

APPLIED RESEARCH

Deep Fuzzy Clustering Network With Matrix Norm Regularization

FEIYU CHEN¹, YAN LI², AND WEI WANG³¹National Center for Applied Mathematics in Chongqing, Chongqing Normal University, Chongqing 401331, China²School of Mathematical Sciences, Chongqing Normal University, Chongqing 401331, China³Department of Computer Science, Army Medical University, Chongqing 400038, China

Corresponding author: Feiyu Chen (fchen_cqnu@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 12101098, in part by the Natural Science Foundation of Chongqing under Grant cstc2021jcyj-msxmX1053 and Grant 2022NSCQ-LZX0040, in part by the Chongqing Education Commission Key Project under Grant KJQN201800116 and Grant KJZD-K20211480, and in part by the Chongqing Ph.D. “Through Train” Scientific Research Project under Grant CSTB2022BSXM-JCX0106.


ABSTRACT Recently, deep clustering networks, which able to learn latent embedding and clustering assignment simultaneously, attract lots of attention. Among the deep clustering networks, the suitable regularization term is not only beneficial to training of neural network, but also enhancing clustering performance. In the paper, we propose a deep fuzzy clustering network with mixed matrix norm regularization (DFCNR). Specifically, DFCNR uses the weighted intra-class variance as clustering loss, $\ell_{1,2}$ norm and the Frobenius norm of soft assignment matrix as regularization term, where the minimization of $\ell_{1,2}$ norm aims to achieve balanced assignment, and maximization of Frobenius norm aims to achieve discriminative assignment. Moreover, by solving the quadratic convex constraint optimization problem about soft assignment, we derive the activation function of clustering layer. Extensive experiments conducted on several datasets illustrate the superiority of the proposed approach in comparison with current methods.

INDEX TERMS Autoencoder, deep fuzzy clustering, deep learning, matrix norm regularization.

I. INTRODUCTION

Clustering analysis is an unsupervised learning method that divides unlabeled datasets into several clusters by some similarity measurement method. It's one of the important technologies in the field of data mining and machine learning, and is widely used in many fields, such as data analysis, visualization, image segmentation [1]. Comparing with the hard clustering methods [2] which assign each data point to a single cluster, fuzzy clustering methods [3] allow data point is assigned to more than one cluster with a certain probability, which offers more flexibility and robustness. However, fuzzy clustering methods still suffers from the difficulties in separate real high-dimensional data with complex intrinsic distribution.

In recent years, the powerful nonlinear fitting ability and feature representation ability of deep learning have shown advantages in unsupervised deep clustering models,

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

which can better alleviate the defects of traditional clustering algorithms. Deep clustering algorithms based on autoencoder(AE) [4], [5], [6], variational autoencoder(VAE) [7], [8], adversarial autoencoder(AAE) [9], and generative adversarial network(GAN) [10], [11], have achieved remarkable success in various unsupervised applications [12], [13]. There are also some deep clustering frameworks based on graph convolutional networks(GCN). For example, Peng et al. [14] proposes an attention-driven graph clustering network to jointly perform feature learning and cluster assignment in an unsupervised manner. Dong et al. [15] proposes an attention-based hierarchical denoised deep clustering model, which enables GCN to learn multiple layers of hidden information and uses the attention mechanism to strengthen the information, and uses denoising autoencoder to reduce the impact of the clustering. In addition, Nguyen et al. [16] improved the transformer based clustering structure on the limitations of GCN to achieve automatic visual clustering via an unsupervised attention mechanism. Mahon et al. [17] proposes an ensemble clustering algorithm, called selective

pseudo-label clustering (SPC), which combines reconstruction and clustering loss, and uses ensemble technique to select different loss functions for different data points.

To enhance the performance of deep clustering networks, various regularization techniques have been proposed in the literature [18], [19], [20], [21]. IMSAT [18] follows the regularized information maximization framework and uses data augmentations to avoid degenerate solutions. Pang et.al [19] uses extended mutual information as regularization to achieve fair but firm assignment. Dizaji et al. [20] proposes a deep embedding regularized clustering algorithm based on relative entropy regularization. Jabi et al. [21] connects several recent discriminative models directly with K-means through theoretical proof, thus leading to a new soft and regularized deep K-means algorithm. Most of these deep clustering algorithms are based on mutual information and entropy as regularization. However, in machine learning techniques, matrix norm regularization is also widely used in many fields, such as regression analysis, feature selection [22]. Specifically, Ming et al. [22] proposes a flexible feature selection method via $\ell_{1,2}$ norm regularization, which obtains features that generally perform better in many real datasets.

