# Cohesive Pair-Wises Constrained Deep Embedding for Semi-Supervised Clustering with Very Few Labeled Samples*

Jing Zhang
School of Computer Science and Artificial Intelligence, Liaoning Normal University, China
zhangjing_0412@lnnu.edu.cn

Guiyan Wei
School of Computer Science and Artificial Intelligence, Liaoning Normal University, China
turbo981226@163.com

Yonggong Ren
School of Computer Science and Artificial Intelligence, Liaoning Normal University, China
ryg@lnnu.edu.cn

**Abstract:** *Semi-supervised learning is a powerful paradigm for excavating latent structures of between labeled and unlabeled samples under the view of models constructing. Currently, graph-based models solve the approximate matrix that directly represent distributions of samples by the spatial metric. The crux lies in optimizing connections of samples, which is achieved by subjecting to must-links or cannot-links. Unfortunately, to find links are rather difficult for semi-supervised clustering with very few labeled samples, therefore, significantly impairs the robustness and accuracy in such scenario. To address this problem, we propose the Cohesive Pair-wises Constrained deep Embedding model (CPCE) to obtain an optimal embedding for representing the category distribution of samples and avoid the failed graph-structure of the global samples. CPCE designs the deep network framework based on CNN-Autoencoder by minimizing reconstruct errors of samples, and build up constrains both of the sample distribution for within-class and the category distribution for intra-class to optimal the latent embedding. Then, leverage the strong supervised information obtained from cohesive pair-wises to pull samples into within-class, which avoid the similarity of high-dimension features located in different categories to achieve more the compact solution. We demonstrate the proposed method in popular datasets and compare the superiority with popular methods.*

**Keywords:** *Semi-supervised learning, clustering, auto-encoder network, pair-wise.*

*Received October 2, 2022; accepted May 24, 2023*
*https://doi.org/10.34028/iajit/21/1/7*

## 1. Introduction

Pattern classification serves as a popular technique of Machine Learning, which is used widely in various fields including medical diagnosis [1, 10], regional science [2], network analysis and so on [22]. Unfortunately, the labeled samples are quite rare in practical scenarios due to the most labels are original from manually annotated, which limits to build up the classification model, and a huge amount of unlabeled data with the wealth information, for instance, latent structures between samples, are wasted. Clustering models pull samples with similar features into the same cluster according to the distance of similarity both of samples in the feature space, which provides a more general way to excavate knowledges without labels. However, clustering models will waste expensive labeled samples that have certainty and guiding significance when datasets containing few labeled samples.

Aiming to above problems, semi-supervised learning is proposed [13], which attracts a lot of interesting from researchers. Semi-supervised clustering excavates latent structures between unlabeled and labeled samples, and utilizes the supervised information from labeled samples to predict classifications of samples [15, 19]. Most approaches perform in the supervised framework with help of large amounts of labeled samples, which greatly improve the clustering performance and have widely applicability. However, these models are limited to solve the scenarios containing unlabeled samples. Graph-based methods [3, 8, 19] are developed in constructing the latent space structures between samples. Early, the kind of methods by solving the similarity matrix follow Laplace segmentation, which has received more attention, e.g., well-known normalized cutting, and ratio cutting. More recent works seek to leverage unsupervised information by constructing accurate pair-wises. Nie *et al.* [14] proposed utilizing cannot-links to optimize the latent structure of the global graph. In order to constrain the semi-supervised clustering model to improve the performance, furthermore, Nie *et al.* [15]. induce the pair-wise constraints that utilize pairs locating in different categories to strengthen the clustering model. In this model, some irrelevant pair-wises from labeled samples provide the robust information to avoid unlabeled samples with similar features that are incorrectly classified. However, cannot-links will be invalid when datasets contain very few labeled samples, especially in multi-classes datasets due to the difficulty of finding a large number of accurate cannot-link pairs through the transitive linking method in a multi-class dataset with very few labels, using this method to obtain inaccurate pairs will weaken the clustering effect. Moreover, in the high-dimension feature space, few constraints of pair-wises provide limited help for clustering. Because of the category distribution is inexact in the original high-dimension feature space, for instance, high-dimension image datasets.

Deep Learning models have achieved the great success for solving tasks of high-dimension data processing, which is derived from the non-linear representation capability of objects. These methods based semi-supervised learning that transfer from the original feature to the optimal feature space by means of supervised information of labeled samples to improve the clustering performance. Ren *et al.* [16] proposed Semi-Supervised Deep Embedded Clustering (SDEC) that follows auto-encoder framework based on convolution unit to obtain high-level features and induces the structure constraints of labeled samples to optimize feature learning, which provide the novel solution for semi-supervised clustering of high-dimension features. Nevertheless, it only considers relationship between samples in the optimal processing of features, and ignores the category distribution from unlabeled samples.

Therefore, the challenges in constructing a semi-supervised model for high-dimensional data with very few labels mainly focus on the following two aspects:

1. How to obtain more discriminative high-dimensional feature representations with very little supervised information.
2. How to fully utilize a very small amount of labeled data to assist in clustering model construction without introducing noise information.

To deal with above challenges, we propose a novel semi-supervised clustering method by excavating the more discriminative embedding to build up correct relationships between of unlabeled and labeled samples. First, we utilize CNN-Autoencoder network bone to extract non-linear features of samples. Second, supplement the constraint of the category distribution to transfer the features into the latent space of clustering. Finally, leverage the regularization constraint from few cohesive pair-wises to strengthen the sample distribution within-category.

