Received October 18, 2019, accepted November 11, 2019, date of publication November 20, 2019, date of current version December 4, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2954559

Self-Paced Multi-View Clustering via a Novel Soft Weighted Regularizer

ZONGMO HUANG[1], YAZHOU REN[1,2], WENLI LIU[1], AND XIAORONG PU[1]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Institute of Electronic and Information Engineering, UESTC, Dongguan 523808, China

Corresponding author: Yazhou Ren (yazhou.ren@uestc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61806043 and Grant 61872062, in part by the Project funded by China Postdoctoral Science Foundation under Grant 2016M602674, and in part by the Guangdong Basic and Applied Basic Research Foundation.

ABSTRACT Multi-view clustering (MVC), which can exploit complementary information of different views to enhance the clustering performance, has attracted people's increasing attentions in recent years. However, existing multi-view clustering methods typically solve a non-convex problem, therefore are easily stuck into bad local minima. In addition, noisy data and outliers affect the clustering process negatively. In this paper, we propose self-paced multi-view clustering via a novel soft weighted regularizer (SPMVC) to address these issues. Specifically, SPMVC progressively selects samples to train the MVC model from simplicity to complexity in a self-paced manner. A novel soft weighted regularizer is proposed to further reduce the negative impact of outliers and noisy data. Experimental results on real-world data sets demonstrate the effectiveness of the proposed method.

INDEX TERMS Multi-view clustering, self-paced learning, soft weighting.

I. INTRODUCTION

The aim of clustering [1] is to divide a set of objects into different groups such that similar objects will be grouped into the same cluster, while dissimilar ones are placed into different clusters. Clustering has been widely used in different fields, including pattern recognition, social network analysis, astronomical data analysis, information retrieval, and bioinformatics, etc.

In the past couple of decades, a large number of clustering models have been proposed, such as k-means [2], fuzzy clustering [3], density-based clustering [4], [5], distribution-based clustering [6], [7], mean shift clustering [8], [9], consensus clustering [10]–[12], clustering based on deep neural networks [13], [14] etc. However, these conventional algorithms can only deal with single view clustering problems. In real-world clustering tasks, data sets are often described by multiple views, each providing a specific aspect of data. To take full advantage of complementary information from different views, multi-view clustering was proposed [15].

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

Recently, a number of multi-view clustering methods [16]–[23] have been proposed and have been proved to be effective in solving multi-view clustering problems. However, existing multi-view clustering methods typically solve a non-convex optimization problem [24], which results in the consequence that they get trapped in bad local minima easily.

To address the non-convexity issue, an effective and efficient way is to use curriculum learning [25] and self-paced learning [26]. The core idea of curriculum learning and self-paced learning is imitating the mechanisms of cognition of humans. At first, the model is trained with easy samples, and then hard samples are involved in the training process gradually. In clustering tasks, easy samples can be interpreted as the data points with smaller loss values, while the hard ones are usually associated with large loss values.

Besides, the existence of noisy data and outliers is another factor that negatively affects the clustering performance of conventional multi-view clustering methods. Kong *et al.* [27] show that using $l_{2,1}$-norm rather than Frobenius norm grants model stronger resistance against noises and outliers. However, since noises and outliers contribute equally with the normal samples to the training, the noisy data issue still exists. To this end, a novel soft weighting regularization term is developed in this work, which reduces the impact of noises

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

and outliers by automatically assigning lower weights to those samples with larger loss values.

Overall, in this paper, we propose self-paced multi-view clustering via a novel soft weighted regularizer (SPMVC) to address the non-convexity issue and noisy data issue. Concretely, our SPMVC solves the former problem by progressively selecting samples to train the MVC model from simplicity to complexity, while a novel soft weighted regularizer is developed to further reduce the impact of noisy data and outliers.

In summary, the contributions of this paper include:

i) Alleviate the non-convexity issue of conventional multi-view clustering algorithms by taking advantage of the self-paced learning.
ii) Reduce the impact of the outliers and noises on the clustering result by developing a novel soft weighting regularization term for self-paced learning.
iii) Derive an efficient optimizing method to solve the proposed model. Experiments on real data are concluded to demonstrate the effectiveness of SPMVC.

II. RELATED WORK

A. MULTI-VIEW CLUSTERING

Multi-view clustering focus on using information given by the multiple views to enhance the clustering performance. In recent years, a lot of multi-view clustering algorithms have been proposed.

Kumar and Daumé [16] proposed a co-trained multi-view spectral clustering method (Co-train), which assumes that a data point should be grouped in the same cluster among all the views. Kumar *et al.* [17] designed two co-regularization strategies and achieved a new spectral clustering structure (Co-reg). To solve the problem of noises and outliers, Tzortzis and Likas [18] assigned a weight to each view based on its quality and proposed multi-view kernel k-means clustering (MVKKM). Cai *et al.* [19] proposed a robust multi-view k-means clustering (RMKMC) which utilizes $l_{2,1}$-norm in the objective function. Huang *et al.* [24] proposed a novel multi-view clustering with multi-view capped-norm k-means (CAMVC). By exploiting the capped-norm loss as the objective, CAMVC could decrease the influence caused by noises and outliers. Huang *et al.* [28] proposed a joint graph-based multi-view clustering model and further boosted the learning performance of multiple kernels.

