World Wide Web (2020) 23:781–798
https://doi.org/10.1007/s11280-019-00723-8

Semi-supervised clustering with deep metric learning and graph embedding

Xiaocui Li[1] · Hongzhi Yin[2] · Ke Zhou[1] · Xiaofang Zhou[2]

Received: 10 June 2019 / Revised: 8 July 2019 / Accepted: 15 August 2019 /
Published online: 24 August 2019

Abstract

As a common technology in social network, clustering has attracted lots of research interest due to its high performance, and many clustering methods have been presented. The most of existing clustering methods are based on unsupervised learning. In fact, we usually can obtain some/few labeled samples in real applications. Recently, several semi-supervised clustering methods have been proposed, while there is still much space for improvement. In this paper, we aim to tackle two research questions in the process of semi-supervised clustering: (i) How to learn more discriminative feature representations to boost the process of the clustering; (ii) How to effectively make use of both the labeled and unlabeled data to enhance the performance of clustering. To address these two issues, we propose a novel semi-supervised clustering approach based on deep metric learning (SCDML) which leverages deep metric learning and semi-supervised learning effectively in a novel way. To make the extracted features of the contribution of data more representative and the label propagation network more suitable for real applications, we further improve our approach by adopting triplet loss in deep metric learning network and combining bedding with label propagation strategy to dynamically update the unlabeled to labeled data, which is named as semi-supervised clustering with deep metric learning and graph embedding (SCDMLGE). SCDMLGE enhances the robustness of metric learning network and promotes the accuracy of clustering. Substantial experimental results on Mnist, CIFAR-10, YaleB, and 20-Newsgroups benchmarks demonstrate the high effectiveness of our proposed approaches.

Keywords Clustering · Semi-supervised learning · Deep metric learning · Graph embedding · k-means

This article belongs to the Topical Collection: *Special Issue on Graph Data Management in Online Social Networks*
Guest Editors: Kai Zheng, Guanfeng Liu, Mehmet A. Orgun, and Junping Du

✉ Hongzhi Yin
h.yin1@uq.edu.au

✉ Ke Zhou
k.zhou@hust.edu.cn

Extended author information available on the last page of the article.

1 Introduction

Data mining has become a research hotspot of great concern to researchers for decades because of its significance in various application fields. Cluster analysis, as one of the most important technologies of data mining, has been developing various algorithms [2, 35, 49, 54] continuously, which is widely applied in a variety of application scenarios, such as social network analysis [9, 32, 33, 53], community detection [15, 45, 50, 57, 60], computer vision [34, 38, 46], natural language processing [24, 28] and knowledge discovery [20, 58, 59]. Clustering and classification are the most two important categories of machine learning, and their major difference is whether supervised learning or not. The objective of clustering is to pull similar data points (according to specific metric in extracted feature space) into the same clusters, while those data points with highly distinct features will be far apart.

Initially, clustering only categories the unlabeled data, which is a branch of unsupervised learning. The unsupervised clustering technique has drawn a tremendous amount of research attention, and many clustering methods have been proposed [11, 12, 19, 51, 55] in the past. These clustering methods can be generally categorized into three types: (1) Feature learning based methods. This kind of methods tries to find more discriminative features by combining with data dimension reduction technique [39, 55] or subspace learning technique [1, 11]. (2) Metric learning based methods. These methods aim to learn an appropriate distance metric for the training data. Under the learned distance metric, it can group similar samples together and separate dissimilar samples apart at the same time [19, 22, 42]. (3) Graph based clustering. This kind of methods partitions the data into different classes according to their pairwise similarities [10, 48]. Recently, deep learning technique has achieved great success in many fields due to its superiority of learning capacity, and some deep learning based methods [26, 40, 52] have been used to solve clustering problems.

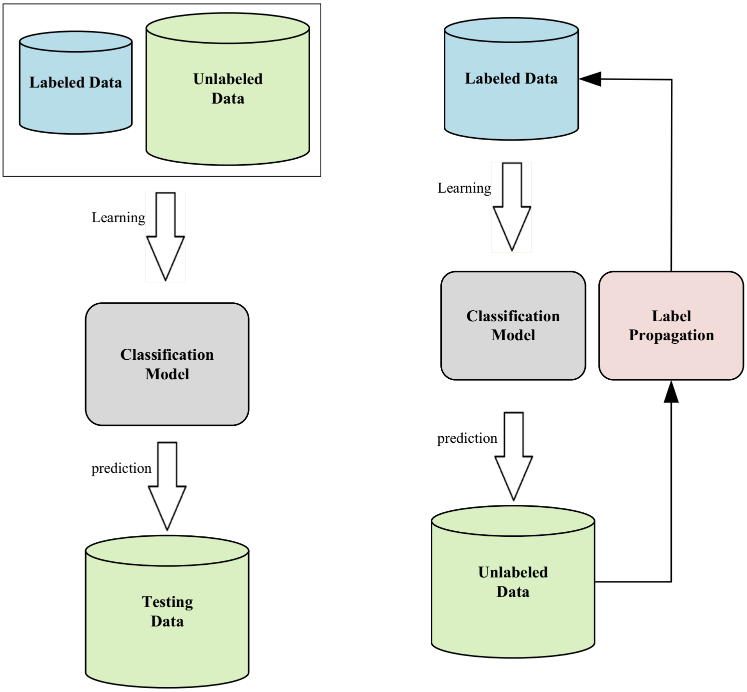
Generally speaking, how to extract useful features and learn an appropriate metric for high-dimensional data without any supervised information is a challenging task. Consequently, some supervised clustering algorithms[13, 17, 56] have been proposed to improve the clustering result. However, most of these methods have great limitations in real practical applications, because it is almost impossible for all data having labels. At the same time, tagging enough sample manually is a waste of human resources and time, and it is also unrealistic. In fact, in most of the real-world applications, we can only obtain limited labeled data while most of the data are unlabeled. Based on the above problems, semi-supervised based clustering methods [4, 14, 44] have emerged more recently. These methods adjust the learning framework through limited label data, so that the clustering process can be executed in the supervised framework, which greatly improve the clustering performance and have widely applicability.

1.1 Motivation

Although the existing semi-supervised clustering algorithms have achieved good results, there are still two important issues that will hinder the performance of clustering. (i) Most of these methods extract features or learn a distance metric through traditional SVM, neural networks or linear mapping, which limits its performance. (ii) They only use the labeled data to guide the process of the clustering, so they can not make full use of the traits of data especially unlabeled data.

Inspired by these problems, we propose a semi-supervised clustering with deep metric learning (SCDML), which can extract discriminative features by using deep metric learning model. At the same time, the unlabeled data is also used to optimize clustering result through k-nearest neighbors label updating strategy to dynamically increase the labeled data set, and then it can promote the performance of the metric learning network. Figure 1a illustrates the existing semi-supervised clustering models which trains the network model with fixed input, while in our network model, the model will be constantly improved by updating the labeled data incrementally, as shown in Figure 1b.