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

1. We design a novel deep-network framework with the helping with CNN-Autoencoder for semi-supervised clustering, which leverages both of the reconstruction errors and the latent category distribution constraint to obtain more discriminative features of samples.
2. We induce the cohesive pair-wise constraint term original from very few labeled samples by a compact criterion to guarantee samples with unsimilar features still located in the same category to optimize the embedding.
3. We verify the effectiveness of proposed model in general semi-supervised datasets, and compare with popular models even in datasets with very few labeled samples. Our model achieves more satisfactory performances.

## 2. Related Work

### 2.1. Semi-Supervised Clustering

Semi-supervised learning utilizes the latent structures information of unlabeled samples and composes with the less supervised information of labeled samples to construct the learning model. From views based on representations of data, there are mainly three kinds of semi-supervised clustering methods, including metric learning-based [12, 18], feature learning-based [20] and graph-based clustering [3, 19]. In metric learning-based clustering, Zhang *et al.* [24] proposed a non-linear transformation clustering method with distance matrix learning, which performed well on non-linearly separable data. Metric learning can autonomously learn task-specific distance metric functions according to different tasks. The dataset $X\in\Re^{N\times d}$ containing the sample $x_i$, $i$=1, ..., $N$, the general metric function $d()$ is defined as:

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T M(x_i - x_j)}, \tag{1}$$

where $M$ is called the metric matrix, which is the inverse of the covariance matrix. Obviously, $M$ is a symmetric matrix. And, the s ample pair $x_i$ and $x_j$ locate in different clusters. Khanali and Vaziri [9] proposed a probabilistic model that combines with fuzzy clustering and metric learning to maximize the distance between centers. However, it is difficult to directly solve the classification information by the symmetric matrix from pair-wises.

Feature learning-based clustering [23] divides samples into different clusters according to the sample distribution in feature spaces. Over the past few decades, many more efficient varieties of k-means have been proposed. Solorio-Fernández *et al.* [17] proposed a useful model for simultaneous clustering based on feature selection and fuzzy data. [4] proposed an adaptive hashing method follow feature clustering to extract more discriminative features and reduces dimensionality of data. This method aims to minimize the following objective functions:

$$J = \sum_{k=1}^{K}\sum_{i=1}^{N} \|x_i - c_k\|_2^2, \tag{2}$$

In Equation (2), euclidean distance between of any sample $x_i$ and the cluster center $c_k$ is used to achieve clustering. The main idea of the kind of method is that construct the samples graph and define the relationship of edges, and use the effective energy function as the evaluation standard.

To address the lack of explicit clustering structures, the constrained Laplacian rank algorithm is proposed by Nie *et al.* [14] for direct multi-class graph clustering. For the dataset $X$ defining the pre-constructed affinity matrix $A \in \Re^{N \times N}$, and will optimize the target $S \in \Re^{N \times N}$. This optimization expression is defined as follows:

$$\min \|S - A\|_F^2 \; s.t. S \geq 0, rank(L_s) = N - C', \tag{3}$$

where $L_s$=$D_s$-$S$ is the Laplace matrix of $S$ and $D_s$ is the diagonal matrix with the sum of each row of $S$. The model takes a rough similarity matrix $A$ and obtains a non-negative normalized approximation $S$ with exactly connected components. However, whether incorrect connections are removed and valid connections are retained will be uncontrolled, Nie *et al.* [15] proposed a non-link graph regularization method to learn pair-wises from a given affinity $A$ by means of the supervised information with pair-wise constraints. Solving the cannot-linked constraint problem by a specific cannot-linked graph regularization can be demonstrated to solve the cannot-linked constraint in graph learning, providing a matchable pairwise constraint selection, significantly improving semi-supervised clustering performance.

## 2.2. Deep Clustering

In order to obtain more efficient representation for clustering, deep clustering integrates the goal of clustering into the powerful representation capability [21]. Typical semi-supervised clustering methods work in the original feature space with poor representation ability, and it is reasonable to use Deep Neural Networks (DNN) for semi-supervised clustering to make SSC more powerful [6]. Li *et al.* [11] proposed a Deep Metric Learning-based Semi-Supervised Clustering method (SCDML) by adopting triplet loss in a deep metric learning network and combined with a label propagation strategy to dynamically update the unlabeled samples. Deep Embedding Clustering (DEC) constructed the KL divergence loss, which jointly learns feature representations and cluster assignments, making the representations learned closer to the cluster centers. However, DEC does not ignore prior knowledge to guide the learning processing. Ren *et al.* [16] improved DEC and proposed a new SDEC scheme to overcome the limitation. The Improved Deep Embedded Clustering (IDEC) adds the constraints to hold the latent structures of samples and autoencoders learn better representations. In semi-supervised deep clustering frameworks, the loss of KL divergence and semi-supervised loss are jointly optimized to obtain the deep representation of clusters, which proves that semi-supervised information indeed improves the deep representation of clusters. This paper proposes a pair-wise prediction considers the relationship between pairings and proposes a new cannot-linked graph regularization to obtain key pairing constraints via Stacked Auto-Encoders (SAE) [5] and key pair constraint clustering extract discriminative features and update cluster centers at the same time, effectively utilize deep learning and semi-supervised learning, greatly improve clustering performance.

