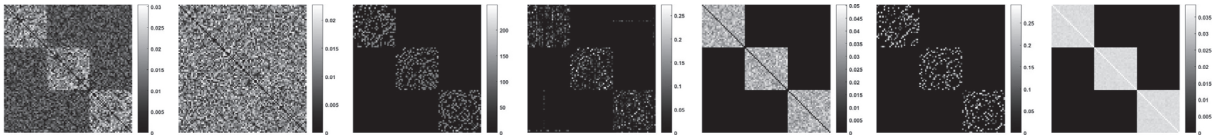
iv. Update α_v, β, η and γ

For the update of the weight coefficient α_v, it has been given in the previous problem (6), which depends on W_v, S.The value of the regularization parameter β can range from 0 to infinity. In this paper, the solution of parameter β is equivalent to solving the following problem:

$$\begin{aligned} \min_{S} \quad & \sum_v \alpha_v \sum_{i,j} \|W_v^T x_i^v - W_v^T x_j^v\|_2^2 s_{ij} + \beta\|S\|_F^2, \\ s.t. \quad & s_i^T \mathbf{1} = 1, 0 \leq s_{ij} \leq 1. \end{aligned} \tag{24}$$

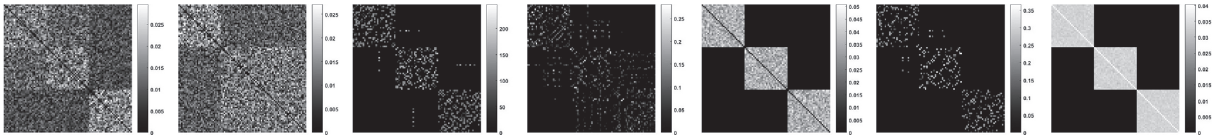
So we can determine the value of the regularized parameter β by (Nie, Wang, and Huang 2014):

$$\beta = \frac{1}{n} \sum_{i=1}^{n} (\frac{K}{2} d_{i,K+1} - \frac{1}{2} \sum_{j=1}^{K} d_{ij}), \tag{25}$$



(a) View 1 (b) View 2 (c) MLAN (d) SwMC (e) RAMC (f) SwMPC (g) RSwMPC

Figure 1: Two initial views and comparison with different methods on toy1 dataset. The results show that the block diagonal matrix obtained by our algorithm is cleaner than the others.



(a) View 1 (b) View 2 (c) MLAN (d) SwMC (e) RAMC (f) SwMPC (g) RSwMPC

Figure 2: Two initial views and comparison with different methods on toy2 dataset. The results show that as the noise increases, our algorithm can still maintain good performance while other algorithms fail.

where K is a nearest neighbor number. The initial value of η can be set to β. Algorithm 1 summarizes the detailed steps to solve the problem (7). Parameter η can be updated by algorithm 1 until the connected components of Laplace matrix L_S is equal to the clustering number k. The value of parameter γ has been given later in the experiment.

v. Time-complexity Analysis

In each iteration, the time complexity of the view weight coefficient α_v, the matrix F, the affinity matrix S and the projection matrix W are $O(d'dn)$, $O(n^2)$, $O(c^2)$ and $O(d^2)$, respectively. Usually we have $c \ll n$, $d' \ll d$, and the number of iterations is less than 10, so the time complexity of our algorithm is $O(d'dn + n^2 + d^2 + c^2) \approx O(n^2 + d^2 + dn)$.

Convergence Analysis

In order to prove the convergence of Algorithm 1, we need to use the following Lemma proposed by (Nie et al. 2010).