In order to further improve the performance of SCDML, through extensive analysis and experimental results, we found that: (i) SCDML takes Siamese CNNs as the metric learning network, in which the contrastive loss function is used to optimize the network. The main objective of contrastive loss is to reduce the distance between positive samples and increase the distance between negative samples. However, contrastive loss treats positive sample and negative sample equally while ignores the difficulty in metric learning brought by negative sample. (ii) In the process of labeling propagation, the k-nearest neighbors of cluster center are tagged as the new labeled data in each cluster, which doesn't fully utilize the results of deep metric learning network. In addition, the parameter k is difficult to be predefined, so it is crude to select k unlabeled data nearest to the center of the labeled data from each cluster. Therefore, we will further improve the performance of our SCDML approach based on the above two aspects.



(a) Existing Semi-supervised Models (b) Label Propagation Model

Figure 1 The difference between existing semi-supervised learning methods and our proposed label propagation model

1.2 Contributions

The key contributions of our work can be summarized as below:

(1) In this work, we design a novel semi-supervised clustering model, which includes a semi-supervised deep metric learning subnetwork and a labeling propagation subnetwork. To the best of our knowledge, the proposed method is a pioneer to address clustering task by combining deep metric learning with semi-supervise learning techniques.
(2) In the metric learning subnetwork, we integrate the Siamese CNNs to extract discriminative features to minimize the cluster error.
(3) In the labeling propagation subnetwork, we design a k-nearest neighbors label updating strategy to transform the unlabeled data into labeled data. As a result, it can reinforce the ability of metric learning network.
(4) We have conducted extensive experiments on three datasets to demonstrate the effectiveness of our proposed approach. Experimental results show that our approach is a robust competitor for the most state-of-the-art clustering methods.

Note that we presented our preliminary study of deep semi-supervised clustering in the prior work [29] as an abstract paper. In this article, we make significant revision and add substantial new materials compared with the prior work. Specifically, this article makes the following new contributions:

(1) We provide a systematically analysis of SCDML and also a more comprehensive review of the related work.
(2) To obtain more discriminative and robust features, we improve our model by applying the triplet loss as the metric learning network's loss function instead of contrastive loss. The triplet CNNs takes three labeled samples (an anchor, a positive sample and a negative sample) as an input. Under the triplet loss function, the positive sample can be pulled closer to anchor point while the negative sample will be pushed away from the anchor at the same time. As a result, all the labeled data can be clustered in learned feature space.
(3) To reinforce the reliable of new labeled data, we propose a more reasonable and effective labeling propagation network. Specifically, by combining the result of classification network and the result of our improved graph clustering algorithm, the unlabeled data can be transformed from weak labeled data into strong labeled data. The new added strong labeled data can positively forward the deep metric learning and classification network, and then improve the accuracy of metric learning network.
(4) We have conducted extensive experiments on four datasets and compared our approaches with more competing methods. In addition, we evaluate the effectiveness of our approaches with its two variants and have provided more verification experiments.

2 Related work

2.1 Clustering methods

In this subsection, we briefly introduce the background of clustering methods, including features learning based methods, metric learning based methods and deep learning based methods.

Features based clustering Features based clustering divides the dataset to clusters according the data's features. The k-means [18] clustering algorithm is a classical features based unsupervised feature learning. This method aims to minimize the following objective function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^j - c_j||^2$$

where $||x_i^j - c_j||^2$ indicates the Euler distance between the data point x_i^j and the cluster center c_j.

Many more efficient varieties of k-means were proposed in the last few decades. Saha et al. [39] proposed a useful model, which performs clustering according to the feature selection and the fuzzy data simultaneously. In literature [55], an adaptive hashing method based on feature clustering is proposed to reduce data dimension.

Metric learning based clustering Metric learning can learn the distance metric function for a specific task autonomously according to different tasks. A common metric distance function is defined as follows:

$$d_M(x, x') = \sqrt{(x - x')^T M (x - x')}$$

where $M \in \mathbb{R}^{d \times d}$ is called the metric matrix which is the inverse of covariance matrix $\sum$. Obviously, M is a symmetric matrix.

Kalintha et al.[22] proposed a non-linear transformation of distance matric learning for clustering, which performs well on non-linear separable data. Heidari et al. [19] proposed a probabilistic model, which combines the fuzzy clustering and metric learning to maximize the distance between clusters and minimize the intra-cluster distance.

Graph based methods As one of the most popular clustering techniques recently, graph clustering has attracted lots of researchers and various graph based clustering methods were proposed[30, 31, 46, 47]. These methods represent entities as vertices in an undirected graph with weighted edges to describe the relationships between entities. In [10], Chen et al. proposed a sparse representation method for graph clustering. In [48], Xie et al. proposed a multi-view graph clustering with global and local graph embedding.

Deep learning based methods These methods can learn more discriminative and robust features by using convolutional neural networks(CNNs) [6, 7, 21]. In [40], Sekmen et al. combined subspace clustering with CNNs to train a deep subspace clustering model. In [27], a nonlinear embedding model is proposed to learn a new representation of examples, so that elements in the same category are organized into the same cluster. In [8], Chen et al. proposed a deep nonparametric clustering method, in which deep learning is used for feature extraction and dimension reduction. Compared with these methods, our proposed approach is semi-supervised, which can improve the performance of clustering with supervised information.

2.2 Semi-supervised learning

Semi-supervised learning is a machine learning technique to improve the performance of the trained model [4, 25, 44]. Different from unsupervised learning, semi-supervised learning train model by utilizing few labeled sample and abundant unlabeled data. Guan et al. [14] proposed a feature space learning model based on semi-supervised framework to better understand and learn feature space. In [37], Laine et al. proposed a temporal ensemble model for semi-supervised learning. In [43], a local density model is proposed to measure the similarity between k-nearest vertex. Kang et al. [23] combined multiple kernel learning with semi-supervised technique to tackle clustering problem. Compared to these traditional semi-supervised learning based clustering methods, our approach can learn more meaningful and discriminative features, which are beneficial to the following clustering.

Recently, some deep semi-supervised based clustering methods have been proposed [3, 36, 41]. In [3], Arshad proposed a semi-supervised deep fuzzy C-mean clustering (DFCM). In [36], Ren et al. proposed a semi-supervised deep embedded model for clustering. In [41], a ClusterNet model is designed by Shukla et al, which uses pair-wise semantic constraints to drive the clustering approach. However, our approach is different from these methods in two aspects. (i) We can make full use of the unlabeled data instead of only utilization for regularization. (ii) We employ label propagation strategy to tag more unlabeled data, while in these methods the number of labeled data is fixed.