## 3. The Proposed Model

Graph-based models develop the latent structures of unlabeled samples to construct the optimal non-negative approximation matrix. Moreover, in order to fully utilize the latent information of correct pairs and to avoid the negative effect of incorrect pairs, Nie *et al.* [15] proposed the information transmission chain constructed cannot-link pairs to strengthen the graph-structures of samples, which achieves satisfactory performances. However, this kind of two-step models from the original space of samples will product the invalid constraints of cannot-link pairs when data contain relatively little ground-truth information, especially these samples locate in the high-dimension feature space. In the kind of scenarios, the incorrect structure graph from few labeled samples will induces the incorrect latent manifold representations. And, graph-based models do not attention the category distribution of data.

To deal with above problems, we propose the Cohesive Pair-wise Constraint Embedding (CPCE) method to achieve semi-supervised clustering with very few labeled samples by solving the optimal embedding. CPCE maps samples containing unlabeled samples and very few labeled samples from the high-dimensional original space to the latent structure space subjected to both of maximizing the distance between categories and minimizing the distance of samples located in the same category. Furthermore, we induce cohesive pair-wise constrains from few labeled samples to optimize the learning process, which obtain the discriminative features and updates cluster centers. In order to solve semi-supervised clustering of high-dimension datasets, for instance, image datasets, we improve above method to build up the deep embedding semi-supervised model through autoencoder network framework, which fully leverages the non-linear representation of the deep network to improve the clustering performance.

### 3.1. Cluster Embedding with Cohesive Pair-Wise Constraints

Despite very few labeled samples provide efficient supervised information for clustering, graph-based clustering models are rather difficult to utilize few samples to construct the accuracy graph structure of global. In order to more efficient leverage few labeled samples, we propose a novel method to solve the embedding with subjecting to optimal conditions of category distribution, and obtain accuracy clustering centers. So, the semi-supervised dataset $X$ is consisted by both parts $\{X_l, X_u\}$ of the labeled dataset $X_l$ and the unlabeled dataset $X_u$ builds up the optimization equation.

According to the metric clustering theory in Equation (1), we solve the clustering centers $c_k$,$k$=1,…,$K$ in dataset $X$ by optimizing the projection model $H$ to ensure the discriminability of the sample distribution. Moreover, we ensure that the distance between samples located in the same categories is as small as possible, and the distance between samples located in different categories is as large as possible in the projection space. In this model, the number of categories is known, which results from the part of labeled samples $X_l$. The following optimization equation is obtained:

$$\min_{H,c} \sum_{k=1}^{K} \sum_{i=1}^{N} \|H^T x_i - c_k\|_2^2 - \lambda \sum_{k,j=1}^{K} \left\|H^T c_k - H^T c_j\right\|_2^2, \qquad (4)$$

where $\lambda$ is a tunable positive parameter, and $N$ is the number of samples, and $K$ is the number of categories. In Equation (4), $c_k$ is the clustering center of the category $k$, and $\|\cdot\|_2^2$ represents the $L_2$-norm and the right term is a conventional term used to prevent overfitting.

The above formulate attention to maximize the distance between category centers for constructing the embedding, and ignore the hard samples located in the classification boundary. These hard samples greatly influence the accuracy of model. Therefore, we utilize the supervised and discriminative information from the labeled part $X_l$ to obtain the more compact solution. We are inspired by Nie *et al*. [15] that proposed and discussed the cohesive pair-wise constraints in solving the embedding. Therefore, we introduce cohesive pair-wise constraints to obtain more optimal the embedding and more accuracy clustering centers. The improved optimization equation with the cohesive pair-wise constrains is expressed as follows:

$$\min_{H,c} \sum_{k=1}^{K} \sum_{i=1}^{N} \|H^T x_i - c_k\|_2^2 - \lambda \sum_{k,j=1}^{K} \left\|H^T c_k - H^T c_j\right\|_2^2, s.t. (y_a + y_b)\|H^T x_a - H^T x_b\| \leq \rho, \qquad (5)$$

where $\rho$ is a rather small constant. The sample $x_a$ or $x_b$ is the ground-truth corresponding to the marked sample $y_a$ or $y_b$. In this paper, we utilize ADMM to solve both of $H$ and $c$ by iterating alternately. The above formula realizes the strong pairwise constraint embedding, and we discuss the rationality as below.

- *Discussion*: labels $y_a$ and $y_b$ corresponding to samples $x_a$ and $x_b$ originating from different categories in binary-classification are +1 and -1. Induce the cohesive pair-wise constraint: $(y_a+y_b)\|H^T x_a-H^T x_b\|\leq\rho$ in the objective in Equation (5) will optimal within-class samples to its cluster. According to Equation (4), the minimized distance-sum of within-class samples is defined the constant $S$, then we can obtain:

$$\sum_{k=1}^{K} \sum_{i=1}^{N} \|H^T x_i - c_k\|_2^2 = S. \qquad (6)$$

In the equation above, if both of samples $x_a$ and $x_b$ belong to the same category, the constraint is transformed into $(H^T x_a-H^T x_b)\leq\rho/(y_a+y_b)$. This combines with the objective of within-class, and can obtain the equation:

$$\sum_{k=1}^{K} \sum_{i=1}^{N} \|H^T x_i - c_k\|_2^2 - (H^T x_a - H^T x_b), < S - \rho \;/\; (y_a + y_b) \qquad (7)$$

where $\rho/(y_a+y_b)$ is rather small value according to the condition $y_a$=1 and $y_b$=-1. Therefore, we can get $S$-$\rho/(y_a+y_b)$<$S$ is more optimal solution for with within-class. If both of samples $x_a$ and $x_b$ belong to different categories, the constraint will equal to 0 that does not distribute to minimize the solution for inter-class.