B. SELF-PACED LEARNING

Similar to the process of human learning, self-paced learning (SPL) chooses simple examples first and then utilizes complex samples until all the examples are selected to train [26]. SPL has been proved that it benefits in alleviating bad local optima [29]. For its effectiveness, SPL has been employed to various machine learning tasks, such as classification [30], clustering [31], [32], computer vision [33]–[35], feature corruption [36], boosting learning [37], diagnosis of disease [38], etc.

Supancic and Ramanan [39] applied self-paced learning to solve the problem of long-term object tracking. Ma *et al.* [40] developed co-training with SPL and proposed a novel co-training algorithm named self-paced co-training (SPaCo). Meng *et al.* [41] provided some theoretical analyses for SPL. Instead of simply dividing the examples into 'easy' and 'complex', a series of self-paced learning algorithms with soft weighting schemes have been proposed [30], [32], [33], [37], [42]. In [32], Ren et al. designed a self-paced learning algorithm with soft weighting for multi-task multi-view clustering (MTMVC), in which the impact of noises and outliers is effectively reduced.

In this paper, a self-paced multi-view clustering method (SPMVC) is proposed, which develops a novel soft weighting SPL scheme for multi-view clustering. In contrast to multi-view self-paced learning (MSPL) [42] which also applies SPL in multi-view clustering, $l_{2,1}$-norm is utilized for objective function instead of Frobenius norm and a novel soft weighted regularizer is further proposed in this paper to enhance the robustness to noisy data and outliers. Experiments on real data also demonstrate our method performs better than MSPL.

III. PROPOSED APPROACH

This section elucidates the proposed self-paced multi-view clustering via a novel soft weighted regularizer (SPMVC).

As mentioned previously, the non-convexity issue and noisy data issue are the mainly factors that cause the bad performance of conventional multi-view clustering algorithms. To address these problems, our method trains the model in a self-paced manner, by gradually selecting samples from simplicity to complexity. Meanwhile, a novel soft regularizer is proposed to address the noisy data issue. The resulting objective function of our model is:

$$
\begin{aligned}
&\min_{C^v,B,W^v} \sum_{v=1}^{m} ||(X^v - C^v B)W^v||_{2,1} + \sum_{v=1}^{m}\sum_{i=1}^{n} f(w_i^v, \lambda^v) \\
&s.t. \quad C^v \geq 0, \quad w_i^v \in [0,1], \\
&\qquad b_{ij} \in \{0,1\}, \sum_{i=1}^{k} b_{ij} = 1, \quad \forall j = 1,2,\ldots,n
\end{aligned}
\tag{1}
$$

$X^v = \{x_1^v, x_2^v, \ldots, x_n^v\}$, $v = 1, 2, \ldots, m$, denotes the v^{th} view of the data set, where n means the number of data points and m is the number of views. $C^v = \{c_1^v, c_2^v, \ldots, c_k^v\}$ denotes the cluster centers of the v^{th} view, k is the predefined number of clusters. The weight matrix $W^v = diag(w_1^v, w_2^v, \ldots, w_n^v)$, where the value of w_i^v represents the weight of the i^{th} sample in the v^{th} view. $B = \{b_1, b_2, \ldots, b_n\} \in R^{k \times n}$ reflects the clustering assignment and is shared by all the m views. The novel soft weighted regularizer $f(w_i^v, \lambda^v)$ is written as:

$$
f(w_i^v, \lambda^v) = (w_i^v + \gamma^v e^{-\lambda^v}) ln(w_i^v + \gamma^v e^{-\lambda^v}) - w_i^v (ln\gamma^v + 1) \tag{2}
$$

This regularizer controls how samples in each view contribute to the training process. In brief, under the influence of this regularizer, those samples with higher loss values contribute less to the training process, and vice versa. As a consequence, noises and outliers are typically associated with large loss values and there negative influence can be reduced.

Our method contains two main parts, i.e., initialization and optimization.

A. INITIALIZATION

Firstly, each X^v, $v = 1, 2, \ldots, m$, is normalized to be non-negative. Then, cluster center matrices C^v and assignment matrix B are initialized by solving the following problem:

$$\min_{C^v, B} \sum_{v=1}^{m} ||(X^v - C^v B)||_{2,1}$$
$$s.t. \quad C^v \geq 0, \quad b_{ij} \in \{0, 1\}, \sum_{i=1}^{k} b_{ij}, \; \forall i = 1, 2, \ldots, n \quad (3)$$

Eq. (3) does not learn weights for different views, and thus can be considered as a simple version of [19]. Following [19], this optimization problem can be solved by alternately updating C^v and B. Actually, Eq. (3) can be also seen as a special case of our model Eq. (1) when all the samples participate in training the model with default weight 1.