3 SCDML

To extract more discriminative features for optimizing the clustering model, we apply the Siamese CNNs and take the contrastive loss as the metric learning's loss function. We also propose the k-nearest neighbors label updating strategy to dynamically transform the unlabeled data into labeled data, which can give full play to the contribution of unlabeled data.

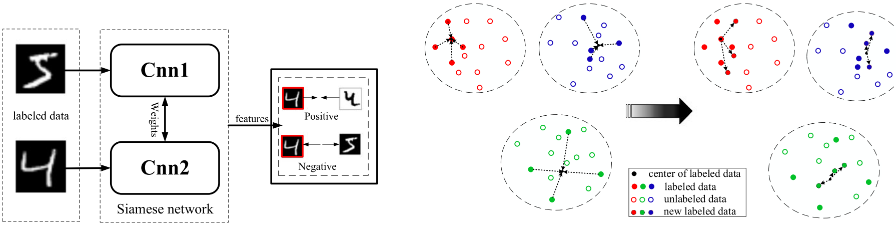
3.1 semi-supervised deep metric learning network

We design a semi-supervised deep metric learning network based on Siamese CNNs, as shown in Figure 2a.

First feed labeled sample pairs to Siamese CNNs to extract discriminable features. In the features learning process, we take the contrastive loss as the objective function of our network. The loss function can be computed as follows:

$$L = y||x_1 - x_2||_2^2 + (1-y)max\left(\alpha - ||x_1 - x_2||_2^2, 0\right) \tag{1}$$

where $||x_1 - x_2||_2$ is the Euclidean Distance between x_1 and x_2. x_1 and x_2 represent the features of input pair samples extracting by metric learning network respectively. $y \in \{0, 1\}$



(a) semi-supervised classification based on metric learning

(b) k-nearest neighbors label updating (k=3)

Figure 2 Illustration of the SCDML approach. The approach consists of two steps: (1) extract discriminative features by Siamese CNNs (the left); (2) obtain more labeled data by k-nearest neighbors algorithm (the right)

(1 if the input pair is from the same class, and 0 if the input pair is from the different classes.) is the corresponding label of input pair samples.

Then encode all the data including labeled data and unlabeled data through the trained metric learning network to obtain their features.

Finally, classify the unlabeled data according to the encoded features, and record the classification results as the label of the unlabeled data.

3.2 *k*-nearest neighbors label updating strategy

In this subsection, we propose a k-nearest neighbors label updating strategy to transform the unlabeled data into labeled data.

As discussed above, all the data are classified in to C clusters and each cluster contains limited labeled data while a lot of unlabeled data. To make full use of the features of unlabeled data, we add $k * C$ new unlabeled data to the labeled dataset each time. The main process of k-nearest neighbors label updating strategy is as follows.

Step 1: Compute the center of each cluster according to the labeled data.

$$c_i = \frac{1}{N_{c_i}^l} \sum_{j=1}^{N_{c_i}} \left\{ \left(s_j^l, l_j \right) | l_j = i \right\} \tag{2}$$

where s_j^l is the labeled data, $N_{c_i}^l$ is the number of labeled samples in cluster c_i, N_c is the number of cluster, l_j is the label of sample s_j^l.

Step 2: Search the k nearest unlabeled data from the center of labeled data in each cluster, and then update their attributes from unlabeled data to labeled data. The new added labeled data Δ in cluster c_i can be computed by:

$$\Delta S = Sort\left(\left\{ Dis\left(\left(s_j^u, l_j \right) | l_j = i, c_i \right) \right\}, k \right) \tag{3}$$

where $(s_j^u, l_j)|l_j = i$ indicates the unlabeled data s_j^u in the i^{th} cluster, $Dis(,)$ is the distance function, $Sort(X, k)$ indicates sorting the elements of X by ascending order and return the top k elements.

For example, in Figure 2b the solid points represent labeled data, and the hollow points represent unlabeled data. After finding each cluster's center of labeled data, each cluster generate three new labeled data which are the nearest unlabeled samples in this center.

As the number of labeled data increases, our proposed metric model can learn more robust and discriminative features, which will further improve the accuracy of the clustering.

4 Improved semi-supervised clustering with deep metric learning

As discussed in the motivation of Section 1, there are still two factors that will affect the performance of clustering: (i) The selection of metric function will influence the accuracy of data feature extraction, then further affect the accuracy of clustering results; (ii) In practical applications, the k-nearest neighbors label updating strategy is not very suitable, due to the different density of each cluster, the number of labeled data and their distribution in each cluster. Moreover, the choice of parameter k also hinders the effectiveness of the algorithm. To enhance the performance of SCDML for the semi-supervised clustering, we improve our SCDML approach from the following two aspects: (i) We take the triplet CNNs as the metric learning model and employ the triplet loss function as the model's loss to train the network. (ii) We design a more reasonable label propagation network to transform the unlabeled data into labeled data dynamically.

The framework of improved semi-supervised clustering with deep metric learning (SCDMLGE) is shown in Figure 3, which contains two subnetworks: a semi-supervised deep metric learning and classification network, and a labeling propagation network. The following subsections will present the details of our proposed approach.

4.1 Semi-supervised deep metric learning and classification network

Unlike the metric learning network used in previous work, we applied triplet network in this work, which contain an anchor, a positive sample and a negative sample. As we discussed above, contrastive loss treats positive sample and negative sample equally while ignores the difficulty in metric learning brought by negative sample, so we introduce the triplet loss function into our CNN model, which pushes away the negative samples from the anchor and pulls the positive samples closer to the anchor.

After training by the triplet metric learning network, the distance between anchor and the positive sample will be shorten and the negative sample will be pushed away from the anchor simultaneously. Therefore, clusters can be better formed in this feature space.

The main training process of the network consists of the following three steps.