### 3.2. Deep Embedding for Semi-Supervised Clustering

According to above model, we solve the optimal embedding and clustering centers according to the sample and category distribution by means of the supervised information from the rather few labeled samples, and utilize the cohesive pair-wise constraints to enhance the clustering performance. However, this linear projection of the model greatly limits the semi-supervised clustering performance when samples locate in high-dimension subspace, for instance, image datasets. The cause can be summarized as the superelevation dimension and sparse features of the image samples. In order to obtain more discriminative features and the optimized clustering model, we proposed a novel semi-supervised deep embedding clustering model.

The proposed model as shown in Figure 1 with the aid of CNN-Autoencoder deep-network framework by constructing the encoder and decoder network structures, which map the samples from the original space to a high-dimensional nonlinear space, and leverage convolutional operations and then restore them by using convolutional operations. The reconstruction error between the restored samples and the original samples is used to train the encoder network. According to Equation (4), maximizing category distribution ensure the global structure of the embedding, moreover, utilize the cohesive pair-wise constraints from few labeled samples to optimize the distribution of samples within-class to enhance the discrimination of embedding from CNN-Autoencoder.

First, we utilize the network structure of CNN-Autoencoder to initialize the non-linear transformation $f_\theta$ that belongs to stronger representation ability. Each layer of the network is a denoising autoencoder that, after training, can reconstruct the output of the previous layer after random corruption. After training, we concatenate all encoders and decoders layer by layer together to form a deep autoencoder. The encoder layer is exactly what we need, as an initial mapping between the original feature space and the latent learned space $f_\theta$ learn the data $X$ embedded in the space $Z$ to be valid feature representations for the original input data examples, $f_\theta$: $X \to Z$.

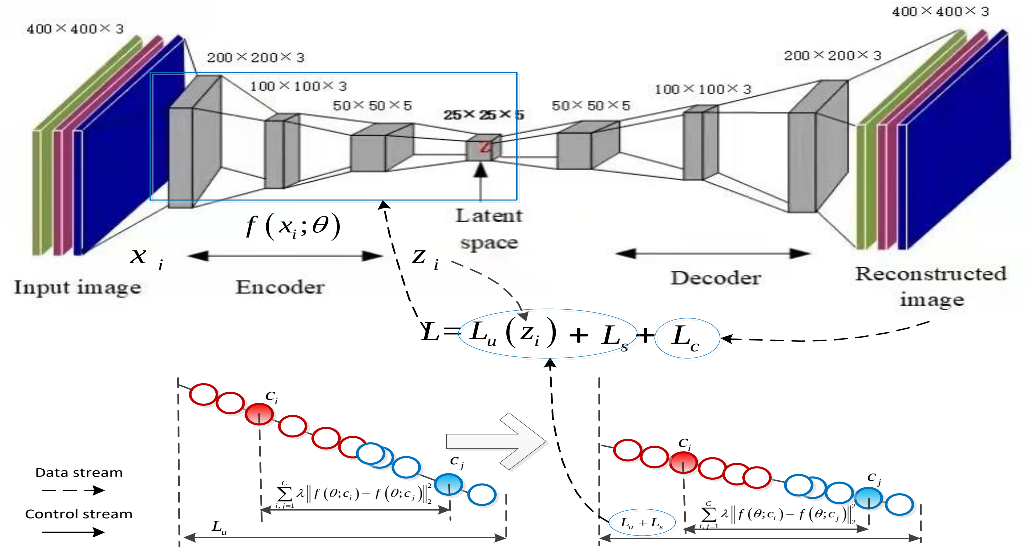


Figure 1. The framework of proposed model.

Second, for semi-supervised clustering, we optimize the deep embedding according the category distribution. Above proposed method that minimize the inter-class distances and maximize the intra-class distances is utilized to measure the clustering loss function between the embedding and cluster centers $c_k$ defined as:

$$L_u = \sum_{k=1}^{K} \sum_{i=1}^{N} \|f_\theta(x_i) - f_\theta(c_k)\|_2^2 - \lambda \sum_{k,j=1}^{K} \left\|f_\theta(c_k) - f_\theta(c_j)\right\| \quad (8)$$

where $f_\theta(x_i) \in Z$ corresponds to the embedded of $x_i \in X$, and $c_k$ is the center of the *i*-th cluster in the embedding space. The first term of $L_u$ indicates the Euclidean distance between $x_i$ and $c_k$, meanwhile, both of $x_i$ and $\underline{c_k}$ represent the features of the input pair samples and cluster centers extracted by the metric learning network respectively. We minimize the within-class and between-class distances after projection to strength the nonlinear transformation $x_i \in X$, and the deep neural network structure initialized by the encoder layer of CNN.