Lemma 1. For any positive number u and v,the following inequality holds:

$$u - \frac{u^2}{2v} \le v - \frac{v^2}{2v} \tag{26}$$

Theorem 1. *In each iteration of Algorithm 1, updated S will decrease the objective value of problem (3), which makes problem (3) convergent.*

Proof. Let $\tilde{S}$ represents the S updated in each iteration, it's easy to have:

$$\begin{aligned}&\sum_v \frac{\sum\limits_{i,j} y_{ij}^v \tilde{s}_{ij}}{2(\sum\limits_{i,j} y_{ij}^v s_{ij})^{1/2}} + \beta\|\tilde{S}\|_F^2 \\ \le &\sum_v \frac{\sum\limits_{i,j} y_{ij}^v s_{ij}}{2(\sum\limits_{i,j} y_{ij}^v s_{ij})^{1/2}} + \beta\|S\|_F^2.\end{aligned} \tag{27}$$

where y_{ij}^v denotes $\|W_v^T x_i^v - W_v^T x_j^v\|_2^2$.

According to Lemma 1, we have

$$\begin{aligned}&\sum_v (\sum_{i,j} y_{ij}^v \tilde{s}_{ij})^{1/2} - \sum_v \frac{\sum\limits_{i,j} y_{ij}^v \tilde{s}_{ij}}{2(\sum\limits_{i,j} y_{ij}^v s_{ij})^{1/2}} \\ \le &\sum_v (\sum_{i,j} y_{ij}^v s_{ij})^{1/2} - \sum_v \frac{\sum\limits_{i,j} y_{ij}^v s_{ij}}{2(\sum\limits_{i,j} y_{ij}^v s_{ij})^{1/2}}.\end{aligned} \tag{28}$$

Combined with problem (27) and (28), we get the sum of them as follows:

$$\begin{aligned}&\sum_v (\sum_{i,j} \|W_v^T x_i^v - W_v^T x_j^v\|_2^2 \tilde{s}_{ij})^{1/2} + \beta\|\tilde{S}\|_F^2 \\ \le &\sum_v (\sum_{i,j} \|W_v^T x_i^v - W_v^T x_j^v\|_2^2 s_{ij})^{1/2} + \beta\|S\|_F^2.\end{aligned} \tag{29}$$

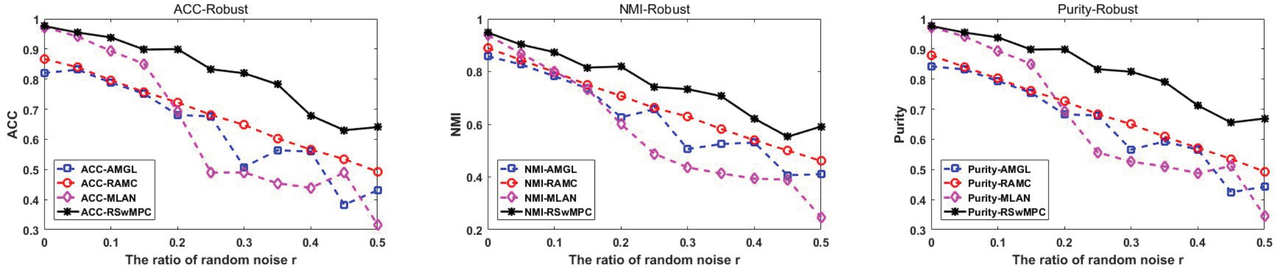
So the convergence of Algorithm 1 is proved by the above formula. □

Experiments

In this section, we prove the effectiveness and robustness of our proposed method on synthetic datasets and real-world datasets, and compare them with other multi-view clustering algorithms. We use the following three evaluation indicators to evaluate clustering performance: Clustering Accuracy (ACC), Normalized Mutual Information (NMI), Purity.

Synthetic Datasets

In this part, we design two datasets to test the effectiveness of our proposed algorithm, toy1 and toy2 datasets, respectively. These two datasets are comprehensive datasets with two views, each view includes a 90 × 90 matrix and three



(a) ACC

(b) NMI

(c) Purity

Figure 3: Robustness comparision with the top four algorithms on HW dataset. Compared with other three algorithms, our algorithm can still maintain good performance with the increase of noise.