Step 1: Train the network with the labeled triplets. First, extract discriminable features through the triplet CNNs, then use the features to train a classifier. To train the feature extracting network and classification network at the same time, we design the loss function for semi-supervised deep metric learning and classification network as follows:

$$\min L = L_M + \lambda_1 L_C + \lambda_2 \|W\|_F^2, \tag{4}$$

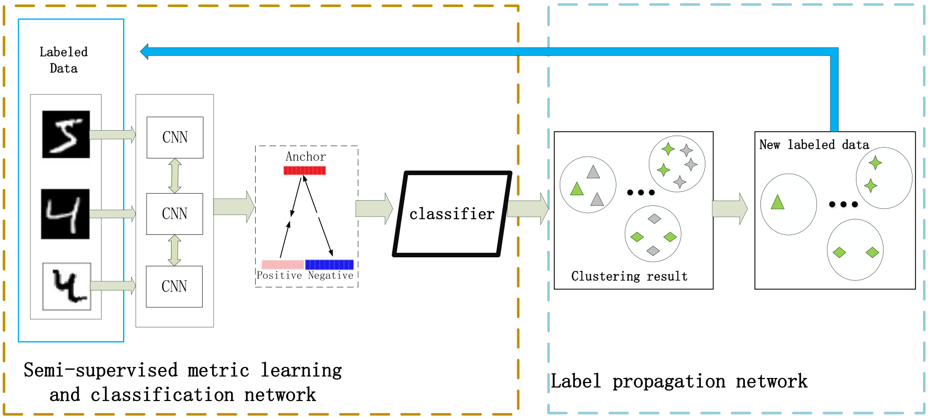


Figure 3 The framework of the clustering with deep semi-supervised metric learning. The framework consists of two subnetworks: (1) a feature extraction subnetwork by Triplet CNNs (the left); (2) a label propagation subnetwork by graph clustering algorithm (the right)

where, λ_1 and λ_2 are a tunable positive parameter. $\|W\|_F^2$ is a regular term to prevent overfitting. L_M and L_C are metric learning loss and classification loss, respectively. They can be computed as follows:

$$L_M = \frac{1}{N}\sum_{i=1}^{N}\{max(||f(x_i^a) - f(x_i^p)||_2 \\ -||f\left(x_i^a\right) - f(x_i^n)||_2 + \alpha, 0)\} \tag{5}$$

where $f(x_i^a)$, $f(x_i^p)$ and $f(x_i^n)$, indicate the feature of anchor, positive sample and negative sample respectively. α is the minimum margin of $||f(x_i^a) - f(x_i^p)||_2$ and $||f(x_i^a) - f(x_i^n)||_2$.

$$L_C = -\sum_{f(x)} p(f(x)) \log q(f(x)) \tag{6}$$

where $p(f(x))$ is the expected outputs, and $q(f(x))$ is the actual outputs of the classification network.

Step 2: Encode the labeled and unlabeled data. Assume that $S_l = \{(s_{li}, l_{li})|i = 1, 2, \ldots, N_l\}$ and $S_u = \{(s_{ui}|i = 1, 2, \ldots, N_u\}$ separately represent the init labeled data and unlabeled data, where N_l is the number of labeled samples, N_u is the number of unlabeled samples, and $l_{li} \in \{1, 2, \ldots, C\}$, where C is the number of classes. We use $S_l' = \{s_{li}'|i = 1, 2, \ldots, N_l\}$ and $S_u' = \{s_{ui}'|i = 1, 2, \ldots, N_u\}$ represent the outputs of the S_l and S_u by CNNs.

Step 3: Tag the unlabeled data according to the classification network. Therefore, S_u can be denoted as $S_u = \{(s_{ui}, l_{ui}^1)|i = 1, 2, \ldots, N_u\}$, where l_{ui}^1 is the classification label of the s_{ui}.

4.2 Semi-supervised clustering labeling propagation network

Through the deep metric learning and classification network, we can obtain the label of the unlabeled data, as called weak label. To acquire the strong label of the unlabeled data, we design a semi-supervised labeling propagation network. It includes two parts: semi-supervised clustering and labeling propagation.

In the process of the semi-supervised clustering, we propose an improved graph clustering algorithm. The details of the algorithm are as follows:

Firstly, we compute the similarity matrix W according to the following equation:

$$w_{ij} = \begin{cases} exp\frac{-||x_i - x_j||^2}{2\sigma^2} & \{x_i, x_j\} \in S_u' \\ 1 & \{x_i, x_j\} \in S_l' \wedge \{l_{x_i} = l_{x_j}\} \\ 0 & \{x_i, x_j\} \in S_l' \wedge \{l_{x_i} \neq l_{x_j}\} \end{cases} \tag{7}$$

where σ represents the neighborhood width of the sample points, i.e., the larger the σ, the greater the similarity between the sample points.

Secondly, we use the following formula to calculate the degree matrix D:

$$d_i = \sum_{j=1}^{n} w_{ij}, \tag{8}$$

and then we can obtain the corresponding Laplacian matrix.

$$L = D - W, \tag{9}$$

Next, we use the top k eigenvectors $u_1, u_2, \ldots, u_k$ of L to form a new matrix U. And then, we obtain the clustering results by using k-means clustering algorithm.

At last, we mark the S'_u according to the clustering results, and record as $S'_u = \{(s'_{ui}, l^2_{ui})|i = 1, 2, \ldots, N_2\}$, where l^2_{ui} is the clustering label of the s'_{ui}.

When both the classification label and clustering label of the unlabeled data S_u are obtained, we can implement labeling propagation strategy. Assume that ΔS represents newly added strong label data, it can be acquired by:

$$\Delta S = \left\{ s_{ui} | \left(l^1_{ui} = l^2_{ui} \right) \right\}, \tag{10}$$

According (10), we can update S_l and S_u untill all of the unlabeled data transform into labeled data.

$$\begin{aligned} S_l &= S_l + \Delta S \\ S_u &= S_u - \Delta S \quad , \end{aligned} \tag{11}$$

Algorithm 1 summarizes the main process of our SCDMLGE approach. It trains a classifier using the labeled data through the semi-supervised deep metric learning and classification network, and then obtains the classification label of unlabeled data L^1_u (line 3∼4). We get the clustering label of unlabeled data L^2_u by applying our improved graph clustering, then compared L^1_u with L^2_u to update the labeled dataset(line 5∼7). This algorithm terminates until all the unlabeled data are transformed into strong label data, or the current iteration error is less than the minimum threshold ε, or the number of iterations reaches the maximum iteration value T.

Algorithm 1 SCDMLGE.

```
Require: labeled data $S_l = \{(s_{li}, l_{li})$ and unlabeled data $S_u = \{s_{ui}\}$; Parameters: $\lambda_1, \lambda_2$,
    the minimum iterative error $\varepsilon$ and the maximum iterative number $T$.
Ensure: clustering results $C_i$
 1: Initialize the parameters of deep metric and classification network;
 2: for $i = 1, 2, \ldots, T$ do
 3:     Generate triplet samples $I = \{(a, p, n)|a, p, n \in S_l \;\&\&\; ||a - p||_2 < ||a - n||_2\}$;
 4:     Train the network with generated triplet samples $I$ ;
 5:     Obtain all the deep features $S_l{}'$ and $S_u{}'$of labeled data $S_l$ and unlabeled data $S_u$;
 6:     Obtain the classification label $L^1_u$ of unlabeled data $S_u$ according $S_u{}'$ ;
 7:     Obtain the clustering result $L^2_u$ of unlabeled data $S_u$ according $S_u{}'$ ;
 8:     Compute the added strong label data $\Delta S$ according to (11);
 9:     Compute the gradients according to (4);
10:     Update the parameters of deep metric and classification network;
11:     Compute the value of loss function $L_i$ by (4);
12:     If $\Delta S = 0$ or $|L_i - L_{i-1}| < \varepsilon$, go to Output.
13: end for
```

5 Experiments

5.1 Datasets and compared methods

Datesets We implement experiments on four publicly available datasets including: Mnist, CIFAR-10 [36], YaleB [11] and 20-Newsgroups [8]. The Mnist dataset consists of 70000 images of hand-written digits from 0 to 9, and widely used for character recognition. The CIFRA-10 dataset consists of 60000 images with 10 categories, and each category includes 6000 samples. The YaleB dataset has 2414 grayscale face images including 38 persons. Each person has 64 samples captured from five different angles. The 20-Newsgroups dataset is often used in text and document classification, and it contains 18846 documents labeled into 20 categories.