Finally, in order to avoid interferences resulting from samples located in the boundary, we leverage cohesive pair-wise constraints to the objective of DEC to learn features transformation and cluster assignment and the use of KL divergence, which uses points with high confidence as supervision and assemble samples located on within-class. However, it does not take into account the relationship between pairs and pairs leading to inability to identify pairwise constraints and limited support for semi-supervised, we introduce cohesive pairwise constraints through a compact criterion to guide clustering and the orientation of the embedding. Pair-wise constraints specify whether a pair of data examples belong to the same class (must link constraints) or belong to different classes (cannot link constraints). We expect that points of the same class should be closer, while points of different classes should be farther apart in the latent feature space. Therefore, we define the structurally constrained loss function as:

$$L_s = (y_a + y_b)\|f_\theta(x_a) - f_\theta(x_b)\|_2^2. \quad (9)$$

Notice, the equation minimizes the cost of violating constraints, enabling simultaneous learning of feature representations and performing cluster assignments to support user-specified constraints. In addition, autoencoder training uses Value-based reconstruction error minimization $L_c$, that is, the mean square error between the model output value and the original input is minimized, so that a deep learning network can be trained unsupervised, and the equation results shown that the overall loss function of the model can be divided into three parts namely unsupervised clustering loss $L_u$, structural constraint loss $L_s$ and reconstruction error loss $L_c$.

$$L = L_c + L_u + L_s, \quad (10)$$

where $L_u$ is the minimum difference loss between the within-class distance and the between-class distance, and it can learn a latent representation of the original data that is beneficial to the clustering task, and the structural constraint loss $L_s$ represents the learned representation between the embedding $f_\theta$ and the prior information. consistency. Intuitively, minimizing the equation, the distance between $f_\theta$ and cluster center $c_k$ will be close in latent space $Z$ if they are in the same class. Similarly, if in different classes, the distance between $f_\theta$ and cluster center $c_k$ will be very far in space $Z$. Therefore, the model not only learns a good representation of the clusters, but also makes points from the same class closer, while points from different classes are separated from each other.

## 4. Experiments

In this section, we verified the proposed model in popular datasets, and demonstrate the efficiency of model even in semi-supervised datasets with very few labeled samples, and analysis advantages when induce cohesive pair-wise constraint in the optimization model. Moreover, we compare with the-state-art methods, which achieve satisfactory performances. The experiments were run on Intel Core i7-8700 and NVIDIA Geforce RTX 2070 hardware environment, as well as Windows 10 operating system, Python3 language environment, and Pytorch deep learning framework.

### 4.1. Experimental Datasets and Evaluation Metrics

To investigate the performances and generality of different algorithms, we perform experiments on four popular datasets under standard experimental conditions:

- **STL-10** is a benchmark dataset commonly used for evaluating unsupervised and semi-supervised learning algorithms. It consists of a total of 10 classes, including airplain, bird, car, cat, deer, dog, horse, monkey, ship, and truck. There are a total of 5,000 training images and 8,000 test images, all in RGB format with the size of 96x96 pixels. [16], we also used HOG features and an 8×8 concatenated colormap as input.
- **MNIST (Modified National Institute of Standards and Technology)** is a commonly used dataset in computer vision and machine learning for evaluating algorithms related to classification and recognition of handwritten digits. It contains 60,000 training images and 10,000 test images, all in grayscale format with the size of 28x28 pixels, and we treat each grayscale image as a 784-dimensional vector. Each dimension is centered and normalized.
- **USPS (United States Postal Service)** is a dataset commonly used for evaluating algorithms related to classification and recognition of handwritten digits, similar to MNIST. It consists of a total of 9,298 training images and 1,979 test images, all in grayscale format with the size of 16x16 pixels. Compared to MNIST, USPS is relatively more challenging due to several factors such as lower image resolution, variations in writing styles, and distortions caused during the scanning process. The dataset also includes significant overlaps between some of the digits, which can make their recognition more difficult.
- **CIFAR-10 (Canadian Institute for Advanced Research)** is a widely used benchmark dataset in computer vision and machine learning for evaluating image classification algorithms. It consists of a total of 60,000 color images, with the size of 32x32 pixels, evenly distributed across 10 classes, including airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks. Each image is associated with a ground-truth label indicating the corresponding class it belongs to. We concatenate HOG features and 8×8 color-maps to represent each image, same as STL-10.

In this paper, we employ Clustering Accuracy (ACC) and Normalized Mutual Information (NMI) to evaluate the performances of the proposed and compared models.

### 4.2. Parameter Selection and Neural Network Structure

In the proposed model, the optimization formulate has been developed in Equation (5) by considering both of the class and intra-class distribution in the deep embedding. The parameter $\lambda$ as the intensively impact to balance both. In light of this, we seek a reasonable value in most datasets in the range [0.1, 1] with the interval of 0.2. As shown in Figure 2, the accuracy value changes with $\lambda$, and achieves the best results when the value is 0.5 in two popular datasets. These results portray the significance of cohesive pair-wise constraints for the problem of semi-supervised learning, especially in very few labeled samples. The $\lambda=0.5$ is use in below experiments. Moreover, in the following sections, we will provide an analysis of how the number of labeled samples affects the performance of the model.

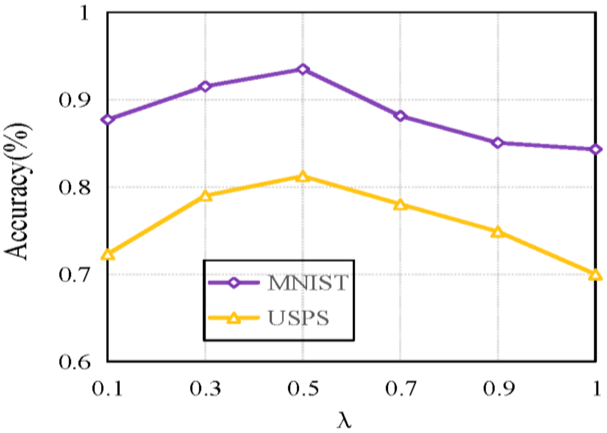


Figure 2. The relationship between the balance impact $\lambda$ and the accuracy value.