In this paper, we propose a deep fuzzy clustering network with mixed matrix norm regularization(DFCNR), which extracts latent features through deep neural networks and takes weighted intra-class variance as the clustering objective to jointly optimize the representation learning process and clustering process. The $\ell_{1,2}$ norm and the Frobenius norm of soft assignment matrix are used as regularization, the minimization of $\ell_{1,2}$ norm and the maximization of Frobenius norm can achieve balanced and discriminative assignment. In addition, the activation function of the cluster layer is derived by solving the quadratic constraint optimization problem. Finally, experiments are performed on several datasets and compared with some classical and advanced clustering algorithms. A large number of experiments show the superiority of the proposed method.

Our main contributions are summarized as follows:

- We propose a simple but efficient end-to-end representation learning and clustering framework with only a few hundred parameters.
- We employ a mixed matrix norm as a regularization term to guide the representation learning and clustering, which achieves balanced and deterministic clustering assignment.
- We derive a novel activation function of fuzzy clustering layer by approximately solving a quadratic optimization problem, where the deterministic of predicted clustering assignment is automatically controlled by a hyper-parameter.

II. DEEP FUZZY CLUSTERING NETWORK WITH MATRIX NORM REGULARIZATION

A. NOTATIONS AND PRELIMINARY

Given a data matrix $A \in \mathbb{R}^{N \times d}$, the Frobenius norm of matrix A is $\|A\|_F = \left(\sum_{n=1}^N \sum_{k=1}^d a_{nk}^2\right)^{1/2}$. In general,

TABLE 1. Basic notations for the proposed DFCNR.

Notations	Explanations
$A \in \mathbb{R}^{N \times d}$	Data matrix
$X \in \mathbb{R}^{N \times d}$	Sample attribute matrix
$Z \in \mathbb{R}^{N \times r}$	Latent embedding
$P \in \mathbb{R}^{N \times K}$	Soft assignment matrix
$D \in \mathbb{R}^{N \times K}$	The distance matrix

the $\ell_{p,q}$ norm of matrix A is defined as $\|A\|_{p,q} = \left(\sum_{k=1}^K \left(\sum_{n=1}^N |a_{nk}|^p\right)^{q/p}\right)^{1/q}$, with the computational mathematics convention that ℓ_p norm on the first index n and ℓ_q norm on the second index k . Denote \odot as the element-wise hadamard product, i.e., $(A \odot B)_{nk} = a_{nk}b_{nk}$ and $A^{(m)} = A \odot \dots \odot A$, i.e., $[A^{(m)}]_{nk} = a_{nk}^m$. And $\mathbf{1}_K$ is the vector in \mathbb{R}^N with all entries equal to 1.

Denote the soft assignment matrix as $P = [p_1, \dots, p_N]^T \in \mathbb{R}^{N \times K}$, where the element p_{nk} represents the probability that the n -th data point is assigned to the k -th class. And denote

$$\pi_k = \frac{1}{N} \sum_{n=1}^N p_{nk} \quad (1)$$

as the soft cluster frequencies, which are equal to the proportion of data points assigned to each class. The notations are summarized in Tab. 1.

B. OBJECTIVE FUNCTION

The objective function of the proposed DFCNR consists of three terms, the weighted intra-class variance for compactness of the representation, $\ell_{1,2}$ norm and Frobenius norm regularization for balanced and deterministic assignment.

1) WEIGHTED INTRA-CLASS VARIANCE

The weighted intra-class variance is used to enhance the compactness of the representation.

$$\begin{aligned} L_c &\triangleq \|P^{(m)} \odot D\|_{1,1} \\ &= \sum_{k=1}^K \sum_{n=1}^N p_{nk}^m \|z_n - \mu_k\|^2 \end{aligned} \quad (2)$$

where the distance matrix D restores the squares of the Euclidean distances of representation z_n and the cluster centers μ_k , i.e., $D_{nk} = \|z_n - \mu_k\|^2$. The hyperparameter $m \in [1, +\infty)$ controls the smoothness of soft assignment. We set $m = 2$ for all experiments.

2) $\ell_{1,2}$ NORM

The $\ell_{1,2}$ norm is applied to obtain balanced clusters. Consider that the probability $p_{ij} \geq 0$, we have

$$\begin{aligned} L_b &\triangleq \|P\|_{1,2}^2 \\ &= \sum_{k=1}^K \left(\sum_{n=1}^N |p_{nk}|\right)^2 \\ &= \sum_{k=1}^K \left(\sum_{n=1}^N p_{nk}\right)^2 \end{aligned}$$

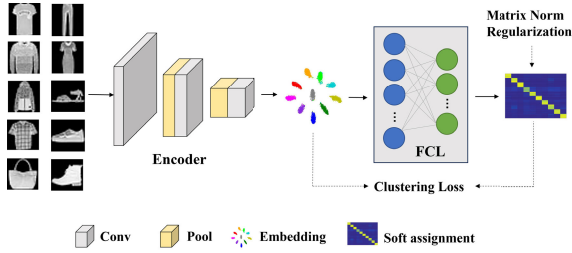


FIGURE 1. Network architecture of DFCNR.

$$= N^2 \sum_{k=1}^K \pi_k^2 \quad (3)$$

Since $\sum_{k=1}^K \pi_k = 1$, the $\ell_{1,2}$ norm (3) reaches a minimum value of N^2/K^2 at $\pi_k = 1/K$ for every k , i.e. each cluster is selected with uniform probability. Therefore, the $\ell_{1,2}$ norm (3) helps to avoid large cluster that most instances are assigned to the same cluster.