Compared methods To evaluate the efficacy of the proposed approaches, we compare our approaches with some state-of-the-art related methods including:

(1) traditional unsupervised based methods: FCH [55], SC-CNMF [11];
(2) traditional semi-supervised (supervised) methods: FSLSC [14], SMKL [23];
(3) deep unsupervised based methods: DCN [5], IDEC [16];
(4) deep semi-supervised based methods: DFCM [3], SDEC [36], ClusterNet [41].

5.2 Evaluation measures and experimental settings

To evaluate the performance of our proposed methods and compared methods, we use two types of measures, namely clustering accuracy (AC), normalized mutual information (NMI).

Table 1 Clustering performance on Mnist, CIFAR-10, YaleB and 20-Newsgroups datasets (the percentage of labeled data is 10%). The best results are shown in bold

Methods	Mnist	CIFAR-10	YaleB	20-Newsgroups
AC(%)				
FCH	66.5	22.4	52.1	31.6
SC-CNMF	68.4	23.2	62.5	30.2
FSLSC	75.2	23.9	59.4	38.6
SMKL	78.3	24.4	59.6	42.5
DCN	81.1	26.3	63.8	49.2
IDEC	88.1	25.0	67.7	50.5
DFCM	90.4	31.4	74.6	52.7
SDEC	89.4	30.3	74.8	**78.1**
ClusterNet	**98.9**	38.5	79.5	67.7
SCDML	92.3	32.7	78.3	57.1
SCDMLGE	98.3	**40.3**	**81.1**	73.7
NMI(%)				
FCH	65.7	14.6	50.6	30.5
SC-CNMF	66.5	14.2	60.2	28.6
FSLSC	73.2	15.8	58.7	36.7
SMKL	76.3	16.3	58.6	40.3
DCN	75.7	17.0	62.4	44.7
IDEC	86.7	17.3	65.3	45.8
DFCM	90.2	21.5	73.6	50.4
SDEC	87.6	23.4	74.1	48.3
ClusterNet	97.0	31.8	77.6	58.6
SCDML	91.9	24.9	75.6	54.1
SCDMLGE	**97.8**	**33.1**	**79.2**	**59.4**

These two measures are widely used to evaluate the performance of clustering in many researches [5, 16, 42, 55].

AC can be computed as follows:

$$AC = \frac{1}{N}\sum_{i=1}^{K}\max(C_i|L_i), \tag{12}$$

where N is the number of samples to be clustered. K is the number of clusters. L_i is the true label information, and C_i is the predicted label information by clustering algorithm.

NMI can be computed as follows:

$$NMI(A, B) = \frac{MI(A, B)}{\sqrt{H(A)H(B)}}, \tag{13}$$

where A is the true cluster set, and B is the predicted cluster set. $MI(A, B)$ is the mutual information between A and B. $H(A)$ and $H(B)$ denote the entropies of A and B. The range of NMI is from 0 (A is independent from B) to 1 (A is equivalent to B).

5.3 Results and analysis

5.3.1 Clustering performance evaluation

In this subsection, we conduct experiment to evaluate the clustering performance of our proposed semi-supervised clustering with deep metric learning approach named SCDML,

Table 2 AC results of proposed methods and three semi-supervised clustering methods with different percentages of labeled data on Mnist, YaleB, CIFAR-10 and 20-Newsgroups datasets. The best results are shown in bold

Datesets	Percentages	FSLSC	SMKL	DFCM	SDEC	ClusterNet	SCDML	SCDMLGE
Mnist	0.5%	65.8	72.9	84.0	83.5	96.8	86.1	**97.1**
	1%	67.1	73.5	84.6	83.8	**98.1**	87.5	97.3
	2%	67.3	74.2	85.7	84.5	**98.3**	88.4	97.4
	5%	69.0	75.5	87.4	86.1	**98.6**	89.6	97.7
	10%	75.2	78.3	90.4	89.4	**98.9**	92.3	98.3
CIFAR-10	0.5%	21.5	22.8	29.7	28.2	37.1	31.1	**38.5**
	1%	21.7	23.3	30.2	28.4	37.2	31.3	**38.8**
	2%	22.6	23.7	30.5	28.6	37.7	31.6	**39.2**
	5%	22.8	24.0	30.7	30.1	38.2	31.8	**39.7**
	10%	23.9	24.4	31.4	30.3	38.5	32.7	**40.3**
YaleB	0.5%	50.3	48.6	65.5	70.8	77.3	74.2	**78.3**
	1%	50.4	48.7	66.2	71.2	77.6	74.5	**79.6**
	2%	51.9	50.5	66.3	71.7	78.5	74.9	**80.0**
	5%	52.8	54.4	68.8	72.3	78.8	75.7	**80.4**
	10%	58.7	59.6	73.8	74.8	79.5	78.3	**81.1**
20-Newsgroups	0.5%	28.3	36.7	47.6	46.3	62.7	51.8	**66.5**
	1%	28.6	36.9	48.4	46.9	63.5	52.3	**66.8**
	2%	28.8	37.2	49.1	47.5	63.8	52.7	**67.2**
	5%	29.1	38.6	50.3	49.2	64.6	53.5	**69.7**
	10%	38.6	42.5	52.7	**78.1**	67.7	57.1	73.7

and its improved version named SCDMLGE. Table 1 shows the clustering results on Mnist, CIFAE-10, YaleB and 20-Newsgroups datasets. According to the experimental results, we observe that: (i) our proposed SCDMLGE outperforms all of state-of-the-art methods. (ii) SCDML can achieve better performance than most of compared methods.

Specifically, compared with traditional clustering methods FCH, SC-CNMF, FSLSC, SMKL, our approaches can learn more meaningful and robust features by using deep metric learning. Moreover, FCH and SC-CNMF are unsupervised methods and the label information is not used in the process of clustering, which further weaken the performance of them. Compared with deep clustering methods DCN, IDEC, DFCM and SDEC, the reasons for the performance improvement as follows: DCN and IDEC ignore the utilization of information of labeled data. The unlabeled data is only use for regularization in DFCM, which limits the performance of deep metric learning. SDEC adopts the pairwise constraints to lead the direction of clustering, which is similar to contrastive loss. In addition, we can see that ClusterNet can outperforms other methods except our SCDMLGE approach.