B. OPTIMIZATION

The objective function Eq. (1) can be optimized w.r.t. one variable while other variables are fixed.

1) STEP 1: FIX C^V AND B, UPDATE W^V.

When C^v and B are fixed, Eq. (1) can be written as:

$$\min_{W} \sum_{v=1}^{m} ||(X^v - C^v B) W^v||_{2,1} + \sum_{v=1}^{m} \sum_{i=1}^{n} f(w_i^v, \lambda^v)$$
$$s.t. \quad w_i^v \in [0, 1] \quad (4)$$

The contribution given by every data point in each view can be calculated separately. Thus, w_i^v can be solved separately by:

$$\min_{w_i^v} l_i^v w_i^v + f(w_i^v, \lambda^v) \quad (5)$$

where

$$l_i^v = ||x_i^v - C^v b_i|| \quad (6)$$

Substituting Eqs. (6) and (2) into Eq. (5), and setting the gradient w.r.t. w_i^v to zero, we can obtain:

$$0 = l_i^v + ln(w_i^v + \gamma^v e^{-\lambda^v}) - ln(\gamma^v) \quad (7)$$

Thus, the optimal value of w_i^v is:

$$w_i^v = \gamma^v (e^{-l_i^v} - e^{-\lambda^v}) \quad (8)$$

Since $w_i^v \in [0, 1]$, w_i^v achieves the minimum 0 when $l_i^v \geq \lambda^v$ and reaches the maximum 1 when

$$l_i^v \leq ln \frac{\gamma^v}{1 + \gamma^v e^{-\lambda^v}} \quad (9)$$

Here, γ^v controls how many samples are associated with the highest weight 1. It is defined as:

$$\gamma^v = \frac{1}{e^{-\alpha\lambda^v} - e^{-\lambda^v}} \quad (10)$$

where $\alpha \in [0, 1]$. Replacing the γ^v in Eq. (9) with Eq. (10), the right side of Eq. (9) is actually equal to $\alpha\lambda^v$. As a result, the number of samples that obtain the highest weight 1 declines as the value of α increases. Specifically, when α is set to 1, the regularizer $f(w_i^v, \lambda^v)$ plays the same role as the traditional hard weighted regularizer. That is, those samples whose loss values are smaller than λ^v will be assigned with weight 1.

For simplicity, in this paper, the parameter α is always set to 0.5. Then, the formula of updating W^v becomes:

$$w_i^v = \begin{cases} 1 & l_i^v \leq \frac{\lambda^v}{2} \\ \gamma^v (e^{-l_i^v} - e^{-\lambda^v}) & \frac{\lambda^v}{2} < l_i^v < \lambda^v \\ 0 & l_i^v \geq \lambda^v \end{cases} \quad (11)$$

From Eq. (11), our novel soft regularizer enables the data points with smaller loss values to get higher weights, as shown in Figure 1. In this way, the impact caused by noisy data and outliers (which are typically with large loss values) can be significantly reduced. Moreover, by increasing the value of λ^v to let more samples join the training process in every iteration, our method trains the MVC model from simplicity to complexity progressively.

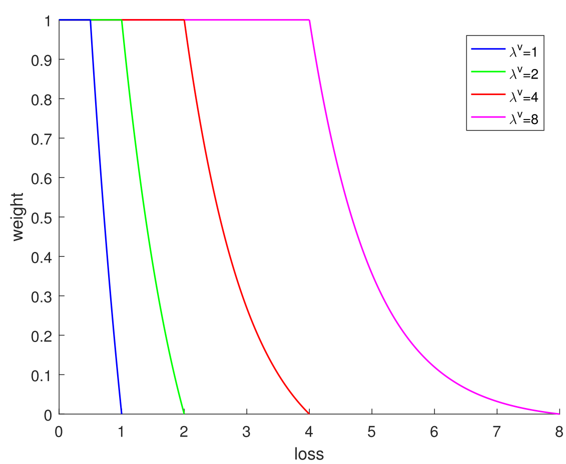

FIGURE 1. Curves correspond to soft weighting of Eq. (11).

2) STEP 2: FIX W^V, ALTERNATELY UPDATE C^V AND B.

a: FIX W^V AND B, UPDATE C^V

When the weight matrix W^v and assignment matrix B are fixed, $f(w_i^v, \lambda^v)$ in Eq. (1) is a constant. Thus, optimizing Eq. (1) is equivalent to solving the following problem:

$$\min_{C^v} \sum_{v=1}^{m} ||(X^v - C^v B) W^v||_{2,1}$$
$$s.t. \quad C^v \geq 0 \quad (12)$$

It is difficult to optimize this function directly. To solve this problem, we firstly define $D^v = diag(d_1^v, d_2^v, \ldots, d_n^v)$ where:

$$d_i^v = \frac{w_i^v}{||x_i^v - C^v b_i||} \tag{13}$$

Then, solving Eq. (12) becomes minimizing the following function for each view:

$$J(C^v) = Tr\big((X^v - C^v B)D^v (X^v - C^v B)^T\big) \tag{14}$$

where $Tr(A)$ denotes the trace of matrix A.