Table 1: Statistics of datasets

View	MSRC-v1	Caltech101-7(20)	HW	NUS-WIDE	ORL Faces
1	CM(24)	Gabor(48)	FOU(76)	CH(64)	GIST(512)
2	HOG(576)	WM(40)	FAC(216)	CORR(144)	HOG(864)
3	GIST(512)	CENT(254)	KAR(64)	EDH(75)	LBP(59)
4	LBP(256)	HOG(1984)	PIX(240)	WT(128)	SURF(500)
5	CENT(254)	GIST(512)	ZER(47)	CM55(225)	-
6	-	LBP(928)	MOR(6)	-	-
#Size	210	1474(2386)	2000	2400	400
#Class	7	7(20)	10	12	40

30×30 diagonal block matrices. In these two datasets, the data in the block represents the correlation between the two corresponding points in a cluster and is randomly set to the number between 0 and 1, and the data outside the block represents noise and is randomly set to the number between 0 and e. In toy1 dataset, the noise value of the first and the second view are set as $e = 0.6$ and $e = 1.0$ respectively. In toy2 dataset, for the first view, the initial noise is $e = 0.6$, the $e = 0.8$ is used for the first and second block matrices; for the second view, the initial noise is $e = 0.7$, and $e = 1.0$ is used for the second and third block matrices.

Figure 1 and Figure 2 show the two original views in the toy1 and toy2 datasets and the grayscale images obtained by clustering with MLAN, SwMC, RAMC, SwMPC and our RSwMPC. The experimental results show that our RSwMPC can achieve best clustering performance which ACC, NMI and Purity are all 1. On the toy1 dataset, these algorithms are able to get a clean block diagonal matrix, while RSwMPC get the clearest block diagonal matrix. On the toy2 dataset, after adding noise, all algorithms except RAMC and RSwMPC are invalid. In comparison, the matrix obtained by our RSwMPC is cleaner than others, which also verifies the robustness of RSwMPC. It indicates that our RSwMPC arrives at the optimal solution with different noise.

Real-world Datasets

In Table 1, we briefly summarize the datasets used in this paper, namely MSRC-v1 (Winn and Jojic 2005), Caltech101 (Fei-Fei, Fergus, and Perona 2007) (here, we use two regular subsets of Caltech101-7 and Caltech101-20), Handwritten numerals (HW) (Asuncion and Newman 2007), NUS-WIDE (Chua et al. 2009) (We select 12 categories of animal images, and the first 200 images are selected for each category), ORL Face (Samaria and Harter 1994) datasets.

On the six real-world datasets, we compare our proposed algorithm with the following algorithms: Co-regularized Multi-view Spectral Clustering (Kumar, Rai, and Daume 2011) (Co-reg), Robust Multi-view Spectral Clustering (Xia et al. 2014) (RMSC), Parameter-Free Auto-Weighted Multiple Graph Learning (Nie et al. 2016) (AMGL), Multi-View Clustering with Adaptive Neighbours(Nie, Cai, and Li 2017) (MLAN), Self-weighted Multi-view Clustering (Nie et al. 2017) (SwMC), Robust Auto-Weighted Multi-Feature Clustering (Ren et al. 2018) (RAMC) and Parameter-Free Weighted Multi-View Projected Clustering (Wang et al. 2019) (SwMPC).

For our proposed algorithm, except that the neighbor numbers K of the MSRC-v1 and ORL Face datasets are set to 30 and 5, respectively, the rest is 15. The parameter γ is searched between $1e-6$ to $1e6$, and the step size is 0.5. In addition, we search for the dimensionality d' corresponding to the best clustering result by searching all the dimensionalities of each view in the original dataset. For the sake of simplicity, we set the step size of the parameter γ to 3. In order to distinguish the proportion of different views in the overall clustering performance, for each feature, we initialize α_v to a random weight between 0 and 1. For other algorithms, we set the parameters to be optimal. All the experiments are repeated for 20 times to obtain the mean and standard deviation of clustering results.