3) FROBENIUS NORM

The Frobenius norm is used to improve the deterministic of clustering assignment. According to the definition of Frobenius norm, we have

$$\begin{aligned} L_f &\triangleq \|P - \frac{1}{2}\|_F^2 \\ &= \sum_{n=1}^N \sum_{k=1}^K (p_{nk} - \frac{1}{2})^2 \end{aligned} \quad (4)$$

Since $\sum_{k=1}^K p_{nk} = 1$ for each n , the above term (4) reaches a maximum value of $NK/4$ if and only if each p_n is an one-hot vector, specifying a deterministic distribution.

4) THE TOTAL LOSS FUNCTION

The overall objective function of the proposed DFCNR is given by

$$\begin{aligned} L &= L_c + \alpha L_b - \beta L_f \\ &= \|P^{(2)} \odot D\|_{1,1} + \alpha \|P\|_{1,2}^2 - \beta \|P - \frac{1}{2}\|_F^2 \end{aligned} \quad (5)$$

where α and β are the hyperparameters that balance the three objective terms, note that the value of α and β are different for different datasets and are very critical.

C. NETWORK ARCHITECTURE

The architecture of the proposed DFCNR consists of two parts of layers. Specifically, the representation learning layer for feature extraction and the clustering layer for clustering. Fig. 1 shows the network architecture of DFCNR.

1) REPRESENTATION LEARNING LAYER

Given the data point x_n , we utilize convolutional neural network to learning its latent representation z_n as follows:

$$z_n = \text{CNN}(x_n; \theta) \quad (6)$$

where CNN represents convolutional neural network, and θ is the parameter of CNN.

2) CLUSTERING LAYER

Given the n -th latent feature vector $z_n \in \mathbb{R}^r$, the clustering layer predicts its clustering assignment $p_n \in \mathbb{R}^k$ via (13) with network parameters $\mu = \{\mu_1, \mu_2, \dots, \mu_k\} \in \mathbb{R}^{r \times k}$, where p_{nk} represents the probability that the n -th data point is assigned to the k -th class. We consider this layer as fully-connected layer with customized activation function shown in (13).

According to the loss function (5), the sub-problem about soft clustering assignment can be rewritten as the following constraint optimization problem:

$$\begin{aligned} \arg \min_P & \|P^{(2)} \odot D\|_{1,1} + \alpha \|P\|_{1,2}^2 - \beta \|P - \frac{1}{2}\|_F^2 \\ \text{s.t. } & P \geq 0; \quad P \mathbf{1}_K = \mathbf{1}_N \end{aligned} \quad (7)$$

In general, $\|P^{(2)} \odot D\|_{1,1}$ and $\|P\|_{1,2}^2$ are non-smooth functions of P due to the existence of absolute value sign in the definition of $\ell_{1,1}$ and $\ell_{1,2}$ matrix norm. Fortunately, the non-negativity constraint of P eliminates the non-smoothness of problem (7).

The above optimization problem (7) does not have closed form solution, but we can derive its approximate solution by the Lagrange multiplier method. The following is the detailed derivation process. First, its Lagrangian function is

$$\begin{aligned} \mathcal{L} &= \sum_{k=1}^K \sum_{n=1}^N p_{nk}^2 \|z_n - \mu_k\|^2 + \alpha \sum_{k=1}^K \sum_{n=1}^N p_{nk}^2 \\ &\quad - \beta \sum_{n=1}^N \sum_{k=1}^K (p_{nk} - \frac{1}{2})^2 + \lambda [\sum_{k=1}^K p_{nk} - 1] \end{aligned} \quad (8)$$

And then we take the derivative of \mathcal{L} with respect to p_{nk} , and we set the derivative to be 0.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p_{nk}} &= 2\|z_n - \mu_k\|^2 p_{nk} + 2\alpha \sum_{n=1}^N p_{nk} - 2\beta(p_{nk} - \frac{1}{2}) + \lambda \\ &= 0 \end{aligned} \quad (9)$$

Assuming that $\pi_k = 1/K$ in (1) and substituting it into (9) for simplification, we obtain:

$$2\|z_n - \mu_k\|^2 p_{nk} + \frac{2\alpha N}{K} - 2\beta p_{nk} + \beta + \lambda = 0 \quad (10)$$