To solve this problem, as in [27], an auxiliary function $Z(C^v, C^{v'})$ of $J(C^v)$ is defined:

$$Z(C^v, C^{v'}) = Tr(X^v D^v X^{vT}) - 2Tr(C^{vT} X^v D^v B^T) + \sum_i \sum_j \frac{(C^{v'} B D^v B^T)_{ij} C_{ij}^{v2}}{C_{ij}^{v'}} \tag{15}$$

The reason why we choose $Z(C^v, C^{v'})$ as the auxiliary function of $J(C^v)$ is that $Z(C^v, C^{v'})$ satisfies the following conditions that have been proved by Kong *et al.* [27]:

$$J(C^v) = Z(C^v, C^v) \tag{16}$$

$$J(C^v) \le Z(C^v, C^{v'}) \tag{17}$$

Let $f(C^{v(t+1)}) = Z(C^{v(t+1)}, C^{v(t)})$, then its gradient is:

$$\frac{\partial f(C^{v(t+1)})}{\partial C_{ij}^{v(t+1)}} = 2\frac{(C^{v(t)} B D^v B^T)_{ij} C_{ij}^{v(t+1)}}{C_{ij}^{v(t)}} - 2(X^v D B^T)_{ij} \tag{18}$$

Setting this gradient to 0, we obtain the optimal solution:

$$C_{ij}^{v(t+1)} = C_{ij}^{v(t)} \frac{(X^v D^v B^T)_{ij}}{(C^{v(t)} B D^v B^T)_{ij}} \tag{19}$$

The above formula is the updating rule of cluster center matrix C^v. It decreases the objective value of Eq. (12), which is proved in Theorem 1.

Theorem 1: Updating rule Eq. (19) decreases the value of objective function Eq. (12).

Proof: The second order derivatives (Hessian matrix) of $f(C^{v(t+1)})$ is:

$$\frac{\partial^2 f(C^{v(t+1)})}{\partial C_{ij}^{v(t+1)} \partial C_{kl}^{v(t+1)}} = 2\frac{(C^{v(t)} B D^v B^T)_{ij}}{C_{kl}^{v(t)}} \delta_{jl}\delta_{ik} \tag{20}$$

where δ_{jl} is equal to 1 when $j = l$, and is equal to 0 otherwise. Thus, the Hessian matrix of $f(C^{v(t+1)})$ is semi-positive definite, which implies that $f(C^{v(t+1)})$ is a convex function. So the optimal solution shown in Eq. (19) is the global minima of $f(C^{v(t+1)})$. Merging this conclusion and the conditions represented in Eq. (16) and Eq. (17), the following unequal relationship can be inferred:

$$\begin{aligned} J(C^{v(t+1)}) &\le Z(C^{v(t+1)}, C^{v(t)}) \\ &\le Z(C^{v(t)}, C^{v(t)}) = J(C^{v(t)}) \end{aligned} \tag{21}$$

With this relationship, we can further prove the following formula is satisfied:

$$\begin{aligned} &||(X^v - C^{v(t+1)}B)W^v||_{2,1} - ||(X^v - C^{v(t)}B)W^v||_{2,1} \\ &\qquad \le \frac{1}{2}[J(C^{v(t+1)}) - J(C^{v(t)})] \end{aligned} \tag{22}$$

To this end, we represent the left side (the first line) of Eq. (22) as *LHS* and the right side (the second line) as *RHS*. Then, we can obtain:

$$\begin{aligned} &LHS - RHS \\ &= \sum_{i=1}^n w_i^v (||X_i^v - C^{v(t+1)} b_i|| - \frac{1}{2}||X_i^v - C^{v(t)} b_i|| \\ &\quad - \frac{||X_i^v - C^{v(t+1)} b_i||^2}{2||X_i^v - C^{v(t)} b_i||}) \\ &= -\frac{1}{2}\sum_{i=1}^n \frac{w_i^v}{||X_i^v - C^{v(t)} b_i||}(||X_i^v - C^{v(t)} b_i|| \\ &\quad - ||X_i^v - C^{v(t+1)} b_i||)^2 \\ &\le 0 \end{aligned} \tag{23}$$