From the results of last two rows in Table 1 we can confirm that SCDMLGE is superior to SCDML. As SCDMLGE takes the triplet CNNs to train the deep metric network, it can extract more discriminative features than Siamese CNNs adopted in SCDML. Better yet SCDMLGE designs an improved labeling propagation network which is more reasonable to transform unlabeled data into labeled data, and then make full use of the contribution of unlabeled data to optimize of classification model.

Table 3 NMI results of proposed methods and three semi-supervised clustering methods with different percentages of labeled data on Mnist, YaleB, CIFAR-10 and 20-Newsgroups datasets. The best results are shown in bold

Datesets	Percentages	FSLSC	SMKL	DFCM	SDEC	ClusterNet	SCDML	SCDMLGE
Mnist	0.5%	65.6	69.6	86.2	85.3	95.2	85.7	**95.7**
	1%	65.8	70.0	86.7	85.5	95.6	86.3	**96.1**
	2%	66.5	71.3	87.3	85.8	95.7	86.5	**96.5**
	5%	68.5	72.4	88.6	86.5	96.3	88.7	**96.8**
	10%	73.2	76.3	90.2	87.6	97.0	91.9	**97.8**
CIFAR-10	0.5%	14.1	14.6	20.3	21.7	**29.5**	22.3	29.4
	1%	14.1	14.8	20.4	21.9	29.7	22.7	**30.0**
	2%	14.3	15.2	20.7	22.2	**30.6**	23.1	**30.4**
	5%	15.6	15.8	21.1	22.5	31.3	23.4	**31.5**
	10%	15.8	16.3	21.5	23.4	31.8	24.9	**33.1**
YaleB	0.5%	51.7	53.2	68.1	70.2	75.7	71.4	**76.1**
	1%	52.2	53.4	68.3	70.6	76.2	71.7	**76.5**
	2%	52.5	53.9	68.5	71.3	76.8	72.4	**77.4**
	5%	53.5	54.4	70.5	72.6	76.9	73.8	**77.7**
	10%	58.7	58.6	73.6	74.1	77.6	75.6	**79.2**
20-Newsgroups	0.5%	24.6	31.7	45.2	44.5	**55.9**	50.6	55.6
	1%	24.6	32.2	45.7	44.9	56.5	51.3	**56.9**
	2%	24.8	32.4	46.6	45.6	56.8	51.7	**57.7**
	5%	25.1	33.3	48.3	47.5	58.3	53.5	**58.8**
	10%	36.7	40.3	50.4	48.3	58.6	54.1	**59.4**

5.3.2 Clustering performance evaluation with different percentages of labeled data

To further evaluate the clustering performance of our proposed approaches, we increase the percentage of labeled data from 0.5% to 10%. Tables 2 and 3 separately report the AC and NMI results of our proposed SCDML and SCDMLGE approaches and five semi-supervised clustering methods on four datasets. From the results of Tables 2 and 3, we can obviously see that our SCDMLGE approach performs better than all compared semi-supervised clustering methods, which indicates that SCDMLGE can learn better discriminative structure features, and simultaneously make full use of the unlabeled data.

5.3.3 Evaluation of the Influence of parameters

This subsection focuses on evaluating the impact of important parameters in SCDMLGE(λ_1, λ_2), and we take Mnist dataset as a example. For parameter λ_1 and λ_2, we separately observe the performance variations of SCDMLGE in the change interval of [0.1,1] with the step size of 0.1 and [0.01,0.1] with the step size of 0.01. From the results in Figure 4a and b, SCDMLGE can reach a stable and good clustering performance when λ_1 is in [0.4,0.7] and λ_2 is in [0.02,0.05]. In addition, other datasets can observe similar results.

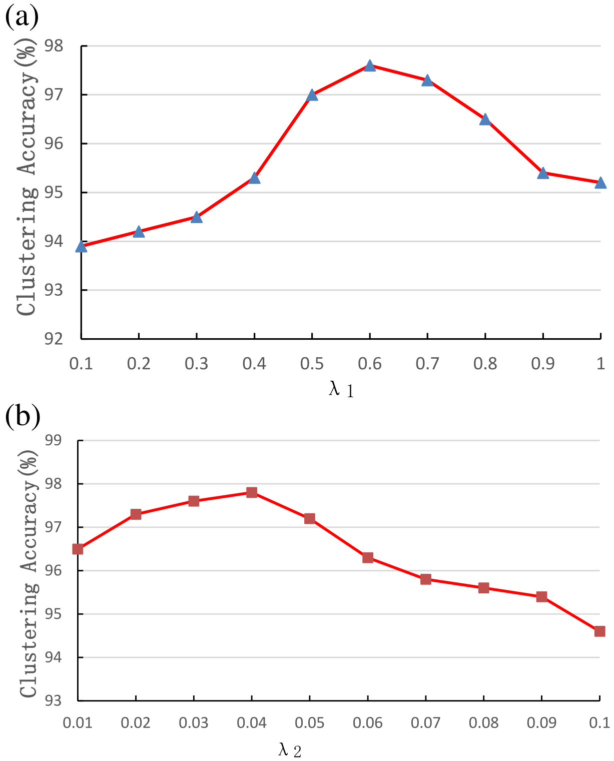


Figure 4 AC results of SCDMLGE versus different values of **a** parameter λ_1, **b** parameter λ_2

Table 4 Clustering performance of our proposed improving strategies on Mnist, YaleB and 20 Newsgroups datasets (the percentage of labeled data is 10%)

Methods	Mnist	CIFAR-10	YaleB	20-Newsgroups
AC(%)				
SCDML	92.3	32.7	78.3	57.1
SCDML+t	95.6	36.9	79.5	68.4
SCDML+p	96.5	37.1	80.0	67.2
SCDMLGE	98.3	40.3	81.1	73.7
NMI(%)				
SCDML	91.9	24.9	75.6	54.1
SCDML+t	94.2	29.4	78.1	57.5
SCDML+p	95.7	28.3	77.7	57.9
SCDMLGE	97.8	33.1	79.2	59.4

5.4 Effectiveness of new strategies

SCDMLGE is the improved version of SCDML which mainly takes two new strategies. In order to evaluate the effectiveness of these two improvements separately, we generate two modified versions of SCDML: (1) "SCDML+t". A variant version of SCDML by employing triplet CNNs as deep metric learning model; (2) "SCDML+p". A variant version of SCDML by using the new labeling propagation network to dynamically increase the labeled data.