Then:

$$p_{nk} = \frac{-\lambda - \beta - \frac{2\alpha N}{K}}{2(\|z_n - \mu_k\|^2 - \beta)} \quad (11)$$

Since $\sum_{k=1}^K p_{nk} = 1$, the sum of k on both sides of the above equation yields an expression for λ :

$$\lambda = -\frac{2}{\sum_{k=1}^K \frac{1}{(\|z_n - \mu_k\|^2 - \beta)}} - \beta - \frac{2\alpha N}{K} \quad (12)$$

Algorithm 1 DFCNR Algorithm

Input: The input data X , the number of clusters K , Maximum iterations $MaxIter$

Output: The soft clustering assignment matrix P

- 1: Initialize θ and μ
- 2: **for** $iter \in 1, 2, \dots, MaxIter$ **do**
- 3: Compute the latent representation z_n by (6)
- 4: Compute the soft assignment p_{nk} by (13)
- 5: Update θ and μ by Adam
- 6: **end for**

TABLE 2. Datasets description.

Dataset	Samples	Dimension	Classes
MNIST-full	70000	1×28×28	10
MNIST-test	10000	1×28×28	10
Fashion-MNIST	70000	1×28×28	10
Reuters	10000	2000	4

Finally, taking the expression of λ into (11). By simplification, the final expression of p_{nk} is obtained as follows:

$$p_{nk} = \frac{(\|z_n - \mu_k\|^2 - \beta)^{-1}}{\sum_{k'} (\|z_n - \mu_{k'}\|^2 - \beta)^{-1}} \tag{13}$$

Note that formula (13) degenerates into the activation function of DEC [4] and DFC [19] if we set $\beta = -1$ or $\beta = 0$. In this article, we set β as a small positive number to ensure p_{nk} is a probability function. The larger value of β we set, the more deterministic assignment it got.

D. DFCNR ALGORITHM

Given an initial guess of parameters of representation learning layer and clustering layer, the proposed DFCNR algorithm conducts the following two steps alternatively. In the forward step, DFCNR estimates the latent embedding and its soft assignment via neural networks. In the backward step, DFCNR updates the parameters via adaptive momentum algorithm(Adam). The whole algorithm is summarized in Algorithm 1.

III. EXPERIMENTAL ANALYSIS

In this section, experiments are carried out on four classical datasets, handwritten digit image datasets (MNIST-full, MNIST-test), a clothing image dataset consisting of images such as skirts and jackets (Fashion-MNIST) and a text dataset (Reuters), respectively. For Reuters, we randomly sampled a subset of 10000 examples and followed DEC [4] using 4 root categories: corporate/industrial, government/social, markets and economics as labels and excluded all documents with multiple labels. We also calculated tf-idf features on the 2000 most frequent words. A brief description of these datasets is summarized in Tab. 2.

The proposed DFCNR algorithm is compared with some classical clustering algorithms (K-means [2], FCM [3], GMM [23], N-Cuts [24]), as well as some recent deep clustering algorithms (SEC [25], DEC [4], IDEC [5], VaDE

[7], DCC [26], JULE [6], DEPICT [20], ClusterGAN [10], DSCDAN [27], SR-K-means [21], S3VDC [28], GrDNFCS [29], ACe/DeC [30], DFC [19], ADEC [9]).

A. EVALUATION METRICS

The commonly used clustering accuracy (ACC) and normalized mutual information (NMI) are used as evaluation metrics to measure the performance of clustering.

ACC measures the similarity between the true label and the predicted label:

$$ACC = \max_m \frac{\sum_{n=1}^N \mathbf{1}\{l_n = m(c_n)\}}{N} \tag{14}$$

where N is the total number of samples, l_n is the ground-truth, c_n is the cluster assignment produced by the algorithm, and m ranges over all possible one-to-one mappings between clusters and labels.

NMI measures the degree of agreement between two data distributions based on the idea of information entropy:

$$NMI = \frac{MI(c, l)}{\max(H(c), H(l))} \tag{15}$$

where $H(l)$ is the entropy of l and MI is the mutual information of c and l .

B. EXPERIMENTAL SETUP

For image and text data, the convolutional autoencoder (ConvAE) and the stacked autoencoder (SAE) are used for feature extraction. For datasets using ConvAE, the structure parameter is $Conv_{32}^5 \rightarrow Conv_{64}^5 \rightarrow Conv_{128}^3$, where $Conv_n^k$ represents a convolution layer with the number of channels n , and the convolution kernel size k , and then a fully connected layer is used to map to the K -dimensional embedding space. For Reuters, the SAE also consists of four fully connected layers. Before training, all raw data is normalized to [0, 1], the learning rate is set to 0.001 and adopts Adam [31] as optimization method.