From Eq. (21), we have $RHS \le 0$. Therefore, $LHS \le 0$, which means that the updating rule Eq. (19) could decrease the value of objective function Eq. (12) monotonically. □

b: FIX W^v AND C^v, UPDATE B

When the weight matrix W^v and cluster center matrix C^v are fixed, optimizing Eq. (1) is equivalent to solving the following problem for each data point separately:

$$\begin{aligned} &\min_{b_i} \sum_{v=1}^m w_i^v ||x_i^v - C^v b_i|| \\ &s.t. \quad b_{ij} \in \{0, 1\}, \quad \sum_i^k b_{ij} = 1 \end{aligned} \tag{24}$$

This problem can be easily solved by exhaustive search method. That is, the optimal solution b_i^* is e_j, where e_j denotes the j^{th} column of the Identity matrix and is obtained by solving:

$$\arg\min_{e_j} \sum_{v=1}^m w_i^v ||x_i^v - C^v e_j|| \tag{25}$$

In **Step 2**, we alternately update C^v of each view by Eq. (19) and update B by Eq. (25) until the terminating condition is satisfied.

The **Step 1** and **Step 2** correspond to an entire iteration. SPMVC keeps the iteration running until all the data points are selected in the training process. At first, for each view, λ^v is initialized to select half of the data points to train the model. Then, in each of the following iterations, λ^v is varied to let 10% more samples to be chosen. Therefore, the algorithm will finish in only 6 iterations. After that, the final cluster center matrix $C^1, C^2, \ldots, C^m$ and assignment matrix B reflect the clustering result. The process of SPMVC is summarized in Algorithm 1.

TABLE 1. The data sets used in the experiments (dimensionality).

View	Handwritten numerals	BBCsport	Movies	Reuters
1	Profile correlations (216)	Segment1 (3183)	Keyword(1878)	English(2000)
2	Fourier coefficients (76)	Segment2 (3203)	Actor(1398)	French(2000)
3	Karhunen coefficients (64)	-	-	German(2000)
4	Morphological (6)	-	-	Italian(2000)
5	Pixel averages (240)	-	-	Spanish(2000)
6	Zernike moments (47)	-	-	-
Data points	2000	554	617	1200
Classes	10	5	17	6

Algorithm 1 The SPMVC Algorithm.

Input: Data set X^v, $v = 1, 2, \ldots, m$; Cluster number k.
Output: The final cluster center matrix C^v, assignment matrix B, $v = 1, 2, \ldots, m$.

```
Initialize C^v and B by solving Eq. (3).
Initialize λ^v for each view, v = 1, 2, ..., m.
repeat
  for each view v do
    Fix C^v and B, update W^v and D^v:
    Update W^v by soft weighting according to Eq. (11).

    Update D^v according to Eq. (13).
  end for
  repeat
    for each view v do
      Fix W^v and B, update C^v:
      Update C^v according to Eq. (19).
    end for
    Fix C and W, update B:
    Update B according to Eq. (25):
  until convergence or exceed the maximal number of
  iterations
  Increase λ^v to select more samples.
until all data points are selected
return C^v and B, v = 1, 2, ..., m.
```

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETUP

1) DATA SETS

Handwritten numerals [1] data set is chosen from UCI machine learning repository. This data set consists of 2000 points with features of handwritten numerals (0-9). For 10 data classes, each class has the same data quantity. Those data points are represented by the following six features: 76 Fourier coefficients of the character shapes, 216 profile correlations, 64 Karhunen-Love coefficients, 240 pixel averages in 2 × 3 windows, 47 Zernike moments, and 6 morphological features.

BBCsport data set originates from sports news reported by the BBC Sport [43]. BBCsport is comprised of 2012 articles with 5 genres. Each article was divided into tow segments, every segment represents a single view and has more than two hundred words which is related to the original article logically.

Movies[2] is collected from IMDb,[3] and contains 617 movies over 17 labels. The two views of data are the 1878 keywords used for more than 3 movies and 1398 actors starred in more than 2 movies.

Reuters[2] selects 1200 articles from 6 categories (C15, CCAT, E21, ECAT, GCAT and M11), each providing 200 articles. Every document is written in five different languages (English, French, German, Italian, and Spanish), corresponding to five different views in the experiments.

The characteristics of data sets is shown in Table 1.

2) COMPARING METHODS

We compare the proposed SPMVC model with seven existing state-of-the-art multi-view clustering approaches:

- Co-train: Co-trained multi-view spectral clustering [16].
- Co-reg: Co-regularized multi-view spectral clustering [17].
- MVKKM: Multi-view kernel k-means clustering [18].
- RMVK: Robust multi-view k-means clustering [19].
- AMGL: Auto-Weighted Multiple Graph Learning [44].
- CAMVC: Robust Capped-Norm Multi-View Clustering [24].
- MSPL: Multi-View Self-Paced Learning for Clustering [42].

In order to make a comprehensive comparison, we employ k-means clustering on each single view (e.g., KM(1) means applying KM on the first view). We also perform k-means on the concatenated features from all the views (KM(Allfea)). Features of each view are assigned with the same weight. The number of clusters is always set to the ground truth number of classes for all methods.