In order to avoid dataset-specific tuning as much as possible, we are inspired by IDEC [7] and SDEC [16] to set the network dimension to d-500-500-2000-10 for all datasets, where d is the data space dimension, which is different between datasets. All layers are tightly (fully) connected. Except for the input layer, output layer, and embedding layer, all inner layers are activated by Relu nonlinear function and optimized by Adam. We pre-train and fine-tune the autoencoder with the same parameter settings as SDEC to minimize the effect of parameter tuning.

### 4.3. Results and Analysis

We randomly select pairs of data points from the dataset: if two data points share the same label, then we generate a constraint that must be linked. Otherwise, and cannot-linked constraint will be generated. Through the proof of cannot-linked pairs regularization, each cannot-linked constraint is guaranteed, so that the associated points enter different clusters. Combined with the must-linked constraint, CPCE significantly improves the clustering performance, so that a large number of connected points in the ambiguous region can be correctly clustered. The learning rate of SGD is 0.001. Set the convergence threshold from to l% to 0.1%. For all algorithms, we set the number of clusters corresponding with the number of ground truth classes. We run each algorithm 10 times independently and report the average results.

- **Results and Comparison**

To evaluate the effectiveness of our proposed algorithm CPCE, we compare it with several benchmark algorithms. We first compare our algorithm with IDEC and SDEC. The classical k-means algorithm is applied to both the original feature space and the embedded feature space. We perform k-means while generating strong constraints through the proof of cannot-linked pairs regularization, so that a large number of connected points in ambiguous regions can be clustered correctly. The details of the comparative clustering methods are as follows: compared with traditional clustering methods k-mean, KM-cst, our method can learn more meaningful and robust features through deep embedding. Furthermore, k-mean and KM-cst are unsupervised methods that do not use label information during the clustering process, which further impairs their performance. The clustering performance of AE+KM in the learning space is significantly better than that of k-mean in the original data space, indicating that its deep neural network with large nonlinear representation ability is indeed beneficial to the clustering task. Compared with the deep clustering methods DEC, IDEC, SDEC, the reasons for the improved performance of CPCE are as follows: DEC and IDEC ignore the utilization of labeled data information, unlabeled samples are only used for regularization, and SDEC adopts pairwise constraints to guide the clustering. Class direction, which is similar to the contrastive loss. On this basis, we add unsupervised clustering loss, structural constraint loss and reconstruction errors to jointly learn representations and cluster assignment. Therefore, the CPCE framework optimizes the embedding and fully utilizes the prior information encapsulated in pairwise constraints is exploited, improving the overall quality of the final result.

As shown in Tables 1 and 2, the clustering results measured by ACC and NMI, respectively. In each row, the best and comparable results are shown in bold. In order to save space, the Standard Deviation (std) is not reported, in fact, the std value of CPEC is very small (i.e., the std value obtained by CPEC on USPS, STL-10, CIFAR-10, MNIST is 0.05%, 0.03%, 0.03%, respectively %, and 0.22%). Some observations can be made from Tables 1 and 2:

1. As shown in the Table 1, the clustering performance of k-means (AE+KM) in the learning space is significantly better than that in the original data space, indicating that the strong nonlinear representation ability of deep neural networks is indeed beneficial to the clustering task.
2. Three algorithms based on the deep embedded clustering framework (i.e., DEC, IDEC and SDEC) jointly deep representations and cluster assignment, which is better than AE+-means(AE+KM) means iteratively update features according to cluster assignment and learn latent representations for clustering.
3. KM-cst generally outperforms k-means in both the original space and the embedding space. This shows that incorporating pairwise information does improve clustering performances.
4. SDEC incorporates pairwise constraints to guide features learning, which suggests that pairwise constraints play an important role in improving performances.
5. The performance of CPCE is the best, and its performance is better than unsupervised deep embedded clustering algorithms DEC, IDEC and semi-supervised clustering SDEC. Specifically, CPCE on MNIST, ACC, and NMI are significantly improved than SDEC. It is shown that the CPCE framework optimizes representations and utilizes the regularization term to fully excavate the prior information, i.e., data samples from the same cluster are forced to be close to each other, while data samples from different clusters are in the learning feature away from each other in space. Combining the three loss functions of unsupervised clustering loss $L_u$, structural constraint loss $L_s$ and reconstruction error loss $L_c$, provides more accurate and robust results, significantly improving clustering performance.

Table 1. Clustering results measured by ACC (%).