Table 2 shows the clustering results (including standard deviation) for all methods, and the best results have been marked in bold. It is easy to see that our algorithm has the best (at least the same) ACC, NMI, Purity on all six datasets. Most of all, the ACC of our proposed algorithm on the Caltech101-7 and Caltech101-20 datasets have increased by about 9.5% and 9.9%. In particular, the ACC, NMI, Purity of our algorithm on the ORL Face dataset have reached an optimal value of 1. Clustering results of our proposed algorithm on other datasets are also improved. In addition, from Table 2, it can be seen that the standard deviation value of all datasets are less than 0.01, and the maximum value is 0.0071, which verify the efficiency and stability of our proposed algorithm. That's because that our algorithm adopts dimensionality reduction processing for high-dimensional

Table 2: Clustering performance comparison. Our algorithm is compared with other algorithms on the MSRC-v1, Caltech101-7 (20), HW, NUS-WIDE and ORL Face datasets. Experimental results show that the overall performance of our proposed algorithm is better than others.

	MSRC-v1			Caltech101-7		
	Acc	NMI	Purity	Acc	NMI	Purity
Co-reg	0.6271 ± 0.0123	0.5399 ± 0.0099	0.6512 ± 0.0105	0.5011 ± 0.0067	0.3867 ± 0.0042	0.8104 ± 0.0022
RMSC	0.8020 ± 0.0212	0.6941 ± 0.0114	0.8110 ± 0.0145	0.4409 ± 0.0101	0.3814 ± 0.0035	0.8033 ± 0.0022
AMGL	0.7095 ± 0.0535	0.6686 ± 0.0330	0.7390 ± 0.0384	0.6538 ± 0.0558	0.5393 ± 0.0482	0.8463 ± 0.0182
MLAN	0.6810 ± 0.0000	0.6299 ± 0.0000	0.7333 ± 0.0000	0.7802 ± 0.0000	0.6304 ± 0.0000	0.8894 ± 0.0000
SwMC	0.8714 ± 0.0000	0.7861 ± 0.0000	0.8714 ± 0.0000	0.6472 ± 0.0000	0.5335 ± 0.0000	0.8365 ± 0.0000
RAMC	0.8871 ± 0.0100	0.8150 ± 0.0169	0.8871 ± 0.0100	0.7497 ± 0.0408	0.6365 ± 0.0243	0.8808 ± 0.0051
SwMPC	0.5571 ± 0.0000	0.5900 ± 0.0000	0.6476 ± 0.0000	0.7436 ± 0.0000	0.5399 ± 0.0000	0.8562 ± 0.0000
RSwMPC	**0.9071 ± 0.0024**	**0.8293 ± 0.0071**	**0.9071 ± 0.0024**	**0.8750 ± 0.0008**	**0.7135 ± 0.0015**	**0.9193 ± 0.0000**

	Caltech101-20			HW		
	Acc	NMI	Purity	Acc	NMI	Purity
Co-reg	0.4470 ± 0.0065	0.5228 ± 0.0040	0.7356 ± 0.0030	0.7199 ± 0.0107	0.6873 ± 0.0057	0.7465 ± 0.0086
RMSC	0.3662 ± 0.0070	0.5037 ± 0.0032	0.7198 ± 0.0042	0.864 ± 0.0118	0.8112 ± 0.0056	0.8730 ± 0.0087
AMGL	0.5067 ± 0.0496	0.5332 ± 0.0263	0.6711 ± 0.0189	0.8007 ± 0.0562	0.8472 ± 0.0307	0.8320 ± 0.0427
MLAN	0.5258 ± 0.0070	0.4744 ± 0.0025	0.6660 ± 0.0000	0.9727 ± 0.0006	0.9384 ± 0.0009	0.9727 ± 0.0006
SwMC	0.5432 ± 0.0000	0.4514 ± 0.0000	0.6676 ± 0.0000	0.7523 ± 0.0531	0.8412 ± 0.0332	0.7865 ± 0.0429
RAMC	0.5928 ± 0.0376	0.6105 ± 0.0222	0.7228 ± 0.0110	0.8574 ± 0.0013	0.8923 ± 0.0016	0.8808 ± 0.0008
SwMPC	0.5281 ± 0.0000	0.4787 ± 0.0000	0.6668 ± 0.0000	0.9605 ± 0.0000	0.9222 ± 0.0000	0.9605 ± 0.0000
RSwMPC	**0.6918 ± 0.0015**	**0.6451 ± 0.0032**	**0.7875 ± 0.0026**	**0.9822 ± 0.0003**	**0.9576 ± 0.0008**	**0.9822 ± 0.0003**