3) EVALUATION MEASURE

We use clustering accuracy (ACC), normalized mutual information (NMI), and purity to evaluate the clustering performance. Bigger values of NMI, ACC, and purity mean better clustering performance. The average results and standard deviations of 10 independent runs are reported in this paper. By utilizing t-test, the statistical significance are evaluated at 5% significance level in our experiments.

[1] https://archive.ics.uci.edu/ml/datasets.php

[2] http://lig-membres.imag.fr/grimal/data.html

[3] http://www.imdb.org

TABLE 2. Results on handwritten numerals.

Methods	ACC(%)	purity(%)	NMI(%)
KM(1)	60.25±4.62	65.86±3.41	61.25±2.08
KM(2)	62.12±7.19	64.73±5.49	63.59±3.52
KM(3)	69.96±9.92	73.52±8.34	70.19±5.82
KM(4)	37.27±1.33	43.05±0.71	48.04±0.54
KM(5)	72.13±4.60	75.33±4.17	72.28±2.76
KM(6)	52.58±4.70	55.89±2.99	49.71±1.63
KM(Allfea)	50.81±6.46	56.39±4.04	57.71±1.79
Co-train	74.62±4.28	76.03±2.49	71.66±1.44
Co-reg	**81.11±6.17**	83.17±4.41	77.04±2.28
MVKKM	60.51±2.36	64.47±1.74	65.31±1.17
RMVK	60.89±6.18	63.70±4.48	65.16±2.16
AMGL	**83.58±2.70**	**85.87±2.22**	**88.09±1.23**
CAMVC	74.08±8.57	78.71±6.41	77.81±4.07
MSPL	76.70±6.38	81.33±4.49	84.55±2.82
SPMVC	**81.13±8.55**	**85.20±6.28**	**86.20±3.61**

TABLE 3. Results on BBCsport.

Methods	ACC(%)	purity(%)	NMI(%)
KM(1)	40.70±4.17	42.00±3.75	11.15±5.64
KM(2)	38.05±4.24	40.50±4.77	10.15±6.47
KM(Allfea)	40.40±5.99	42.19±5.72	12.95±9.15
Co-train	36.08±1.54	38.25±1.33	4.06±1.01
Co-reg	29.61±0.39	36.21±0.09	2.17±0.30
MVKKM	39.30±5.74	41.34±6.06	10.97±9.26
RMVK	36.07±1.17	36.53±1.01	2.51±1.60
AMGL	35.97±0.26	36.51±0.16	2.64±0.34
CAMVC	37.28±3.43	37.90±3.27	4.60±5.11
MSPL	36.76±2.31	37.15±2.23	4.06±3.13
SPMVC	**44.44±6.43**	**46.71±4.56**	**19.25±7.59**

B. CLUSTERING RESULTS ON REAL DATA

In this section, we evaluate the performance of the proposed method and the comparing approaches on real data sets. The ACC, NMI, and purity values of different data sets are given in Tables 2 - 5. In each column, the best and comparable results are highlighted in boldface. From these tables, the following observations can be concluded:

(i) Multi-view clustering methods generally perform better than the single-view algorithm, i.e., k-means, indicating the superiority of employing comprehensive information from multiple views.
(ii) k-means achieves different performance on different views. The main reason is that different views exert different influence on the clustering result.
(iii) Our SPMVC method always obtain the best or comparable clustering results. Specifically, SPMVC performs better than MSPL on all data sets, which demonstrates the effectiveness of $l_{2,1}$-norm and the novel SPL soft weighted regularizer.

C. STUDY ON THE CONVERGENCE

This section shows the convergence trend of our method. Figure 2 shows the convergence curve on different data sets when all the samples participate in the training process. Here, the abscissa means the number of iterations and the ordinate is the objective value of Eq. (1). It is obvious that SPMVC converges very fast when all the samples are selected for training, empirically revealing the efficiency of our model.

TABLE 4. Results on movies.

Methods	ACC(%)	purity(%)	NMI(%)
KM(1)	13.13±1.97	14.33±2.06	11.97±3.47
KM(2)	11.00±1.04	12.04±0.90	8.98±1.20
KM(AllFea)	11.72±2.75	12.79±2.92	9.32±3.84
Co-train	9.08±0.44	10.15±0.38	5.02±0.32
Co-reg	11.04±0.56	12.27±0.43	7.28±0.49
MVKKM	11.85±3.59	12.97±3.65	9.24±4.96
RMVK	12.30±2.80	13.35±2.84	10.01±3.88
AMGL	7.97±0.10	9.61±0.08	4.72±0.01
CAMVC	12.01±2.43	13.11±2.46	9.34±3.03
MSPL	14.23±2.59	15.04±2.57	12.65±2.69
SPMVC	**18.27±1.48**	**19.48±1.35**	**17.01±1.90**

TABLE 5. Results on reuters.