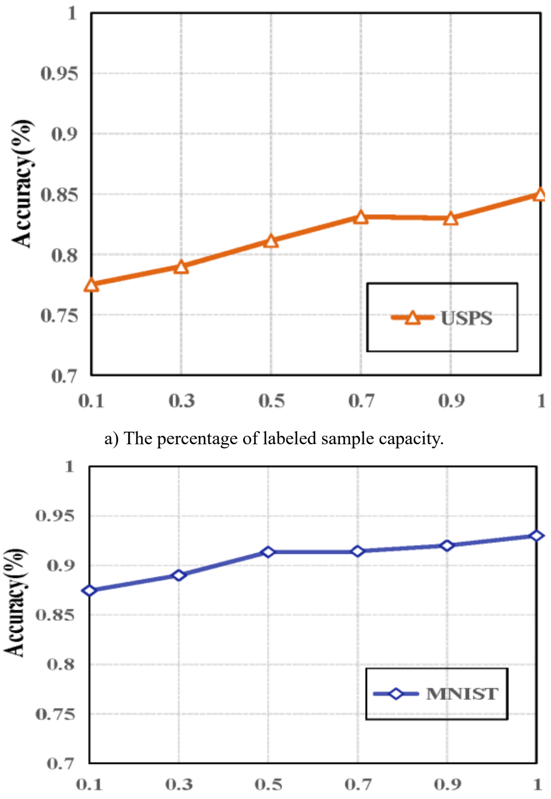
| Data | k-mean | KM-cst | AE+KM | DEC | IDEC | SDEC | CPEC |
|---|---|---|---|---|---|---|---|
| **USPS** | 65.67 | 68.18 | 70.28 | 75.81 | 75.86 | 76.39 | **79.01** |
| **STL-10** | 28.31 | 29.09 | 34.00 | 37.40 | 36.99 | 38.86 | **52.20** |
| **CIFAR-10** | 23.75 | 23.91 | 23.89 | 26.26 | 25.02 | 27.26 | **50.58** |
| **MNIST** | 52.98 | 54.27 | 54.27 | 84.94 | 83.85 | 78.12 | **91.35** |

Table 2. Clustering results measured by NMI (%).

| Data | k-mean | KM-cst | AE+KM | DEC | IDEC | SDEC | CPEC |
|---|---|---|---|---|---|---|---|
| **USPS** | 62.00 | 63.94 | 66.38 | 76.91 | 77.68 | 77.68 | **78.60** |
| **STL-10** | 24.40 | 24.79 | 29.37 | 32.43 | 32.53 | 32.84 | **45.44** |
| **CIFAR-10** | 14.67 | 14.21 | 15.80 | 16.99 | 17.27 | 17.20 | **41.37** |
| **MNIST** | 49.74 | 50.47 | 72.26 | 81.60 | 77.89 | 82.89 | **88.51** |

- **Semi-Supervised Clustering Analysis**

The dilemma of semi-supervised clustering lies in that very few supervised information is difficult to guide construction the global structure, especially in weakly discriminative features. Our model solves the optimal embedding with the metrics both of maximizing the distance of intra-class samples and minimizing the distance of within-class samples, and only use very few labeled samples to construct pair-wise constrains to control the samples located in the classification boundary, which enhance the performances. In general datasets, we demonstrate this conduct by experiments that set different environments with different sizes of labeled samples to execute the proposed model. As shown in Figure 3, the accuracies appear ever-increasing with the size of labeled samples growing. In MNIST dataset, the accuracy of the proposed model gets 0.87% even the percentage of labeled sample capacity is 10%, and the accuracy gets 0.91% when the percentage is 50%, which is rather closing to supervise learning. Aiming to the features of samples with low discrimination in difference-class, such as USPS dataset, our model also achieves the satisfactory performance using 10% of labeled samples, as shown in Figure 3.



a) The percentage of labeled sample capacity.

b) The percentage of labeled sample capacity.

Figure 3. The accuracy variation with the scale minimization of labeled samples in USPS and MNIST datasets.

## 5. Conclusions

In this paper, we propose model to optimize the local and global clustering structures simultaneously, and induce must pair-wise constrains to solve ambiguous samples and obtain the compact solution and the better embedding than the original space. Furthermore, we utilize CNN-Autoencoder with constraints to overcome the negative influence of high-dimension on the global structure in the calculation. The proposed model is verified in popular datasets, improvement the performance of semi-supervised clustering significantly. In the future, the research will proceed and focus on the remained challenge that is incremental semi-supervised learning.

## Acknowledgments

This document is the results of the research project funded by National Science Foundation of China (No.61902165, No.61976109); Provincial Department of Education (No. LJKMZ20221425, No. JYTMS20231040).

## References

[1] Allam M. and Malaiyappan N., “Hybrid Feature Selection based on BTLBO and RNCA to Diagnose the Breast Cancer,” *The International Arab Journal of Information Technology*, vol. 20, no. 5, pp. 727-737, 2023. https://doi.org/10.34028/iajit/20/5/5

[2] Butler K., Davies D., Cartwright H., Isayev O., and Walsh A., “Machine Learning for Molecular and Materials Science,” *Nature*, vol. 559, no. 7715, pp. 547-555, 2018. DOI: 10.1038/s41586-018-0337-2

[3] Chen C., Wang Z., Wu J., Wang X., Guo L., and Li Y., “Interactive Graph Construction for Graph-Based Semi-Supervised Learning,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 9, pp. 3701-3716, 2021. doi: 10.1109/TVCG.2021.3084694.

[4] Chen L. and Zhong Z., “Adaptive and Structured Graph Learning for Semi-Supervised Clustering,” *Information Processing and Management*, vol. 59, no. 4, pp. 102949, 2022. doi: 10.1016/j.ipm. 2022.102949

[5] Diallo B., Hu J., Li T., Khan G., Liang X., and Zhao Y., “Deep Embedding Clustering Based on Contractive Autoencoder,” *Neurocomputing*, vol. 433, pp. 96-107, 2021. doi: 10.1016/j.neucom.2020.12.094.