	NUS-WIDE			ORL Face		
	Acc	NMI	Purity	Acc	NMI	Purity
Co-reg	0.2772 ± 0.0019	0.1743 ± 0.0012	**0.2974 ± 0.0017**	0.7732 ± 0.0095	0.9181 ± 0.0038	0.8261 ± 0.0074
RMSC	0.2546 ± 0.0052	0.1632 ± 0.0016	0.2870 ± 0.0023	0.7251 ± 0.0132	0.8566 ± 0.0062	0.7590 ± 0.0117
AMGL	0.1605 ± 0.0250	0.0799 ± 0.0247	0.1615 ± 0.0250	0.9605 ± 0.0203	0.9874 ± 0.0070	0.9698 ± 0.0149
MLAN	0.2108 ± 0.0100	0.1475 ± 0.0063	0.2356 ± 0.0096	0.9475 ± 0.0000	0.9761 ± 0.0000	0.9575 ± 0.0000
SwMC	0.1488 ± 0.0000	0.0818 ± 0.0000	0.1629 ± 0.0000	0.9625 ± 0.0000	0.9906 ± 0.0000	0.9750 ± 0.0000
RAMC	0.2015 ± 0.0062	0.1217 ± 0.0056	0.2155 ± 0.0075	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**
SwMPC	0.1679 ± 0.0000	0.0899 ± 0.0000	0.1845 ± 0.0000	0.9475 ± 0.0000	0.9802 ± 0.0000	0.9600 ± 0.0000
RSwMPC	**0.2778 ± 0.0060**	**0.1810 ± 0.0061**	**0.2974 ± 0.0049**	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**

data, the noise in the datasets is processed by $\ell_{2,1}$-norm, and a non-parametric self-weighting method is adopted to successfully solve the clustering problem on high-dimensional and noisy multi-view datasets.

Robustness Assessment

In this part, we verify the robustness of the algorithm by adding different proportions of noise to real-world datasets. First, we add different proportions of noise based on the original dataset to construct a set of datasets containing noise. Let r be the ratio of noise (r is 0 to 0.5, step size is 0.05), n is the number of data points in the original dataset. In the randomly selected $n \times r$ data points, we add a normal distribution noise with an average of 300 and a standard deviation of 30 to form a set of noisy datasets.

Due to space constraints, we only show the robustness comparison on HW dataset. As shown in Figure 3, for simplicity, we only show the top four algorithms in clustering performance. From Figure 3, we can see that as the noise ratio increases, the performance of all algorithms decrease, but the performance of our algorithm is always in the lead. Furthermore, the performance gain of our algorithm is more significant. When the random noise increases from 0 to 0.5, the performance gain of our algorithm on ACC, NMI, Purity increase from 12.57%, 6.6%, 11.05% to 29.95%, 28.29%, 35.77% compared with the results of RAMC. Compared with the MLAN algorithm, the performance gain of our algorithm on ACC, NMI, Purity increase from 0.32%, 0.96%, 0.32% to 102.34%, 139.27%, 93.65% respectively. Compared with the AMGL algorithm, the performance gain of our algorithm on ACC, NMI, Purity increase from 19.04%, 10.45%, 15.86% to 48.58%, 44.02%, 51.27%, respectively. These results show that our algorithm can suppress noise and maintain good clustering performance when the dataset contains noise.