C. CLUSTERING RESULTS

The experimental results on four datasets are shown in Tab. 3 (in terms of ACC and NMI), where the best results are marked in bold. Most of the clustering results by different methods were taken from their published articles. Mark (*) indicates results from their own published paper, mark (†) indicates results presented in other authors' published articles, and mark (–) indicates that no actual results could be obtained. From the experimental results, we can see that DFCNR achieves better performance on all datasets, and obtains the best ranking on MNIST-full and Fashion-MNIST. Compared to traditional clustering methods such as K-means, it clearly has superior clustering performance. Compared with other existing deep clustering methods such as DEC, IDEC, GrDNFCS, the superiority of the method is still demonstrated. Especially on Fashion-MNIST, we surpassed ADEC 12.3% on ACC.

TABLE 3. Comparison of the Clustering Performance.

Method	MNIST-full		MNIST-test		Fashion-MNIST		Reuters	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
K-means [2]	0.535 [†]	0.501 [†]	0.541 [†]	0.503 [†]	0.471 [†]	0.508 [†]	0.550 [†]	0.323 [†]
FCM [3]	0.547 [†]	0.482 [†]	-	-	0.529 [†]	0.516 [†]	0.601 [†]	0.337 [†]
GMM [23]	0.540 [†]	0.509 [†]	0.537 [†]	0.501 [†]	0.523 [†]	0.409 [†]	0.547 [†]	0.410 [†]
N-cuts [24]	0.647 [†]	0.710 [†]	0.648 [†]	0.719 [†]	0.489 [†]	0.551 [†]	0.594 [†]	0.369 [†]
SEC [25]	0.627 [†]	0.604 [†]	-	-	0.542 [†]	0.558 [†]	0.651 [†]	0.366 [†]
DEC [4]	0.843*	0.829 [†]	0.853 [†]	0.821 [†]	0.523 [†]	0.551 [†]	0.722*	0.468 [†]
IDEC [5]	0.881*	0.867*	0.846 [†]	0.802 [†]	0.529 [†]	0.557 [†]	0.756*	0.498*
VaDE [7]	0.945*	0.876 [†]	-	-	0.578 [†]	0.630 [†]	0.798*	0.541 [†]
DCC [26]	0.963 [†]	0.913 [†]	-	-	-	-	0.596 [†]	0.572 [†]
JULE [6]	0.964 [†]	0.913*	0.931 [†]	0.915*	0.563 [†]	0.608 [†]	-	-
DEPCT [20]	0.965*	0.917*	0.963*	0.915*	0.392 [†]	0.392 [†]	-	-
ClusterGAN [10]	0.950*	0.890*	-	-	0.630*	0.640*	-	-
DSCDAN [27]	0.978*	0.941*	0.980*	0.946*	0.662*	0.645*	-	-
SR-K-means [21]	0.939*	0.866*	0.863*	0.873*	0.507 [†]	0.548 [†]	-	-
S3VDC [28]	0.950*	0.902*	0.961*	0.911*	0.613*	0.610*	-	-
GrDNFCS [29]	0.915*	0.907*	-	-	0.635*	0.661*	0.778*	0.527*
ACe/DeC [30]	0.980*	0.940*	0.980*	0.940*	0.600*	0.640*	-	-
DFC [19]	0.980*	0.952*	0.982*	0.953*	0.708*	0.617*	0.823*	0.578*
ADEC [9]	0.986*	0.961*	0.985*	0.957*	0.586*	0.662*	0.821*	0.605*
DFCNR	0.986	0.976	0.983	0.958	0.709	0.662	0.832	0.600

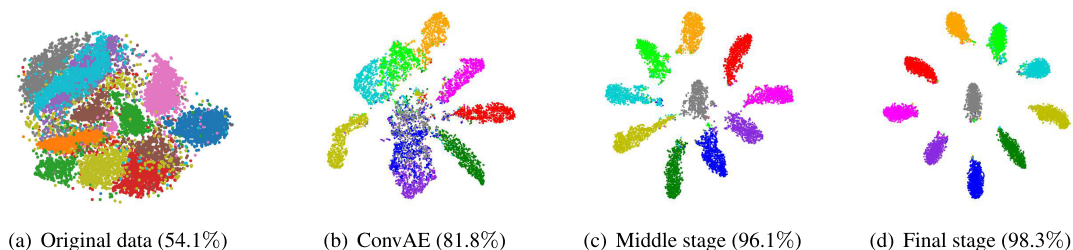


FIGURE 2. The evolution of features and their corresponding clustering accuracy (in brackets) across the whole training stage on MNIST-test dataset, the colors indicate the clustering assignment obtained from DFCNR.