Methods	ACC(%)	purity(%)	NMI(%)
KM(1)	21.28±4.82	21.59±4.83	6.77±6.04
KM(2)	26.27±9.09	26.66±9.20	12.14±9.35
KM(3)	26.36±6.91	26.54±6.96	10.65±6.20
KM(4)	28.70±4.85	29.02±4.91	12.03±6.49
KM(5)	28.21±5.74	28.40±5.76	10.96±6.83
KM(Allfea)	24.95±8.92	25.06±8.95	9.69±8.53
Co-train	16.93±0.07	17.11±0.04	0.87±0.08
Co-reg	19.98±1.45	20.43±1.51	2.31±0.36
MVKKM	21.65±4.71	22.20±5.10	7.98±6.43
RMVK	25.56±18.42	26.25±19.67	9.27±5.61
AMGL	18.34±0.10	19.83±0.68	5.92±1.25
CAMVC	21.87±4.69	22.39±4.69	7.61±4.50
MSPL	29.57±7.44	30.30±7.31	14.21±5.77
SPMVC	**39.79±2.91**	**40.93±2.48**	**22.65±2.32**

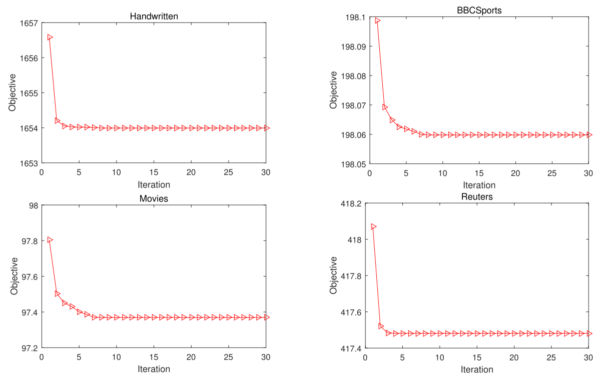

FIGURE 2. Convergence curve of SPMVC on all data sets.

V. CONCLUSION

In this paper, a novel clustering method named self-paced multi-view clustering via a novel soft weighted regularizer (SPMVC) is proposed. Self-paced learning is applied in multi-view model to address the non-convexity issue by gradually choosing samples for training from simplicity to complexity. Meanwhile, $l_{2,1}$-norm and a novel SPL soft weighted regularizer are used to significantly reduce the negative impact of noises and outliers. Experiments on multi-view data sets demonstrate the effectiveness and efficiency of the proposed SPMVC.