Table 4 shows the effectiveness of our proposed new strategies. From the experimental results, we can see that the clustering performance of SCDML+t and SCDML+p are better than SCDML, which means that our proposed new strategies in SCDMLGE are beneficial to improve the performance of clustering.

6 Conclusion

In this paper, we propose a novel semi-supervised clustering with deep metric learning approach(SCDML) to address the problem of extracting more discriminative features with deep metric learning network and making full use of the unlabeled data features. In order to further improve the effectiveness and practicability of SCDML, we propose an improved semi-supervised clustering related to SCDML, named SCDMLGE, which embeds triplet CNNs in deep metric learning network instead of siamese network and comprises a new labeling propagation network simultaneously. The semi-supervised deep metric learning network adopted triplet loss function can extract more powerful features, and then learn a more discriminative metric. After that, labeling propagation network is used to label new data which is more suitable for real applications. Experimental results on Mnist, CIFAE-10, YaleB and 20-Newsgroups datasets have shown the high performance and effectiveness of our proposed semi-supervised clustering with deep metric learning approaches, and the SCDMLGE performs better than SCDML.

In our proposed approach, labeled data must cover all class, which should hinder its application value. For the future work, we will further enhance the performance of our proposed method, and apply it to solve incremental clustering problem.

Acknowledgements This work was supported by the National Natural Science Foundation of China under grants No. 61821003, the National Key Research and Development Program of China under grant No. 2016YFB0800402, ARC Discovery Early Career Researcher Award (DE160100308) and ARC Discovery Project (DP170103954;DP190101985).

References

1. Abavisani, M., Patel, V.M.: Multimodal sparse and low-rank subspace clustering. Information Fusion **39**, 168–177 (2018)
2. Afzalan, M., Jazizadeh, F.: An automated spectral clustering for multi-scale data. Neurocomputing **347**, 94–108 (2019)
3. Arshad, A., Riaz, S., Jiao, L., Murthy, A.: Semi-supervised deep fuzzy c-mean clustering for software fault prediction. IEEE Access **6**, 25675–25685 (2018)
4. Basu, S.: Semi-supervised Clustering: Probabilistic Models, Algorithms and Experiments. PhD thesis University of Texas at Austin (2005)
5. Bo, Y., Fu, X., Sidiropoulos, N.D., Hong, M.: Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August, 2017, pp. 3861–3870 (2017)
6. Chang, J., Wang, L., Meng, G., Xiang, S., Pan, C.: Deep adaptive image clustering. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pp. 5880–5888 (2017)
7. Chen, D., Lv, J., Yi, Z.: Unsupervised multi-manifold clustering by learning deep representation. In: 2017 The Workshops of the The Thirty-First AAAI Conference on Artificial Intelligence, Saturday, February 4-9. San Francisco, California, USA (2017)
8. Chen, G.: Deep learning with nonparametric clustering. arXiv:1501.03084 (2015)
9. Chen, H., Yin, H., Wang, W., Wang, H., Nguyen, Q.V.H., Li, X.: PME: projected metric embedding on heterogeneous networks for link prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, pp. 1177–1186 (2018)
10. Chen, M., Qi, W., Chen, S., Li, X.: Capped l_1-norm sparse representation method for graph clustering. IEEE Access **7**, 54464–54471 (2019)
11. Cui, G., Li, X., Dong, Y.: Subspace clustering guided convex nonnegative matrix factorization. Neurocomputing **292**, 38–48 (2018)
12. Fard, M.M., Thonet, T., Gaussier, É.: Eric Deep k-means: Jointly clustering with k-means and learning representations. arXiv:1806.10069 (2018)
13. Gan, H., Huang, R., Luo, Z., Xi, X., Gao, Y.: On using supervised clustering analysis to improve classification performance. Inf. Sci. **454-455**, 216–228 (2018)
14. Guan, R., Wang, X., Marchese, M., Liang, Y., Yang, C.: A feature space learning model based on semi-supervised clustering. In: 2017 IEEE International Conference on Computational Science and Engineering, CSE 2017, and IEEE International Conference on Embedded and Ubiquitous Computing, EUC 2017, Guangzhou, China, July 21-24, 2017, vol. 1, pp. 403–409 (2017)
15. Guo, L., Yin, H., Qinyong W., Tong Chen: Alexander Zhou, and Nguyen Quoc Viet Hung. Streaming session-based recommendation. SIGKDD (2019)
16. Guo, X., Gao, L., Liu, X., Yin, J.: Improved deep embedded clustering with local structure preservation. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, pp. 1753–1759 (2017)
17. Haponchyk, I., Uva, A., Yu, S., Uryupina, O., Moschitti, A.: Supervised clustering of questions into intents for dialog system applications. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels Belgium, October 31 - November 4, 2018, pp. 2310–2321 (2018)
18. Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm (1979)
19. Heidari, N., Moslehi, Z., Mirzaei, A., Safayani, M.: Bayesian distance metric learning for discriminative fuzzy c-means clustering. Neurocomputing **319**, 21–33 (2018)
20. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics–state-of-the-art, future challenges and research directions. BMC Bioinf. **15**(S6), I1 (2014)
21. Huang, P., Huang, Y., Wang, W., Wang, L.: Deep embedding network for clustering. In: 22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, August 24-28, 2014, pp. 1532–1537 (2014)