D. EMBEDDING VISUALIZATION

The proposed DFCNR algorithm is capable of simultaneously learning latent embedding and clustering assignments. In order to observe how DFCNR converges to the final result, we perform t-SNE [32] at four different stages on MNIST-test dataset. Based on the results shown in Fig. 2, the original data all mixed together before training. As the training process goes, it is clear to see that the learned embedding gather and scatter more distinctly, and the intra(inter)-cluster variance has a significant decreases(increase), which means that the learned embedding become more and more suitable for clustering, also the clustering accuracy is keep increasing.

E. PARAMETERS SENSITIVITY

As mentioned before, the choice of α and β is crucial for DFCNR. We experimented by sampling the effect of different α and β on the performance. For the Reuters dataset, we conduct experiments by sampling $\alpha = \{0.001, 0.01, 0.1, 1, 10\}$ and for the other datasets by sampling $\alpha = \{0.1, 1, 10, 50, 100\}$. All datasets are experimented by sampling $\beta = \{0.001, 0.01, 0.1, 1, 10\}$. Fig. 3 shows the clustering accuracy of the proposed model on the datasets MNIST-full, MNIST-test, Fashion-MNIST, and Reuters with the two varying hyper parameters. From Fig. 3, we can observe that the model achieves the best performance when

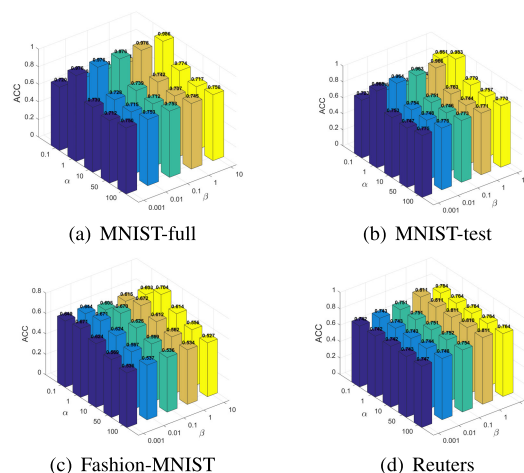


FIGURE 3. ACC with different α and β .

setting $\alpha = 1$ and $\beta = 10$ on MNIST-full, MNIST-test and Fashion-MNIST datasets, $\beta = 1$ on Reuters dataset. We suggest to choose $\alpha = 1$ for the proposed model and vary β in $[1, 10]$ for difference datasets in the paper.

F. ABLATION ANALYSIS FOR REGULARIZATION

In order to prove the effectiveness of the proposed mixed matrix norm regularization, experiments are carried out on

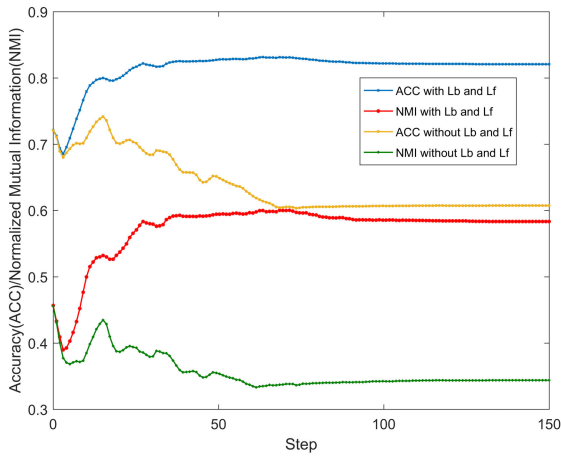


FIGURE 4. Performance comparison with and without L_b and L_f regularization terms on the Reuters dataset.

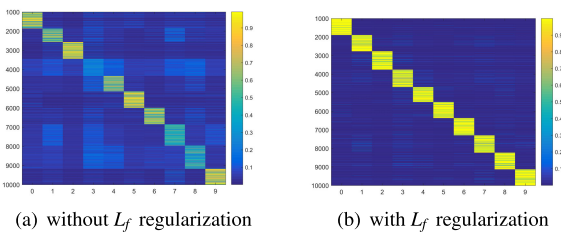


FIGURE 5. Plot of clustering assignment matrix with and without L_f regularization on the MNIST-test dataset.

the Reuters dataset to evaluate the influence of regularization (L_b and L_f) on the clustering performance, where L_b represents the $\ell_{1,2}$ norm and L_f represents the Frobenius norm. Fig. 4 shows the effect of the presence or absence of regularization on ACC and NMI. It is clear that ACC and NMI are much higher after using two terms as regularization, which indicates that the mixed matrix norm regularization term is beneficial to improve the clustering performance. As mentioned above, the mixed matrix norm regularization aims to achieve discriminative and balanced cluster assignments. We conducted experiments with and without mixed matrix norm regularization on the MNIST-test dataset, and Fig. 5 plots the the ordered clustering assignment matrix with and without L_f on MNIST-test dataset. Notice that the assignment obtained with L_f regularization has much more clearly block structure and larger color difference, which reflects the discriminant of assignment; on the other hand, comparing with Fig. 5(a), the size of block is almost uniform in Fig. 5(b), which implies the balance of assignment.