REFERENCES

[1] J. A. Hartigan, *Clustering Algorithms*. Hoboken, NJ, USA: Wiley, 1975.
[2] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, 1967, pp. 281–297.
[3] Y. Suo, T. Liu, F. Yu, and X. Jia, "Application of clustering analysis in brain gene data based on deep learning," *IEEE Access*, vol. 7, pp. 2947–2956, 2018.
[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, 1996, pp. 226–231.
[5] Y. Ren, X. Hu, K. Shi, G. Yu, D. Yao, and Z. Xu, "Semi-supervised denpeak clustering with pairwise constraints," in *Proc. 15th Pacific Rim Int. Conf. Artif. Intell.*, 2018, pp. 837–850.
[6] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006, pp. 430–439.
[7] J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, vol. 49, pp. 803–821, Sep. 1993.
[8] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.
[9] Y. Ren, C. Domeniconi, G. Zhang, and G. Yu, "A weighted adaptive mean shift clustering algorithm," in *Proc. SIAM Int. Conf. Data Mining*, 2014, pp. 794–802.
[10] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Dec. 2002.
[11] Y. Ren, C. Domeniconi, G. Zhang, and G. Yu, "Weighted-object ensemble clustering: Methods and analysis," *Knowl. Inf. Syst.*, vol. 51, no. 2, pp. 661–689, 2017.
[12] J. Tian, Y. Ren, and X. Cheng, "Stratified feature sampling for semi-supervised ensemble clustering," *IEEE Access*, vol. 7, pp. 128669–128675, 2019.
[13] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 478–487.
[14] Y. Ren, K. Hu, X. Dai, L. Pan, S. C. H. Hoi, and Z. Xu, "Semi-supervised deep embedded clustering," *Neurocomputing*, vol. 325, pp. 121–130, Jan. 2019.
[15] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2004, pp. 19–26.
[16] A. Kumar and H. Daumé, III, "A co-training approach for multi-view spectral clustering," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 393–400.
[17] A. Kumar, P. Rai, and H. Daumé, III, "Co-regularized multi-view spectral clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1413–1421.
[18] G. Tzortzis and A. Likas, "Kernel-based weighted multi-view clustering," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 675–684.
[19] X. Cai, F. Nie, and H. Huang, "Multi-view K-means clustering on big data," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 2598–2604.
[20] Y.-M. Xu, C.-D. Wang, and J.-H. Lai, "Weighted multi-view clustering with feature selection," *Pattern Recognit.*, vol. 53, pp. 25–35, May 2016.
[21] G.-Y. Zhang, C.-D. Wang, D. Huang, and W.-S. Zheng, "Multi-view collaborative locally adaptive clustering with minkowski metric," *Expert Syst. Appl.*, vol. 86, pp. 307–320, Nov. 2017.
[22] D. Xie, Q. Gao, Q. Wang, and S. Xiao, "Multi-view spectral clustering via integrating global and local graphs," *IEEE Access*, vol. 7, pp. 31197–31206, 2019.
[23] Y.-P. Zhao, L. Chen, M. Gan, and C. L. P. Chen, "Multiple kernel fuzzy clustering with unsupervised random forests kernel and matrix-induced regularization," *IEEE Access*, vol. 7, pp. 3967–3979, 2018.
[24] S. Huang, Y. Ren, and Z. Xu, "Robust multi-view data clustering with multi-view capped-norm K-means," *Neurocomputing*, vol. 311, pp. 197–208, Oct. 2018.
[25] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Int. Conf. Mach. Learn.*, 2009, pp. 41–48.
[26] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1189–1197.
[27] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using L21-norm," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 673–682.
[28] S. Huang, Z. Kang, I. W. Tsang, and Z. Xu, "Auto-weighted multi-view clustering via kernelized graph learning," *Pattern Recognit.*, vol. 88, pp. 174–184, Apr. 2019.
[29] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, "Self-paced curriculum learning," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2694–2900.
[30] Y. Ren, P. Zhao, Y. Sheng, D. Yao, and Z. Xu, "Robust softmax regression for multi-class classification with self-paced learning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2641–2647.
[31] Y. Ren, X. Que, D. Tao, and Z. Xu, "Self-paced multi-task clustering," *Neurocomputing*, vol. 350, no. 1, pp. 212–220, Jul. 2019.
[32] Y. Ren, X. Yan, Z. Hu, and Z. Xu, "Self-paced multi-task multi-view capped-norm clustering," in *Proc. Int. Conf. Neural Inf. Process.*, 2018, pp. 205–217.
[33] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann, "Easy samples first: Self-paced reranking for zero-example multimedia search," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 547–556.
[34] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller, "Shifting weights: Adapting object detectors from image to video," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 647–655.
[35] W. Xu, W. Liu, X. Huang, J. Yang, and S. Qiu, "Multi-modal self-paced learning for image classification," *Neurocomputing*, vol. 309, pp. 134–144, Oct. 2018.
[36] Y. Ren, P. Zhao, Z. Xu, and D. Yao, "Balanced self-paced learning with feature corruption," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2017, pp. 2064–2071.
[37] T. Pi, X. Li, Z. Zhang, D. Meng, F. Wu, J. Xiao, and Y. Zhuang, "Self-paced boost learning for classification," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1932–1938.
[38] X. Que, Y. Ren, J. Zhou, and Z. Xu, "Regularized multi-source matrix factorization for diagnosis of Alzheimer's disease," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 463–473.
[39] J. S. Supancic, III, and D. Ramanan, "Self-paced learning for long-term tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2379–2386.
[40] F. Ma, D. Meng, Q. Xie, Z. Li, and X. Dong, "Self-paced co-training," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 2275–2284.
[41] D. Meng, Q. Zhao, and L. Jiang, "A theoretical understanding of self-paced learning," *Inf. Sci.*, vol. 414, pp. 319–328, Nov. 2017.
[42] C. Xu, D. Tao, and C. Xu, "Multi-view self-paced learning for clustering," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 3974–3980.
[43] D. Greene and P. Cunningham, "A matrix factorization approach for integrating multiple data views," in *Machine Learning and Knowledge Discovery in Databases*, W. Buntine, M. Grobelnik, D. Mladenić, and J. Shawe-Taylor, Eds. Berlin, Germany: Springer, 2009, pp. 423–438.
[44] F. Nie, J. Li, and X. Li, "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1881–1887.

ZONGMO HUANG is currently pursuing the bachelor's degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His current research interests include multi-view clustering and self-paced learning.

YAZHOU REN received the B.Sc. degree in information and computation science and the Ph.D. degree in computer science from the South China University of Technology, Guangzhou, China, in 2009 and 2014, respectively. He visited the Data Mining Laboratory, George Mason University, USA, from 2012 to 2014. He is currently an Associate Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. He has published more than 30 research articles. His current research interests include clustering, self-paced learning, and deep learning.

WENLI LIU is currently pursuing the bachelor's degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His current research interests include unsupervised learning and multi-view learning.

XIAORONG PU received the Ph.D. degree in computer application from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2007. She is currently a Professor with the School of Computer Science and Engineering, UESTC. Her current research interests include neural networks, computer vision, computer-aided diagnosis (CAD), e-Health, and machine learning.