22. Kalintha, W., Ono, S., Numao, M., Fukui, K.: Kernelized evolutionary distance metric learning for semi-supervised clustering. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, February 4-9 San Francisco, California, USA, pp. 4945–4946 (2017)
23. Kang, Z., Lu, X., Yi, J., Xu, Z.: Self-weighted multiple kernel learning for graph-based clustering and semi-supervised classification. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, pp. 2312–2318 (2018)
24. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R.: Ask me anything: dynamic memory networks for natural language processing (2015)
25. Kumar, N., Kummamuru, K.: Semisupervised clustering with metric learning using relative comparisons. IEEE Trans. Knowl. Data Eng. **20**(4), 496–503 (2008)
26. Lal Bhatnagar, B., Singh, S., Arora, C., Jawahar, C.V.: Unsupervised learning of deep feature representation for clustering egocentric actions. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, pp. 1447–1453 (2017)
27. Law, M.T., Urtasun, R., Zemel, R.S.: Deep spectral clustering learning. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, pp. 1985–1994 (2017)
28. Li, H.: Learning to rank for information retrieval and natural language processing. Synthesis Lectures on Human Language Technologies **4**(1), 113 (2014)
29. Li, X., Yin, H., Zhou, K., Chen, H., Sadiq, S.W., Zhou, X.: Semi-supervised clustering with deep metric learning. In: Database Systems for Advanced Applications - DASFAA 2019 International Workshops: BDMS, BDQM, and GDMA, Chiang Mai, Thailand, April 22-25, 2019. Proceedings, pp. 383–386 (2019)
30. Lian, D., Zheng, K., Ge, Y., Cao, L., Chen, E., Xie, X.: Geomf++: Scalable location recommendation via joint geographical modeling and matrix factorization. ACM Trans. Inf. Syst. **36**(3), 33:1–33:29 (2018)
31. Liu, G., Liu, Y., Zheng, K., Liu, A., Li, Z., Wang, Y., Zhou, X.: MCS-GPM: multi-constrained simulation based graph pattern matching in contextual social graphs. IEEE Trans. Knowl. Data Eng. **30**(6), 1050–1064 (2018)
32. Liu, G., Wang, Y., Orgun, M.A.: Optimal social trust path selection in complex social networks. In: Proceedings of the 24th Conference on Artificial Intelligence (2010)
33. Liu, G., Wang, Y., Orgun, M.A.: Finding K optimal social trust paths for the selection of trustworthy service providers in complex social networks. In: IEEE International Conference on Web Services, pp. 41–48 (2011)
34. Liu, G., Zheng, K., Wang, Y., Orgun, M.A., Liu, A., Zhao, L., Zhou, X.: Multi-constrained graph pattern matching in large-scale contextual social graphs. In: 31st IEEE International Conference on Data Engineering, pp. 351–362 (2015)
35. Liu, G., Zhu, F., Zheng, K., Liu, A., Li, Z., Zhao, L., Zhou, X.: TOSI: A trust-oriented social influence evaluation method in contextual social networks. Neurocomputing **210**, 130–140 (2016)
36. Ren, Y., Hu, K., Dai, X., Pan, L., Hoi, S.C.H., Xu, Z.: Semi-supervised deep embedded clustering. Neurocomputing **325**, 121–130 (2019)
37. Laine S., Aila, T.: Temporal ensembling for semi-supervised learning. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings
38. Magalhães S., Netto, B., Otàvio, J., Diniz, B., Corrèa Silva, A., Cardoso de Paiva, A., Nunes, R.A., Gattass, M.: Modified quality threshold clustering for temporal analysis and classification of lung lesions. IEEE Trans. Image Processing **28**(4), 1813–1823 (2019)
39. Saha, A., Das, S.: Clustering of fuzzy data and simultaneous feature selection: A model selection approach. Fuzzy Set. Syst. **340**, 1–37 (2018)
40. Sekmen, A., Koku, A.B., Parlaktuna, M., Abdul-Malek, A., Vanamala, N.: Unsupervised deep learning for subspace clustering. In: 2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11-14, 2017, pp. 2089–2094 (2017)
41. Shukla, A., Cheema, G.S., Anand, S.: Clusternet: Semi-supervised clustering using neural networks. arXiv:1806.01547 (2018)
42. Sui, X.L., Xu, L., Qian, X., Liu, T.: Convex clustering with metric learning. Pattern Recogn. **81**, 575–584 (2018)
43. Vu, V.-V.: An efficient semi-supervised graph based clustering. Intell. Data Anal. **22**(2), 297–307 (2018)
44. Wang, F., Li, T., Zhang, C.: Semi-supervised clustering via matrix factorization. In: Proceedings of the SIAM International Conference on Data Mining, SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA, pp. 1–12 (2008)
45. Wang, Q., Yin, H., Hu, Z., Lian, D., Wang, H., Huang, Z.: Neural memory streaming recommender networks with adversarial training. In: Proceedings of the 24th ACM SIGKDD International Conference

on Knowledge Discovery & Data Mining, KDD, 2018. London, UK, August 19-23, 2018, pp. 2467–2475 (2018)

46. Wang, Q., Yin, H., Wang, W., Zi, H., Guo, G., Nguyen, Q.V.H.: Multi-hop path queries over knowledge graphs with neural memory networks. In: DASFAA, pp. 777–794 (2019)
47. Wang, Y., Yin, H., Chen, H., Wo, T., Xu, J., Zheng, K.: Origin-destination matrix prediction via graph convolution: a new perspective of passenger demand modeling. SIGKDD (2019)
48. Xie, D.-Y., Gao, Q., Wang, Q., Xiao, S.: Multi-view spectral clustering via integrating global and local graphs. IEEE Access **7**, 31197–31206 (2019)
49. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**(3), 645–678 (2005)
50. Yang, J., Mcauley, J., Leskovec, J.: Community detection in networks with node attributes. In: IEEE International Conference on Data Mining (2013)
51. Yang, J., Parikh, D., Batra, D.: Joint unsupervised learning of deep representations and image clusters. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 5147–5156 (2016)
52. Yao, D., Zhang, C., Zhu, Z., Hu, Q., Wang, Z., Huang, J.-H., Bi, J.: Learning deep representation for trajectory clustering. Expert Systems, 35(2) (2018)
53. Yin, H., Wang, Q., Zheng, K., Li, Z., Yang, J., Zhou, X.: Social influence-based group representation learning for group recommendation. In: ICDE, pp. 566–577 (2019)
54. Yu, S.-S., Chu, W.S., Wang, C.-M., Chan, Y.-K., Chang, T.-C.: Two improved k-means algorithm. Appl. Soft Comput. **68**, 747–755 (2018)
55. Yuan, T., Deng, W., Hu, J., An, Z., Tang, Y.: Unsupervised adaptive hashing based on feature clustering. Neurocomputing **323**, 373–382 (2019)
56. Zhang, S., Li, J., Jiang, M., Yuan, P., Zhang, B.: Scalable discrete supervised multimedia hash learning with clustering. IEEE Trans. Circuits Syst. Video Techn. **28**(10), 2716–2729 (2018)
57. Zhao, Y., Zheng, K., Li, Y., Su, H., Liu, J., Zhou, X.: Destination-aware task assignment in spatial crowdsourcing: A worker decomposition approach. IEEE Trans. Knowl Data Eng. (2019)
58. Zheng, B., Su, H., Hua, W., Zheng, K., Zhou, X., Li, G.: Efficient clue-based route search on road networks. IEEE Trans. Knowl Data Eng. **29**(9), 1846–1859 (2017)
59. Zheng, K., Yu, Z., Yuan, N.J., Shang, S., Zhou, X.: Online discovery of gathering patterns over trajectories. IEEE Trans. Knowl Data Eng. **26**(8), 1974–1988 (2014)
60. Zheng, K., Zhao, Y., Defu, L., Zheng, B., Liu, G., Zhou, X.: Reference-based framework for spatio-temporal trajectory compression and query processing. IEEE Trans. Knowl. Data Eng. (2019)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Xiaocui Li[1] · Hongzhi Yin[2] · Ke Zhou[1] · Xiaofang Zhou[2]

Xiaocui Li
LXC@hust.edu.cn

Xiaofang Zhou
zxf@itee.uq.edu.au

[1] Wuhan National Laboratory for Optoelectronics and School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China

[2] School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD, Australia