We also carry out corresponding experiments to illustrate the balance of allocation. We conduct experiments on MNIST-full, MNIST-test, Fashion-MNIST and Reuters datasets respectively. Fig.6 shows the impact of L_b with respect to the number of points assigned into each class, which calculated from predicted clustering assignment. As can be seen from Fig.6, taking the Fashion-MNIST dataset as an example, the ground truth distribution of clustering label is shown in green line, which are uniformly distributed. The red line represents the result according to our objective

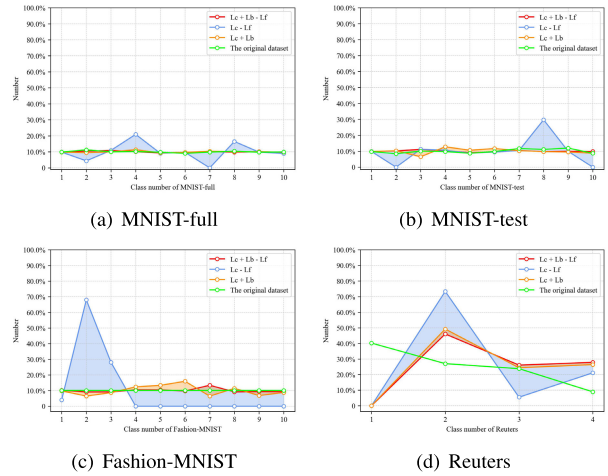


FIGURE 6. The impact of L_b on clustering results on different datasets. The horizontal axis represents the number of clusters in the dataset, and the vertical axis represents the proportion of the number of each class to the total number.

TABLE 4. Performance comparison of three different soft assignment estimation methods based on ACC and NMI.

Method		FCM	t-distribution	formula (13)
MNIST-full	ACC	0.9728	0.9728	0.9837
	NMI	0.9504	0.9553	0.9667
MNIST-test	ACC	0.9702	0.9705	0.9834
	NMI	0.9339	0.9350	0.9575
Fashion-MNIST	ACC	0.6892	0.7004	0.7090
	NMI	0.6506	0.6514	0.6620
Reuters	ACC	0.8307	0.8305	0.8315
	NMI	0.5982	0.5986	0.5999

function, the orange one without L_f (Frobenius norm), and the blue one without L_b ($\ell_{1,2}$ norm). It is obvious that the distribution of the blue line points is very uneven, and most of the data points are assigned to the second and third classes. The red and orange lines with L_b are more evenly distributed than the blue line, and the same conclusion can be drawn for the other two datasets. Therefore, it also verifies the effectiveness of L_b ($\ell_{1,2}$ norm) to measure the balanceness.

G. INFLUENCE OF ACTIVATION FUNCTIONS

We further investigate the impact of activation functions, reporting model performance (ACC and NMI) using three different activation functions on different datasets. The activation functions used include Fuzzy C-means(FCM), t-distribution and formula (13). In Tab.4 we mark the best results in bold, and it can be seen that the formula (13) outperforms the other activation functions in all tests.

IV. CONCLUSION

In this paper, we propose a mixed matrix norm regularized deep fuzzy clustering network (DFCNR) for feature learning and cluster assignment in an end-to-end manner. The encoder network is used to map the original data into a more appropriate latent space, and the learned latent representation is fed to the designed fuzzy clustering layer (FCL) to predict the soft assignment. For clustering, weighted intra-class variance is used to enhance the compactness

of the learned embedding. From the perspective of matrix norms, $\ell_{1,2}$ norm and Frobenius norm are proposed as regularizations to obtain more balanced and discriminative soft assignments. The effectiveness of the proposed method is verified by experiments on different datasets. Compared with the traditional clustering methods and the current popular deep clustering methods, the proposed method has higher clustering accuracy. It effectively solves the performance defects of traditional algorithms, improves the clustering performance of fuzzy clustering on high-dimensional complex datasets, and verifies the superiority of the proposed model.