Conclusion

In this paper, a new robust self-weighted multi-view projection clustering algorithm is proposed. It can simultaneously study the projection matrix, similar matrix and weight coefficients to obtain low-dimensional subspaces with cluster structure. The introduction of the $\ell_{2,1}$-norm on the term not only suppresses noise, but also makes the line sparse and easy to solve. At the same time, the obtained optimal graph can be directly used for clustering without further processing. Experiments on the synthetic datasets and real-world datasets demonstrate the superiority and robustness of our method for processing high-dimensional data with noise. In future work, the framework can be extended to semi-supervised clustering.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 61602379, 61972315 and 61906109) and the Shaanxi Science and Technology Innovation Team Support Project under grant agreement (No. 2018TD-O26). The authors are grateful for the constructive advice on the revision of the manuscript from the anonymous reviewers.

References

Asuncion, A., and Newman, D. 2007. Uci machine learning repository.

Cai, X.; Nie, F.; Huang, H.; and Kamangar, F. 2011. Heterogeneous image feature integration via multi-modal spectral clustering. In *CVPR 2011*, 1977–1984. IEEE.

Chen, X.; Yuan, G.; Wang, W.; Nie, F.; Chang, X.; and Huang, J. Z. 2018. Local adaptive projection framework for feature selection of labeled and unlabeled data. *IEEE transactions on neural networks and learning systems* 29(12):6362–6373.

Chua, T. S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: A real-world web image database from national university of singapore. In *Acm International Conference on Image & Video Retrieval*.

Fan, K. 1949. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences of the United States of America* 35(11):652.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2007. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer vision and Image understanding* 106(1):59–70.

Gao, H.; Nie, F.; Li, X.; and Huang, H. 2015. Multi-view subspace clustering. In *Proceedings of the IEEE international conference on computer vision*, 4238–4246.

Huang, J.; Nie, F.; and Huang, H. 2013. Spectral rotation versus k-means in spectral clustering. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Kumar, A., and Daumé, H. 2011. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 393–400.

Kumar, A.; Rai, P.; and Daume, H. 2011. Co-regularized multi-view spectral clustering. In *Advances in neural information processing systems*, 1413–1421.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in neural information processing systems*, 1813–1821.

Nie, F.; Li, J.; Li, X.; et al. 2016. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In *IJCAI*, 1881–1887.

Nie, F.; Li, J.; Li, X.; et al. 2017. Self-weighted multiview clustering with multiple graphs. In *IJCAI*, 2564–2570.

Nie, F.; Cai, G.; and Li, X. 2017. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Nie, F.; Tian, L.; and Li, X. 2018. Multiview clustering via adaptively weighted procrustes. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2022–2030. ACM.

Nie, F.; Wang, X.; and Huang, H. 2014. Clustering and projected clustering with adaptive neighbors. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 977–986. ACM.

Ren, P.; Xiao, Y.; Xu, P.; Guo, J.; Chen, X.; Wang, X.; and Fang, D. 2018. Robust auto-weighted multi-view clustering. In *IJCAI*, 2644–2650.

Samaria, F. S., and Harter, A. C. 1994. Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, 138–142. IEEE.

Tang, W.; Lu, Z.; and Dhillon, I. S. 2009. Clustering with multiple graphs. In *2009 Ninth IEEE International Conference on Data Mining*, 1016–1021. IEEE.

Wang, R.; Nie, F.; Wang, Z.; Hu, H.; and Li, X. 2019. Parameter-free weighted multi-view projected clustering with structured graph learning. *IEEE Transactions on Knowledge and Data Engineering*.

Winn, J. M., and Jojic, N. 2005. Locus: Learning object classes with unsupervised segmentation. In *Proc International Conference on Computer Vision*.

Xia, R.; Pan, Y.; Du, L.; and Yin, J. 2014. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Xiao, Y.; Ren, P.; Li, Z.; Chen, X.; Wang, X.; and Fang, D. 2019. Rs3cis: Robust single-step spectral clustering with intrinsic subspace. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5482–5489.

Yang, Y.; Shen, H. T.; Ma, Z.; Huang, Z.; and Zhou, X. 2011. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised. In *Twenty-Second International Joint Conference on Artificial Intelligence*.