[6] Goel S. and Tushir M., “A New Semi-Supervised Clustering for Incomplete Data,” *Journal of Intelligent and Fuzzy Systems*, vol. 42, no. 2, pp. 727-739, 2022. DOI:10.3233/JIFS-189744.

[7] Guo X., Gao L., Liu X., and Yin J., “Improved Deep Embedded Clustering with Local Structure Preservation,” *in Proceedings of the 26th International Joint Conference on Artificial Intelligence*, Melbourne, pp. 1753-1759, 2017. DOI:10.24963/ijcai.2017/243

[8] Han Y. and Wang T., “Semi-Supervised Clustering for Financial Risk Analysis,” *Neural Processing Letters*, vol. 53, no. 5, pp. 3561-3572, 2021. https://doi.org/10.1007/s11063-021-10564-0

[9] Khanali H. and Vaziri B., "An Improved Approach to Fuzzy Clustering Based on FCM Algorithm and Extended VIKOR Method," *Neural Computing and Applications*, vol. 32, no. 2, pp. 473-484, 2020. https://doi.org/10.1007/s00521-019-04035-w

[10] Kononenko I., "Machine Learning for Medical Diagnosis: History, State of the Art and perspective," *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89-109, 2001. https://doi.org/10.1016/S0933-3657(01)00077-X

[11] Li X., Yin H., Zhou K., and Zhou X., "Semi-Supervised Clustering with Deep Metric Learning and Graph Embedding," *World Wide Web*, vol. 23, no. 2, pp. 781-798, 2020. DOI:10.1007/s11280-019-00723-8

[12] Lu J., Hu J., and Zhou J., "Deep Metric Learning for Visual Understanding: An Overview of Recent Advances," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 76-84, 2017. DOI:10.1109/MSP.2017.2732900

[13] Miyato T., Maeda S., Koyama M., and Ishii S., "Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979-1993, 2018. DOI:10.1109/TPAMI.2018.2858821

[14] Nie F., Wang X., Jordan M., and Huang H., "The Constrained Laplacian Rank Algorithm for Graph-Based Clustering," *in Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Phoenix, pp. 1969-1976, 2016. https://doi.org/10.1609/aaai.v30i1.10302

[15] Nie F., Zhang H., Wang R., and Li X., "Semi-Supervised Clustering Via Pairwise Constrained Optimal Graph," *in Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3160-3166, Yokohama, 2020. https://doi.org/10.24963/ijcai.2020/437

[16] Ren Y., Hu K., Dai X., Pan L., Hoi S., and Xu Z., "Semi-Supervised Deep Embedded Clustering," *Neurocomputing*, vol. 325, pp. 121-130, 2019. https://doi.org/10.1016/j.neucom.2018.10.016

[17] Solorio-Fernández S., Carrasco-Ochoa J., and Martínez-Trinidad J., "A Review of Unsupervised Feature Selection Methods," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 907-948, 2020. https://doi.org/10.1007/s10462-019-09682-y

[18] Suárez J., García S., and Herrera F., "A Tutorial on Distance Metric Learning: Mathematical Foundations, Algorithms, Experimental Analysis, Prospects and Challenges," *Neurocomputing*, vol. 425, pp. 300-322, 2021. https://doi.org/10.1016/j.neucom.2020.08.017

[19] Tanemura K., Das S., and Merz K., "AutoGraph: Autonomous Graph-Based Clustering of Small-Molecule Conformations," *Journal of Chemical Information and Modeling*, vol. 61, no. 4, pp. 1647-1656, 2021. https://doi.org/10.1021/acs.jcim.0c01492

[20] Wen J., Varol E., Sotiras A., and Yang Z., "Multi-Scale Semi-Supervised Clustering of Brain Images: Deriving Disease Subtypes," *Medical Image Analysis*, vol. 75, pp. 102304, 2022. https://doi.org/10.1016/j.media.2021.102304

[21] Xie J., Girshick R., and Farhadi A., "Unsupervised Deep Embedding for Clustering Analysis," *in Proceedings of the 33rd International Conference on Machine Learning*, New York, pp. 478-487, 2016. https://proceedings.mlr.press/v48/xieb16.html

[22] Xu Z., Liu B., Zhe S., Bai H., and Wang Z., "Variational Random Function Model for Network Modeling," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 318-324, 2019. DOI:10.1109/TNNLS.2018.2837667

[23] Yu Z., Luo P., Liu J., Wong H., and You J., "Semi-Supervised Ensemble Clustering Based on Selected Constraint Projection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 12, pp. 2394-2407, 2018. DOI:10.1109/TKDE.2018.2818729.

[24] Zhang Y., Wang H., Yang Y., and Zhou W., "Deep Matrix Factorization with Knowledge Transfer for Lifelong Clustering and Semi-Supervised Clustering," *Information Sciences*, vol. 570, pp. 795-814, 2021. https://doi.org/10.1016/j.ins.2021.04.067

**Jing Zhang** born in 1984. PhD, associate professor. Her main research interests include Machine Learning and Reinforcement Learning.

**Guiyan Wei** born in 1997. Master. Her main research interests include Deep Learning and Machine Learning.

**Yonggong Ren** born in 1972. PhD, professor. His main research interests include Data Mining and Artificial Intelligence.