REFERENCES

- [1] S. Arora and I. Chana, "A survey of clustering techniques for big data analysis," in *Proc. 5th Int. Conf.-Confluence Next Gener. Inf. Technol. Summit (Confluence)*, Sep. 2014, pp. 59–65.
- [2] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, 1967, pp. 281–297.
- [3] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, Jan. 1973.
- [4] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 478–487.
- [5] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1753–1759.
- [6] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5147–5156.
- [7] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational deep embedding: An unsupervised and generative approach to clustering," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1965–1972.
- [8] K.-L. Lim, X. Jiang, and C. Yi, "Deep clustering with variational autoencoder," *IEEE Signal Process. Lett.*, vol. 27, pp. 231–235, 2020.
- [9] N. Mrabah, M. Bouguessa, and R. Ksantini, "Adversarial deep embedded clustering: On a better trade-off between feature randomness and feature drift," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 1603–1617, Apr. 2022.
- [10] S. Mukherjee, H. Asnani, E. Lin, and S. Kannan, "ClusterGAN: Latent space clustering in generative adversarial networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 4610–4617.
- [11] X. Chen, Y. Duan, R. Houthoof, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2016, pp. 2172–2180.
- [12] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, Apr. 2017.
- [13] E. Karamatli, A. T. Cemgil, and S. Kirbiz, "Audio source separation using variational autoencoders and weak class supervision," *IEEE Signal Process. Lett.*, vol. 26, no. 9, pp. 1349–1353, Sep. 2019.
- [14] Z. Peng, H. Liu, Y. Jia, and J. Hou, "Attention-driven graph clustering network," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 935–943.
- [15] Y. Dong, Z. Wang, J. Du, W. Fang, and L. Li, "Clusformer: A transformer based clustering approach to unsupervised large-scale face and visual landmark recognition," *World Wide Web*, vol. 26, no. 1, pp. 441–459, 2023.
- [16] X. B. Nguyen, D. T. Bui, C. N. Duong, T. D. Bui, and K. Luu, "Attention-driven graph clustering network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2021, pp. 10847–10856.
- [17] L. Mahon and T. Lukasiewicz, "Selective pseudo-label clustering," in *Proc. German Conf. Artif. Intell.*, 2021, pp. 158–178.
- [18] W. Hu, T. Miyato, S. Tokui, E. Matsumoto, and M. Sugiyama, "Learning discrete representations via information maximizing self-augmented training," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1558–1567.
- [19] Y. Pang, F. Chen, S. Huang, Y. Ge, W. Wang, and T. Zhang, "Deep fuzzy clustering with weighted intra-class variance and extended mutual information regularization," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2020, pp. 464–471.
- [20] K. G. Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang, "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5747–5756.
- [21] M. Jabi, M. Pedersoli, A. Mitiche, and I. B. Ayed, "Deep clustering: On the link between discriminative models and K-means," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 1887–1896, Jun. 2021.
- [22] D. Ming, C. Ding, and F. Nie, "A probabilistic derivation of LASSO and $L_{1,2}$ -norm feature selections," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 4586–4593.
- [23] G. McLachlan and D. Peel, *Finite Mixture Models*. New York, NY, USA: Wiley, 2000.
- [24] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
- [25] F. Nie, D. Xu, I. W. Tsang, and C. Zhang, "Spectral embedded clustering," in *Proc. Int. Joint Conf. Artif. Intell.*, 2009, pp. 1181–1186.
- [26] S. A. Shah and V. Koltun, "Deep continuous clustering," 2018, *arXiv:1803.01449*.
- [27] X. Yang, C. Deng, F. Zheng, J. Yan, and W. Liu, "Deep spectral clustering using dual autoencoder network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4061–4070.
- [28] L. Cao, S. Asadi, W. Zhu, C. Schmidli, and M. Sjberg, "Simple, scalable, and stable variational deep clustering," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2021, pp. 108–124.
- [29] Q. Feng, L. Chen, C. L. P. Chen, and L. Guo, "Deep fuzzy clustering—A representation learning approach," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 7, pp. 1420–1433, Aug. 2020.
- [30] L. Miklautz, L. G. M. Bauer, D. Mautz, S. Tschitschek, C. Böhm, and C. Plant, "Details (don't) matter: Isolating cluster information in deep embedded spaces," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 2826–2832.
- [31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [32] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.



FEIYU CHEN received the B.Sc. degree from the School of Mathematics, Liaoning Normal University, China, in 2007, the M.S. degree from the School of Mathematic Sciences, Dalian University of Technology, China, in 2009, and the Ph.D. degree from the Department of Applied Mathematics, Delaware State University, USA, in 2013. He is currently an Associate Professor with the National Center for Applied Mathematics in Chongqing, Chongqing Normal University, China. His current research interests include data clustering, machine learning, and optimization.



YAN LI received the B.Sc. degree from the School of Mathematic Sciences, Chongqing Normal University, China, in 2021, where she is currently pursuing the master's degree in applied mathematics with the School of Mathematical Sciences. Her current research interests include machine learning, deep learning, and data clustering.



WEI WANG received the B.Sc. degree from the Department of Computer and Information Technology, Nanyang Normal University, China, in 2012, the M.S. degree from the College of Computer Science, Chongqing University, China, in 2016, and the Ph.D. degree from the School of Big Data and Software Engineering, Chongqing University, in 2020. His current research interests include clustering, machine learning, and deep learning.

...